Aline Villavicencio · Viviane Moreira
Alberto Abad · Helena Caseli · Pablo Gamallo
Carlos Ramisch · Hugo Gonçalo Oliveira
Gustavo Henrique Paetzold (Eds.)

LNAI 11122

Computational Processing of the Portuguese Language

13th International Conference, PROPOR 2018
Canela, Brazil, September 24–26, 2018
Proceedings

Springer

Lecture Notes in Artificial Intelligence 11122

Subseries of Lecture Notes in Computer Science

More information about this series at http://www.springer.com/series/1244

Aline Villavicencio · Viviane Moreira
Alberto Abad · Helena Caseli · Pablo Gamallo
Carlos Ramisch · Hugo Gonçalo Oliveira
Gustavo Henrique Paetzold (Eds.)

Computational Processing of the Portuguese Language

13th International Conference, PROPOR 2018
Canela, Brazil, September 24–26, 2018
Proceedings

Springer

Editors

Aline Villavicencio
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre
Brazil

and

CSEE, University of Essex
Colchester
UK

Viviane Moreira
Instituto de Informática - UFRGS
Porto Alegre
Brazil

Alberto Abad
INESC-ID
Lisbon
Portugal

Helena Caseli
UFSCAR
Sao Carlos
Brazil

Pablo Gamallo
Centro Singular de Investigación en Tecnoloxías
Universidade de Santiago de Compostela
Santiago de Compostela, La Coruña
Spain

Carlos Ramisch
Université de Toulon
Parc Scientifique Technologique Luminy
Marseille
France

Hugo Gonçalo Oliveira
Centro de Informática e Sistemas
Universidade de Coimbra
Coimbra
Portugal

Gustavo Henrique Paetzold
Federal University of Technology
Dois Vizinhos, Paraná
Brazil

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-99721-6 ISBN 978-3-319-99722-3 (eBook)
https://doi.org/10.1007/978-3-319-99722-3

Library of Congress Control Number: 2018952343

LNCS Sublibrary: SL7 – Artificial Intelligence

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The International Conference on Computational Processing of Portuguese (PROPOR) is the most important scientific event in natural language processing dedicated to the Portuguese language. It covers both theoretical and technological advances, simultaneously dealing with spoken as well as with written dimensions. PROPOR 2018 was the 13th edition of the event, which is held every two years, alternating between Brazil and Portugal. Previous events were held in Lisbon/Portugal (1993), Curitiba/Brazil (1996), Porto Alegre/Brazil (1998), Évora/Portugal (1999), Atibaia/Brazil (2000), Faro/Portugal (2003), Itatiaia/Brazil (2006), Aveiro/Portugal (2008), Porto Alegre/ Brazil (2010), Coimbra/Portugal (2012), São Carlos/Brazil (2014), and Tomar/Portugal (2016).

The meeting is a rich forum for the exchange of ideas and partnerships for the research communities dedicated to the automated processing of Portuguese, promoting the development of methodologies, resources, and projects that can be shared among researchers and practitioners in the field. This 13th edition of PROPOR took place in Canela, in the south-east of Brazil, during September 24–26, 2018, and was organized by the Institute of Computer Science of the Federal University of Rio Grande do Sul (UFRGS).

The event featured the fifth edition of the MSc/MA and PhD Dissertation Contest, which rewards the best academic work in Portuguese language processing by young researchers. Moreover, PROPOR 2018 included a workshop for the demonstration of software and resources for Portuguese processing, as well as the Second Job-Shop and Innovation Forum (JIF), aiming at the creation of a fruitful environment to promote the exchange of results, skills, and partnerships between researchers in the field of language technologies and companies using and developing these technologies. In 2018, PROPOR also had two workshops in addition to the main program, namely, the Student Research Workshop (SRW), which provided a venue for students in computational linguistics, linguistic resources, and natural language processing to present their work, with a focus on Portuguese and related languages, and the First Workshop on Linguistic Tools and Resources for Paraphrasing in Portuguese (POP@PROPOR2018). PROPOR also hosted the first edition of the Latin American and Iberian Languages Open Corpora Forum (OpenCor).

Three keynote speakers honored the event with their lectures: Marie-Catherine de Marneffe (Linguistics Department, The Ohio State University, USA), Lori Lamel (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, LIMSI, France), and Osvaldo N. Oliveira Jr (Interinstitutional Center for Computational Linguistics, NILC, São Carlos Institute of Physics, University of São Paulo, Brazil).

A total of 92 submissions were received for the main event, involving 165 authors from many institutions worldwide, such as Belgium, Brazil, France, Macau, Portugal, Spain, UK, or USA. This volume brings together a selection of the 45 papers accepted at the main conference: 42 long papers and three short papers. Each submission was reviewed by three reviewers and the overall acceptance rate (including both long and short papers) was 49%. To these, four papers corresponding to the selected submissions of the MSc/MA and PhD dissertation contest were added. In this volume, the papers are organized thematically and include the most recent developments in corpus linguistics, information extraction, language applications, language resources, sentiment analysis and opinion mining, speech processing, and syntax and parsing.

Our sincere thanks to every person and institution involved in the complex organization of this event, especially to the members of the Program Committee of the main event, the dissertations contest and the associated workshops, the invited speakers, and the general organization staff. We are also grateful to the agencies and organizations that supported and promoted the event.

July 2018

Aline Villavicencio
Viviane Moreira
Alberto Abad
Helena Caseli
Pablo Gamallo
Carlos Ramisch
Hugo Gonçalo Oliveira
Gustavo Henrique Paetzold

The information in the original version of the preface was not complete. The corrected preface will have the below information at the end of third paragraph.
PROPOR also hosted the first edition of the Latin American and Iberian Languages Open Corpora Forum (OpenCor).

Organization

General Chairs

Aline Villavicencio	Institute of Informatics, UFRGS, Brazil and CSEE, University of Essex, UK
Viviane P. Moreira	Institute of Informatics, UFRGS, Brazil

Program Chairs

Alberto Abad	IST/INESC-ID, Portugal
Carlos Ramisch	Aix-Marseille University, France
Helena Caseli	Universidade Federal de São Carlos, Brazil
Pablo Gamallo	University of Santiago de Compostela, Spain

Editorial Chairs

Gustavo Henrique Paetzold	Federal University of Technology Paraná, Brazil
Hugo Gonçalo Oliveira	CISUC, University of Coimbra, Portugal

Best MSc/MA and PhD Dissertation Contest Chair

António Teixeira	University of Aveiro, Portugal

Demos Committee

Fernando Batista	INESC-ID and ISCTE-IUL, Portugal
Rodrigo Wilkens	Université Catholique de Louvain, Belgium
Valeria de Paiva	Nuance Comms and University of Birmingham, UK

Student Research Workshop Chairs

Amália Mendes	Centro de Linguística da Universidade de Lisboa, Portugal
Daniel Beck	University of Melbourne, Australia
Livy Real	University of São Paulo, Brazil

Co-located Workshops Committee

Ana R. Luís	CELGA, University of Coimbra, Portugal
Fernando Perdigão	IT, DEEC, University of Coimbra, Portugal
Jorge Baptista	University of Algarve and L2F-Spoken Language Lab, INESC ID, Lisboa, Portugal

Osvaldo Novais de Oliveira Jr.	USP, Brazil
Vládia Pinheiro	University of Fortaleza, Brazil

Tutorial Chairs

Alberto Simões	2Ai Lab, IPCA, Portugal
Diana Santos	Linguateca and University of Oslo, Norway
Diego Amâncio	USP-SC, Brazil

Job-Shop and Innovation Forum Chairs

David Martins de Matos	INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Fabio Kepler	Unbabel, Portugal

Advisory Committee for Diversity and Inclusion in Language Technologies

António Branco	Universidade de Lisboa, Portugal
Margarita Correia	CELGA/ILTEC, Portugal
Gilvan Müller de Oliveira	UFSC, Brazil

Organization Committee

Aline Villavicencio	Institute of Informatics, UFRGS, Brazil and CSEE, University of Essex, UK
Leonardo Zilio	Université Catholique de Louvain, Belgium
Rodrigo Wilkens	Université Catholique de Louvain, Belgium
Roger Prates de Pelle	UFRGS, Brazil
Viviane P. Moreira	Institute of Informatics, UFRGS, Brazil

Social Media and Communications Chairs

Marcely Boito	UFRGS, Brazil
Renata Ramisch	UFSCAR, Brazil

Webmaster

Roger Prates de Pelle	UFRGS, Brazil

Steering Committee

Alexandre Rademaker	IBM Research Brazil and EMAp/FGV, Brazil
António Branco	Universidade de Lisboa, Portugal

André Adami	Universidade de Caxias do Sul, Brazil
Sara Candeias	Microsoft, Portugal

Program Committee

Alberto Simões	2Ai Lab, IPCA, Portugal
Alexandre Rademaker	IBM Research Brazil and EMAp/FGV, Brazil
Aline Villavicencio	Institute of Informatics, UFRGS, Brazil and CSEE, University of Essex, UK
Amália Mendes	Centro de Linguística da Universidade de Lisboa, Portugal
Anabela Barreiro	INESC-ID, Portugal
André Adami	Universidade de Caxias do Sul, Brazil
António Branco	Universidade de Lisboa, Portugal
Antonio Serralheiro	INESC ID and Academia Militar, Portugal
Ariani Di Felippo	Universidade Federal de São Carlos, Brazil
Augusto Soares Da Silva	Catholic University of Portugal, Portugal
Berthold Crysmann	CNRS – Laboratoire de linguistique formelle, France
Brett Drury	Scicrop, Brazil
Bruno Cuconato	FGV/EMAp, Brazil
Bruno Martins	INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Carla Griggio	Inria, France
Carlos A. Prolo	UFRN, Brazil
Carolina Scarton	The University of Sheffield, UK
David Martins de Matos	INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Diana Santos	Linguateca and University of Oslo, Norway
Eraldo Rezende Fernandes	FACOM/UFMS, Brazil
Eric Laporte	Université Paris-Est Marne-la-Vallée, France
Erick Fonseca	University of São Paulo, Brazil
Erick Galani Maziero	University of São Paulo, Brazil
Evandro Gouvea	Interactions, USA
Fabio Kepler	Unbabel, Portugal
Fai Wong	University of Macau, SAR China
Fernando Batista	INESC-ID and ISCTE-IUL, Portugal
Fernando Perdigão	IT, DEEC, University of Coimbra, Portugal
Gaël Dias	Normandie University, France
Helena Caseli	Universidade Federal de São Carlos, Brazil
Helena Moniz	INESC/FLUL, Universidade de Lisboa, Portugal
Henrique Santos	PUCRS, Brazil
Hugo Gonçalo Oliveira	CISUC, DEI, University of Coimbra, Portugal
Hugo Rosa	INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Irene Rodrigues	Universidade de Évora, Portugal

Isabel Falé	Universidade Aberta/Centro de Linguística da Universidade de Lisboa, Portugal
Isabel Trancoso	INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Ivandre Paraboni	University of São Paulo, Brazil
João Silva	Universidade de Lisboa, Portugal
João Balsa	BioISI/MAS, Universidade de Lisboa, Portugal
Joaquim Llisterri	Universitat Autònoma de Barcelona, Spain
Jorge Baptista	University of Algarve and L2F-Spoken Language Lab, INESC ID Lisboa, Portugal
José David Lopes	Heriot Watt University, UK
Jose Ramom Pichel	imaxin—software, Spain
Larissa Freitas	UFPEL, Brazil
Laura Alonso Alemany	Universidad Nacional de Córdoba, Argentina
Leandro Henrique Mendonça de Oliveira	Empresa Brasileira de Pesquisa Agropecuria (Embrapa) – Secretaria de Pesquisa e Desenvolvimento (SPD), Brazil
Leonardo Zilio	Université Catholique de Louvain, Belgium
Luísa Coheur	INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Lucia Specia	The University of Sheffield, UK
Luciana Benotti	Universidad Nacional de Cordoba, Argentina
Luis Felipe Uebel	SIDIA – Samsung Instituto de Desenvolvimento para a Informática da Amazônia, Brazil
Mário Silva	Instituto Superior Técnico, Universidade de Lisboa, INESC-ID, Portugal
Magali Duran	University of São Paulo, Brazil
Marcelo Finger	University of São Paulo, Brazil
Marcos Garcia	University of Coruna, Spain
Marcos Treviso	University of São Paulo, Brazil
Marcos Zampieri	University of Wolverhampton, UK
Maria Das Graças Volpe Nunes	NILC- ICMC, University of São Paulo at São Carlos, Brazil
Maria Jose Bocorny Finatto	UFRGS, Brazil
Martín Pereira-Fariña	University of Santiago de Compostela, Spain
Mikel Forcada	DLSI, Universitat d'Alacant, Spain
Muntsa Padró	Eloquant, France
Nelson Neto	Universidade Federal do Pará, Brazil
Norton Roman	USP, Brazil
Nuno Cavalheiro Marques	DI-FCT, Universidade NOVA de Lisboa, Portugal
Nuno Mamede	INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Oto Araújo Vale	Universidade Federal de São Carlos, Brazil

Pável Calado	INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
Palmira Marrafa	Universidade de Lisboa, Portugal
Patricia Martin-Rodilla	Institute of Heritage Sciences Spanish National Research Council, Spain
Patrick Blackburn	Roskilde Universitet, Denmark
Paula López Otero	Universidade da Coruña, Spain
Paulo Quaresma	Universidade de Évora, Portugal
Plinio Barbosa	University of Campinas, Brazil
Ranniery Maia	Federal University of Santa Catarina, Brazil
Renata Vieira	PUCRS, Brazil
Ricardo Ribeiro	INESC ID Lisboa/ISCTE-IUL, Portugal
Ricardo Rodrigues	CISUC and IPC, Coimbra, Portugal
Rodrigo Wilkens	Université Catholique de Louvain, Belgium
Rubén Solera-Ureña	INESC-ID Lisboa, Portugal
Sandra Aluisio	University of São Paulo/ICMC/NILC, Brazil
Sara Candeias	Microsoft, Portugal
Silvio Ricardo Cordeiro	Aix Marseille University, France
Teresa Gonçalves	University of Évora, Portugal
Thiago Pardo	University of São Paulo, Brazil
Thomas Pellegrini	Université de Toulouse, IRIT, France
Valéria Feltrim	Universidade Estadual de Maringá, Brazil
Valeria de Paiva	Nuance Comms and University of Birmingham, UK
Violeta Quental	PUC-Rio, Brazil
Vitor Rocio	Universidade Aberta/INESC TEC, Portugal

Contents

Corpus Linguistics

Information Extraction

Natural Language Processing Applications

Language Resources

Sentiment Analysis & Opinion Mining

Speech Processing

Syntax & Parsing

Corpus Linguistics

Analyzing the Rhetorical Structure of Opinion Articles in the Context of a Brazilian College Entrance Examination

Karina Soares dos Santos, Mariana Soder, Bruna Stefany Batista Marques,
and Valéria Delisandra Feltrim(✉)

State University of Maringá, Maringá, PR, Brazil
{ra89149,ra95381,ra103404,vdfeltrim}@uem.br

Abstract. In this paper we present a study about the rhetorical structure of opinion articles that have been written as part of college entrance examination. For that, we defined a set of rhetorical categories that aim at modeling the structure of opinion articles produced in this specific context and used it to manually annotate a corpus. Results of the annotation experiment showed substantial agreement among the annotators and disagreements were settled using the majority vote to build a gold standard corpus. This corpus was then used to build automatic classifiers that assign one of the possible categories to each sentence of the opinion article. Experimental results regarding the classification were promising considering the model's simplicity and the reduced number of training instances.

Keywords: Opinion article · Rhetoric structure · Classification

1 Introduction

Since 1998 the *Brazilian National Curricular Parameters* propose the teaching of the Portuguese language to be based on textual genres. In this practice, students are encouraged to interpret and produce texts that are contextualized in a defined culture and social situation. Besides bringing changes in teaching practices, this proposal also brought changes in relation to criteria for evaluation of written production, especially in the context of selective processes for entrance in higher education institutions [2].

As a consequence, some Brazilian universities have been requiring proficiency in different textual genres in their admission selective processes. An example is the State University of Maringá. Since 2008, candidates participating in its entrance examination process are required to produce texts of specific genres, which are selected among a list that is disclosed in advance and periodically updated by the institution.

According to Menegassi [8], textual genres are materialized in texts produced in a defined social situation and that present relative stability of its elements,

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 3–12, 2018.
https://doi.org/10.1007/978-3-319-99722-3_1

namely: thematic content, style, and compositional construction. While style is related to the linguistic resources used to build the text, making it a meaningful, cohesive, and coherent unit, compositional construction refers to the organization given to the content, following a structure that can be somewhat standardized, and which is characteristic of the genre.

Aspects related to the compositional construction of a discourse has been addressed by Dijk and Kintsch [13] as the concepts of macro and superstructure. According to the authors, within the levels of semantic organization of discourse, these concepts schematize the structure of the text and characterize certain types of discourse, thus defining schematic categories that are used to organize it. These concepts have been used as theoretical basis for studies that aim at characterizing textual genres in terms of a schematic/rhetorical structure that can be learned and reproduced.

In the context of the texts produced as part of admission selective processes, their conformance to a rhetorical structure expected for the textual genre in question is one of the elements that is usually considered in the evaluation of these texts. Therefore, being able to automatically detect such structures is one of the steps towards the construction of automated scoring systems [4,10] and writing tools [5,11] focusing on a particular genre.

In this paper we present a study about the rhetorical structure of opinion articles that have been written as part of UEM's entrance examinations. This genre has already been requested in previous selective processes and, according to the exams' program published by the institution[1], it may be requested in 2018 winter entrance examinations. To conduct this study, we defined a set of rhetorical categories that aim at modeling the structure of opinion articles and used it to manually annotate a corpus. Results of the annotation experiment showed substantial agreement among the annotators and the majority vote was used to build a gold standard (GS) corpus. The GS corpus was then used to build automatic classifiers based on superficial features. The classification results were promising considering the model's simplicity and the reduced number of training instances used. It is worth noticing that such classifiers may be used as part of a future automated scoring system for texts produced in UEM's entrance examinations.

The remaining of this paper is organized as follows. Section 2 details the opinion article genre and presents the rhetorical structure proposed in this study. Section 3 describes our corpus and its annotation. Section 4 presents the built classifiers, and experimental results are presented in Sect. 5. Final remarks and directions for future works are presented on Sect. 6.

2 Opinion Article as a Genre

As stated by Zanini [15], the opinion article is a dissertative-argumentative text materialized in a journalistic context. Its thematic covers topics of interest to the

[1] http://www.cvu.uem.br/.

society (at least at the time that the text was written), expressing the author's opinion on the approached subject. As in other dissertative-argumentative genres, authors make use of facts and evidences to build their argument. However, despite bringing objective evidence of convincing, it is also permeated by subjectivity, due to the expression of the author's point of view.

Regarding the subjectivity of the genre, it is important to point out that when produced in the context of a college entrance examinations opinion articles tend to be more subjective, since they are written by students who are required to position themselves in a predefined non-realistic context.

Zanini [15] defines a compositional structure for the opinion article organized as: title, introduction, expansion (development), conclusion, and signature (of the author). The author points out that while some of these rhetorical movements may occur in other argumentative genres as well, title and signature are characteristics of opinion articles, inherited from its journalistic context.

2.1 Rhetorical Structure

Our search for a prototypical rhetorical structure for opinion articles were based on the studies of Bakhtin [3] and Van Dijk [14]. The theoretical analysis of textual genre created by Bakhtin [3] contributed to understanding the relationship between the notion of genre and the compositional structure aimed in this study. The studies of van Dijk [14] on the structure of discourse at the macrostructural and superstructural levels guided us in the elaboration of a rhetorical structure that would model characteristics of the genre, and at the same time, be appropriate for computational applications. The balance between these two aspects guided our study at all times.

With that in mind, we manually analyzed a sample of opinion articles that have been produced as part of an entrance examination. This analysis led to several refinements in our proposal and helped us finding a structure that could be considered prototypical. The resulting rhetorical structure, composed of seven categories, is presented in Fig. 1.

In Fig. 1, the dotted lines indicate optional categories, namely Title (s0) and Author (s4), which specify the starting and ending limits of the opinion article. Although these two categories are especially characteristic of the genre in a journalistic context, they are not always mandatory in the context of entrance examinations' opinion articles. When used, the title statement gives a preview of the subject approached and the author's position, and the signature/author serves to emphasize the author and her social relevance, which may bring credibility to the opinion expressed in the article.

At the center of Fig. 1 are the categories (or sections, therefore named as si) that define the conventional organization expected for this genre, namely: (s1) Introduction, (s2) Argumentation, and (s3) Conclusion. Note that the sequence s0, s1, ..., s4 equals the compositional structure defined by Zanini [15], but differently from the author, we refined Introduction and Conclusion into other categories: (t1) Theme and (t2) Thesis for the introduction, and (t3) Background

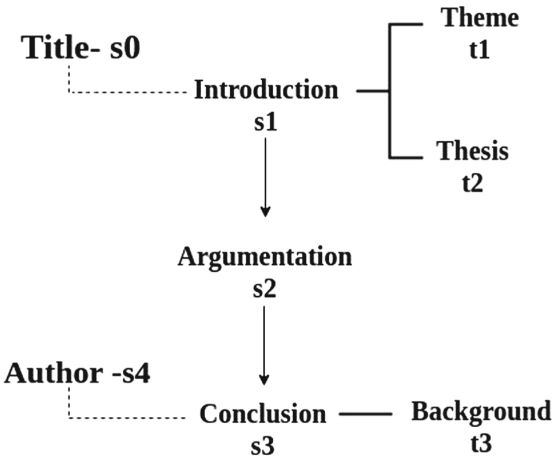


Fig. 1. Proposed rhetorical structure for opinion articles

for the conclusion. These refinements were included to cover aspects of the genre that are observed in the context of a college entrance examinations.

In this context, candidates are given instructions to write opinion articles about a specific subject and from a specific point of view. Therefore, introductions may have sentences that contextualize and set the subject treated in the article. We categorize these sentences as (t1) Theme. Besides setting the article's theme, the candidate should also state her own point of view, thus defining the thesis that she will defend. We categorize these sentences as (t2) Thesis.

Following the introduction, (s2) Argumentation presents the arguments that support the author's thesis. Sentences in this section present facts and ideas usually organized in a logical way. Argumentation tend to be the longer section of the article, since it has the purpose of persuading the reader about the author's point of view.

After presenting her arguments, the author must conclude. We observed that is a common practice in this section to resumption the initial thesis, not yet as a final conclusion, but as a mean of "binding" the different parts of the text together. In such cases, we categorize these sentences as (t3) Background. Finally, (s3) Conclusion brings a conclusive argument that does not overlap with the thesis, but introduce a final statement, which may or may not provide solutions to the problems discussed in the article.

3 Corpus

The corpus used in this study is composed of 271 texts produced by candidates to the 2014 and 2016 UEM entrance examinations. The texts were provided by the institution as scanned images of the original texts under a term of responsibility. Each text was then manually converted to text format respecting all the particulars of the original image, such as marks of punctuation, paragraphs, and spelling and grammatical errors. After that, the texts were converted into an XML format in which each text is represented as a sequence of sentences, totaling 2,562 sentences.

Three human annotators participate in the annotation of the corpus, two of them with a background in computer science and one in linguistics. They were previously informed about the proposed rhetorical structure and the function that each category has in it, and three randomly selected texts were annotated together to clarify possible doubts. After that, the annotators were asked to independently assign one of the categories presented in Fig. 1 to each sentence of the remaining of the corpus. As regards sentences reflecting more than one category, the annotators were instructed to assign the one they identified as more prominent for that particular sentence.

The agreement among the annotators measured by the Kappa statistics [9] was 0.78 (N = 2,532, k = 3, n = 7), which indicates substantial agreement [7]. To get a better understanding on the annotation results, we also estimated agreement for each category in isolation and the resulting Kappa values (N = 2,532, k = 3, n = 2) are presented in Table 1.

Table 1. Kappa value by category

Category	Kappa
s0 (title)	1.00
t1 (theme)	0.73
t2 (thesis)	0.70
s2 (argumentation)	0.77
t3 (background)	0.56
s3 (conclusion)	0.73
s4 (author)	1.00

As can be seen in Table 1, Kappa indicates at least substantial agreement for all categories except t3 (background), for which Kappa indicates moderate agreement. To verify our hypothesis that t3 might be confused with s3 (conclusion), we collapsed these two categories. The Kappa for this new setup raised to 0.81 (N = 2,532, k = 3, n = 6), evidencing difficulties of the annotators to distinguish between categories t3 and s3. Nevertheless, we decided to keep our seven-category structure since we believe that it better represent the corpus observations, and work on the improvement of the annotation guidelines, especially in the characterization of these two categories, for future studies.

All in all, we can conclude that our rhetorical structure annotation is reproducible and reliable enough to be used as training data for an automatic classifier. However, to do so, disagreements should be settled. We solved disagreements by assuming the category assigned by the majority of the annotators as correct. For cases in which the three annotators disagreed, we assumed the category assigned by the linguist as the correct one. Despite appearing arbitrary, this decision was made due to the consolidated knowledge of the rhetorical categories by the linguist.

The resulting annotated corpus, which we named golden standard (GS) corpus, was then used for training and testing the classifier described in the next section. The categories distribution in the GS corpus is presented in Fig. 2.

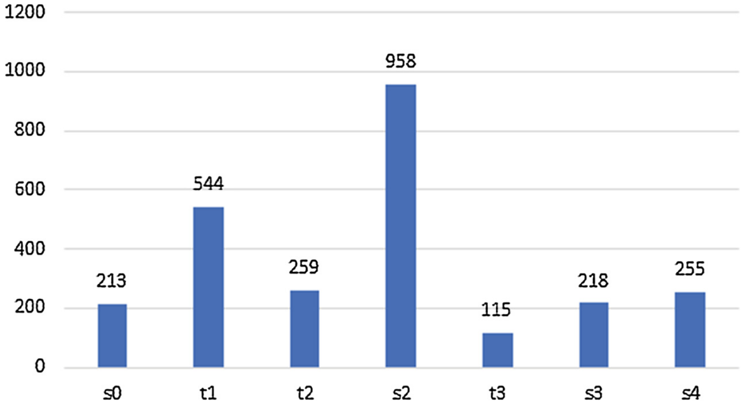


Fig. 2. Category distribution in the GS corpus

4 Classifier

Considering related works that approach the rhetorical structure identification as a classification problem [1,4–6,12], we trained a classifier that assigns one of the seven possible categories to the sentences of an opinion article. We used scikit-learn[2] implementations for preprocessing de corpus, extracting features, training and testing the classifiers.

The preprocessing of the corpus included sentence segmentation and tokenization. Our features were TF-IDF values, and we used a chi-squared distribution to select the most significant features. We have experimented with nine separate feature extraction pipelines by combining TF-IDF calculations and chi-squared thresholds: TF-IDF was estimated for unigrams, bigrams, and trigrams. For each of the three n-gram sets, we have used the chi-squared distribution to select 50, 100, and 1000 best features. We also evaluated the addition of the relative position of the sentence as a feature. In all cases (whenever possible), the feature vector for a sentence s_i contains, besides its own features, the features of sentences s_{i-1} and s_{i+1}.

We experimented with two different machine learning algorithms: support vector machines (SVM) and conditional random fields (CRF). While the first is

[2] http://scikit-learn.org/stable/.

a traditional approach for various pattern recognition tasks, the latter focus on sequence labeling problems, and rhetorical structures can be modeled as a sequence of labels. The SVM model used a linear kernel on a one-vs-the-rest scheme, and parameters were set to their default values. For the CRF model, parameters were set as follows: algorithm = 'lbfgs', c1 = 0.1, c2 = 0.1, max_iterations = 100, all_possible_transitions = True. Algorithm specifies the training algorithm, in this case, gradient descent using the L-BFGS method; c1 and c2 are coefficients for L1 and L2 regularization, respectively; max_iterations sets the maximum number of iterations for optimization; and all_possible_transitions defines if transition features that do not occur in the training data should be generated.

All models were evaluated using 10-fold cross-validation on the GS corpus. For the SVM, cross-validation was applied over a set of sentences, while for the CRF model cross-validation was applied over a set of texts (sentences sequences). Performance was measured in terms of precision, recall, and f1-score.

5 Results

We conducted several experiments combining the described feature sets and the mentioned classification algorithms. Both SVM and CRF performed better using unigrams, chi-squared to select the 100 best features, and the relative position of the sentence. The CRF model performed better than SVM for all metrics. The averaged f1-score was 0.75 for CRF and 0.71 for SVM. Therefore, we focused our analysis on the CRF classifier, whose results by class are shown in the Table 2.

Table 2. Results for the CRF classifier

Category	Precision	Recall	F1-score
s0 (title)	0.84	0.99	0.91
t1 (theme)	0.75	0.60	0.66
t2 (thesis)	0.56	0.49	0.53
s2 (argumentation)	0.77	0.91	0.83
t3 (background)	0.49	0.33	0.39
s3 (conclusion)	0.65	0.59	0.62
s4 (author)	1.00	0.97	0.98
Average	0.75	0.76	0.75

For most categories, the precision values were slightly higher than recall. Exceptions were categories s0 (title) and s2 (argumentation), for which recall was higher than precision. Despite these differences, precision and recall were reasonably balanced for all categories.

The two best f1-scores (0.91 and 0.98) were obtained for categories s0 (title) and s4 (author), respectively. This was expected, since these two categories are very prototypical and can be easily identified by their relative position in the text. Considering the remaining categories, results were better for category s2 (argumentation) (0.83), which is the most frequent category, and worse for categories t2 (thesis) and t3 (background) (0.53 and 0.39, respectively), which are among the less frequent ones.

Despite these observations, we have noted that f1-scores values were more related to the subjectivity associated with each category during the manual annotation than to their frequency in the corpus. In other words, categories for which the annotators disagreed more were also the most difficult for the classifier. This relation can be observed in Fig. 3, which compares the proportion of sentences, f1-scores and Kappa values for the categories in the corpus.

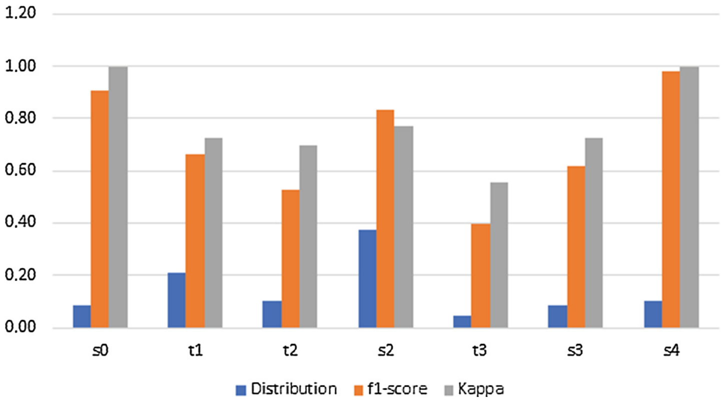


Fig. 3. Proportion of sentences, f1-scores, and Kappa values per category

6 Conclusion

This paper presented a study about the rhetorical structure of opinion articles produced in the context of a college entrance examinations. Based on related works concerning textual genres and discourse, as well as on the manual analysis of a corpus, we proposed a set of rhetorical categories that aim at modeling the structure of opinion articles produced in this specific context.

We manually annotated the corpus based on the rhetorical structure proposed in this study and experimental results showed that our annotation scheme is reproducible. However, we note that there is still room for improvement, especially regarding the refined categories proposed for the introduction and conclusion sections.

The resulting corpus was used to build classifiers that assign rhetoric categories to sentences of opinion articles. The best performance was obtained by a CRF classifier based on superficial features and results have shown that f1-scores for the classification and Kappa values for the manual annotations follow a similar distribution. From that we theorize that improvements in the GS corpus, especially regarding the categories pairs t1-t2 and t3-s3, will also lead to better classification results.

It is worth notice that this study was not concerned in evaluate the rhetorical structures of the opinion articles in the corpus. Instead, we focused on finding a prototypical structure that would be appropriate for describing our corpus in terms of its rhetorical moves, as well as for the construction of automatic classifiers. Therefore, future work includes a descriptive analysis of the corpus in terms of the structures found in it and how close/distant they are from a prototypical ideal one.

Regarding the automatic classification, we intent to focus future works on experimenting with a richer set of features, such the ones proposed by Teufel and Moens [12], and Andreani and Feltrim [1]. Also, as soon as we can expand the size of our training corpus, we intent to experiment with other learning schemes.

Acknowledgements. We would like to thank Araucária, State of Paraná Research Foundation, for the financial support.

References

1. Andreani, A.C., Feltrim, V.D.: Campos Aleatórios Condicionais Aplicados à Detecção de Estrutura Retórica em Resumos de Textos Acadêmicos em Português. In: Proceedings of Symposium in Information and Human Language Technology, pp. 111–120. Sociedade Brasileira de Computação (2015)
2. Antonio, J.D., Navarro, P.: Gêneros textuais em contexto de vestibular. EDUEM, Maringá (2017)
3. Bakhtin, M.: Estética da Criação Verbal. Wmf Martins Fontes, São Paulo (1997)
4. Burstein, J., Marcu, D., Knight, K.: Finding the WRITE stuff: automatic identification of discourse structure in student essays. IEEE Intell. Syst. **18**(1), 32–39 (2003)
5. Feltrim, V.D., Teufel, S., Nunes, M.G.V., Aluísio, S.M.: Argumentative zoning applied to critiquing novices' scientific abstracts. In: Shanahan, J.G., Qu, Y., Wiebe, J. (eds.) Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series, vol. 20, pp. 233–246. Springer, Dordrecht (2006). https://doi.org/10.1007/1-4020-4102-0_18
6. Fisas, B., Ronzano, F., Saggion, H.: On the discoursive structure of computer graphics research papers. In: Proceedings of the 9th Linguistic Annotation Workshop held in conjunction with NAACL 2015, pp. 42–54. Association for Computational Linguistics (2015)
7. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
8. Menegassi, R.J.: Aspectos sobre o gênero discursivo. In: Antonio, J.D., Navarro, P. (Org.) Gêneros textuais em contexto de vestibular, pp. 17–42. EDUEM, Maringá (2017)
9. Siegel, S., Castellan, N.J.J.: Nonparametric Statistics for the Behavioral Sciences, 2nd edn. McGraw-Hill, Berkeley (1988)
10. Shermis, M.D., Burstein, J.: Handbook of Automated Essay Evaluation Current Applications and New Directions. Routledge, New York (2013)

11. Souza, V.M.A., Feltrim, V.D.: A coherence analysis module for SciPo: providing suggestions for scientific abstracts written in Portuguese. J. Braz. Comput. Soc. **19**(1), 59 –73 (2012)
12. Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Comput. Linguists **28**(4), 409–445 (2002)
13. van Dijk, T.A., Kintsch, W.: Strategies of Discourse Comprehension. Academic Press, New York (1983)
14. van Dijk, T.A: Macrostructures: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition. Lawrence Erlbaum Associates, New Jersey (1980)
15. Zanini, M.: Artigo de opinião: do ponto de vista à argumentação. In: Antonio, J.D., Navarro, P. (Org.) Gêneros textuais em contexto de vestibular, pp. 43–58. EDUEM, Maringá (2017)

SMILLE for Portuguese: Annotation and Analysis of Grammatical Structures in a Pedagogical Context

Leonardo Zilio(✉), Rodrigo Wilkens, and Cédrick Fairon

Centre de traitement automatique du langage – CENTAL,
Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium
{leonardo.zilio,rodrigo.wilkens,cedrick.fairon}@uclouvan.be

Abstract. In Second Language Acquisition (SLA), the exposure of learners to authentic material is an important learning step, but the use of raw text may pose problems, because the information that the learner should be focusing on may be overlooked. In this paper, we present SMILLE for Portuguese, a system for detecting pedagogically relevant grammatical structures in raw texts. SMILLE's rules for recognizing grammatical structures were evaluated in random sentences from three different genres, achieving an overall precision of 84%. The automatic recognition of pedagogically relevant grammatical structures can help teachers and course coordinators to better inform the choice of texts to be used in language courses, while also allowing for the analysis of grammar profiles for SLA. As a case study, we used SMILLE to analyze pedagogical material used in a Portuguese as foreign language course and to observe how the predominance of grammatical content in the texts is related to the described grammatical focus of the language levels.

Keywords: Second Language Acquisition · Grammatical structures
Natural Language Processing · Grammatical parsing for Portuguese

1 Introduction

Research on the field of Second Language Acquisition (SLA) has already shown that the mere presentation of input to a language learner is not enough for ensuring that something will be learned [9]. This means that the language learner may process the input for its meaning alone, without noticing its linguistic structures, because there is no salient language information. Input is understood as language data that is potentially processable and made available, by chance or by design, to language learners [15]. On the other hand, the intake is the part of the input which is actually internalized by a learner and that can potentially be stored in the long-term memory [11].

Supported by the Walloon Region (Projects BEWARE 1510637 and 1610378) and Altissia International.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 13–23, 2018.
https://doi.org/10.1007/978-3-319-99722-3_2

An input in its raw form has lower chances of being converted into intake by the learner, and may thus not bring any new linguistic information. In the early 90's, Schmidt [12] developed the hypothesis that, in order for language learners to convert input into intake, they have to notice the relevant information in the input. Schmidt would later state, in a less controversial way, that "people learn about the things that they attend to and do not learn much about the things they do not attend to" [13]. There is much discussion regarding the assumptions of the Noticing Hypothesis, and it has its contesters (e.g. Truscott [18]). Nevertheless, it seems to be of general agreement that noticing is at least a facilitator of the language learning process, even though there are differences in the way that authors view the process of noticing, either as a purely conscious process or as a possibly unconscious process [5].

In this context, raw input may not present properties for drawing the learner's attention, especially when one deals with authentic texts, which are normally not meant as a language learning object. In an attempt to solve the lack of salience in raw input, Smith and Truscott [16] suggested the use of what they called "input enhancements", so as to give prominence to the relevant linguistic information. This "focus-on-form strategy" [6] has provided a new way to assist language learners, and some studies have shown that input enhancements represent a positive step in transforming input into intake (e.g. [10,14]).

In this paper, we present SMILLE (Smart and Immersive Language Learning Environment) for Portuguese[1], a system that can analyze and enhance written texts by employing Natural Language Processing (NLP) techniques for automatically retrieving pedagogically relevant grammatical structures. By highlighting pedagogically interesting structures in texts, SMILLE can be used by language teachers to automatically locate specific grammatical structures in texts and to evaluate if the texts that they are going to use with their learners are adequate to their language level. For instance, some structures introduced in a given moment may not be fully understood by the learners until they reach a more advanced level – this is specially the case for structures that are seen at the end of each level. As such, using a text that exposes the learner to all structures of a level may lead to over-exposure and cause the learner to lose focus. In this context, it is important to analyze to which structures the learner is being exposed during the second language learning process. Besides the more directly pedagogical application of SMILLE (i.e., enhancing grammatical structures), its association with a second language learning curriculum makes of it a useful tool for analyzing the pedagogical material of second language courses.

Since SMILLE prioritizes a pedagogical approach to information extraction, some of the automatically annotated grammatical structures are not directly recognized from part-of-speech-tagger or dependency-parser information (such as hidden and explicit subject and passive voice) and, thus, they are recognized through rules based on the more generic parser information. These rules are not trivial, and, thus, in this paper we want to especially observe the precision of

[1] The system is available for testing at https://cental.uclouvain.be/resources/smalla_smille/smille/.

SMILLE's rules, because, as stated by Meurers et al. [9], in language learning, precision tends to be more important than recall.

In addition to observing SMILLE's precision, we also put it to test in analyzing the pedagogical material used in a Portuguese as foreign language course. This analysis was designed to take into account the distribution of grammatical structures in the pedagogical material, in order to observe how the textual content presented to the learners aligns with the grammatical content that is taught in the different levels. In brief, this analysis is a profile of grammatical structures that occur in the texts of the pedagogical material, and this profile will then be contrasted with the structures presented in the course's handbooks. Our hypothesis is that the grammatical content of the handbook for the basic level will be significantly more prominent (95% confidence) in the basic level texts. Conversely, the grammar of the advanced level will be significantly more prominent in the texts for the advanced level.

This paper is organized as follows: we present information on related work in Sect. 2; next, we describe SMILLE's annotated structures in Sect. 3; Sect. 4 presents evaluation of SMILLE's annotation in different corpora; in Sect. 5, we explain and present the results of our experiment with the pedagogical material; and, finally, in Sect. 6 we present our final remarks and future work.

2 Related Work

In this section, we describe applications developed in the context of Second Language Acquisition (SLA) that can retrieve pedagogical information from raw texts and enhance it to learners.

The REAP project [3] is a tutoring system for English that focuses on finding authentic, Web-based texts that are suitable for the user in terms of reading level. The system also highlights words that are supposedly not known by the user. There is also a REAP.PT project [8] that was ported from English to European Portuguese and then further developed to also encompass gamification and user interaction in a 3D environment.

The WERTi system [9] allows for text enhancements of selected linguistic elements of English, Spanish and German. WERTi uses NLP tools combined with rules and regular expressions to retrieve text information for each of seven linguistic subjects: articles, determiners, gerunds, noun countability, phrasal verbs, prepositions, wh-questions.

SmartReader, developed by Azab et al. [1,2], provides a reading assistant tool that is fully based on the structures annotated by Stanford CoreNLP [7]. Each word in the text can also be clicked on to display semantic, syntactic and other information. It also displays syntactic function of selected words in the given sentence and generates simple questions about named entities, provided the answers are in the near context.

The *FLAIR* system is an online information retrieval system that annotates and re-ranks Web documents based on user-selected grammatical constructions [4]. It can recognize 87 different types of grammatical structures described in the official curriculum of English as foreign language used in German schools.

SMILLE distinguishes itself from these systems in how the enhancements are selected. For instance, in WERTi, only a few grammatical structures are available, while SmartReader and REAP present information more relevant to the meaning of the words and lexical units, and rely exclusively on parsing for retrieving grammatical structures. And, although FLAIR and SMILLE share a bigger scope in terms of detected grammatical information, the focus of FLAIR lies on text retrieval, while SMILLE focuses on the recognition process. In addition, with the exception of REAP, the other systems do not focus on Portuguese.

3 Grammatical Structures in SMILLE

SMILLE was originally developed for English [20,22,23], and then was further extended to Portuguese, both understood as a foreign language for the learner. This extension to Portuguese included a fully new set of rules that were developed based on the grammatical structures that are deemed important in a Portuguese-as-foreign-language course[2].

SMILLE links the detected information to the guidelines of the Common European Framework of Reference for Languages (CEFR) [19], so that the grammatical enhancements are not limited to isolated linguistic structures, but covers the needs for a given language level and for specific linguistic knowledge that is required from the learner in proficiency tests. By applying rules on top of the parser annotation[3], SMILLE also detects grammatical structures that are not directly retrieved from parsing. As such, for instance, teachers can select texts that are interesting according to their learner's preferences, while keeping an eye on important information in terms of linguistic structures that are relevant for their process of acquiring a second language.

SMILLE uses the PassPort system [21], a dependency parsing system based on Universal Dependency tags and PALAVRAS part-of-speech tags, as basis. Thus, much of the grammatical information that is detected by SMILLE for Portuguese requires only that the underlying parser correctly analyze the word or structure in question. Such is the case, for instance, of some adverbs, adjectives and simple verb tenses. However, some structures require rules on top of part-of-speech and dependency tags for retrieving more complex grammatical constructions, such as compound verb tenses, passive voice and relative clauses. And other structures still, such as comparatives and some adverbial phrases, are retrieved based on specific rules. As such, SMILLE combines the analysis done by the parser with hand-written rules to extract text information that would not be easily identified, and would not be salient, in a raw input. Several rules in SMILLE are the mere association of different part-of-speech tags and dependency tags or attachments (e.g., compound tenses), other rules require much more complex pattern matching, with multiple possibilities, especially in the case of comparatives, which can appear in several different forms.

[2] Our grammatical structures were based on the course developed by Altissia International (www.altissia.com).

[3] SMILLE for Portuguese makes use of the PassPort system [21].

While developing SMILLE, we had to make a decision regarding the granularity of grammatical structures and the escalation of knowledge associated to each language level. In a language course, different grammatical structures can be learned in progressive steps, so, for instance, today a language learner may study the relative pronoun "que" (approx. "that/which") and later, during another session, it is possible to learn the relative pronoun "quem" (approx. "who"). In authentic texts, the chances are that different pronouns will appear at the same time, interwoven in the text. To address this fine-grained differentiation, SMILLE would have to encompass specific rules for each case, sometimes for each word in a grammatical category. This would require more processing and an increase in the number of rules. So, although SMILLE respects the escalation related to different language levels (e.g., grammatical structures from different levels were separated in specific rules), the progression of content in the same level was overruled and generalized in overarching classes of grammatical structures, such as "relative pronouns".

SMILLE for Portuguese contains a total of 71 rules for recognizing pedagogically relevant grammatical structures in written texts. These rules encompass both the Brazilian and the European variants and are based on the CEFR levels from A1 to B2, and each rule is linked to a specific level. Here is a list of grammatical structures that SMILLE can detect in Portuguese texts[4]: prepositions, articles, use of pronouns "tu" and "você", pronouns used as indirect and direct complements, possessive pronouns, demonstrative pronouns, comparatives, adjectives, plural forms, nouns, expression of preferences, imperative, expressions of obligation, various verb tenses (including progressive ones), interrogative sentences, irregular verbs, uses of "ser", "estar", "ter" and "haver", diminutives, direct and indirect complements, superlative, final clauses, relative clauses and pronouns, verbal periphrases, numbers, possessives, indefinite pronouns, use of the pronoun "si", several types of adverbs and adverbs derived from adjectives, passive voice, hidden and explicit subjects, and use of clitics.

4 Evaluation of Selected Structures

For evaluating the quality of the rules used for recognizing the different, pedagogically relevant grammatical structures that SMILLE can display to the user, and to see which ones can be trusted for further analyses, we used random sentences from three different genres and applied SMILLE's pipeline of parsing and rules annotation.

First, we selected three different genres: literature[5], newspaper articles (from the Diário Gaúcho corpus[6]) and subtitles (from the Portuguese corpus of subti-

[4] We do not present here the 71 rules because many of the grammatical structures are divided along the CEFR levels, presenting some basic content in lower levels and reinforcing them in higher levels, and others are divided in different categories, such as the verb tenses, the comparative forms, the types of adverbs, etc.

[5] Selected romances from www.dominiopublico.gov.br.

[6] This corpus was compiled in the scope of the project PorPopular (www.ufrgs.br/textecc/porlexbras/porpopular/index.php).

tles compiled by Tiedemann [17]). We then annotated the corpora with SMILLE and randomly extracted 25 sentences for each structure from each corpus to be manually evaluated, totaling 75 sentences/instances per structure[7]. Finally, a manual evaluation was carried out by one linguist. Here are example sentences for the 20 structures that were manually evaluated in terms of precision (the main words associated with the structure are marked in *italic*):

1. **Adverbs of manner:** A mulher bateu *fortemente* no assaltante.
2. **Adverbs derived from adjectives:** Os papéis desapareceram *rapidamente*.
3. **Interrogative pronouns and adverbs:** *De que* estás a falar?
4. **Relative pronouns:** A árvore *que* está no centro do parque é a mais bonita.
5. **Superlative form of adjectives:** O Benfica é *o mais forte*.
6. **Adjectival comparative forms:** A Paula é *mais velha do que* a Susana.
7. **Extended comparative forms:** Um bebé *dorme mais do que* um adulto.
8. **Compound future tense:** *Vou estudar* muito para o exame.
9. **Forms of expressing obligation:** Elas *têm de* estudar muito
10. **Interrogative clauses in the present tense:** *Onde está* o teu amigo*?*
11. **Verbal periphrases:** *Começo a traduzir* agora mesmo.
12. **Compound pluperfect tense:** Eu já *tinha jantado* quando tu chegaste.
13. **Present continuous tense:** A Sofia *está trabalhando*.
14. **Reflexive pronouns:** O professor explica-*se* aos seus superiores.
15. **Final clauses:** Vim *para te ver*.
16. **Relative clauses:** O rapaz *com quem te encontraste* é muito giro.
17. **Hidden and explicit subjects:** *Disseram*-me que *eles* iam dormir aqui.
18. **Progressive tenses:** A Rita *esteve lavando* a cara.
19. **Passive voice:** O trabalho *foi terminado* ontem.

Table 1 shows results of the annotation divided by structure and genre. As we can see, most of the structures have high precision, so that the mean precision lies at 84.07% for the evaluated structures, and the median is 88%. The literature genre seems to pose more problems for the annotation, with a mean of 83.58% and a median of 84%; newspaper articles were worse in the mean precision, with 82.32%, but the median was much higher, at 92%; finally, the subtitles had the best mean precision, at 86.32%, and median, at 96%. In terms of individual structures, there were very few for which the genre seems decisive, and most of them had either generally bad performance, like reflexives, or generally good performance, like the progressive tenses. Even so, we see some structures, like the compound future or the hidden or explicit pronominal subjects, that present an unbalance in the precision evaluation towards one genre.

This precision evaluation showed us which structures can be used in further analyzing the pedagogical material in terms of content and organization per level. The material and the analysis are described in the next section.

[7] Sentences with more than one instance of the selected structure were evaluated only based on the first instance.

Table 1. Precision of automatically annotated grammatical structures

#	Structure	Newspaper	Literature	Subtitles	Total
1	Adverbs of manner	72%	92%	76%	80.0%
2	Adverbs from adjectives	96%	100%	100%	98.7%
3	Interrogative pronouns	92%	80%	96%	89.3%
4	Relative pronouns	80%	80%	52%	70.7%
5	Superlative	92%	100%	100%	97.3%
6	Comparative	100%	100%	100%	100.0%
7	Extended comparative	80%	76%	84%	80.0%
8	Compound future	92%	48%	92%	77.0%
9	Expression of obligation	96%	100%	100%	98.7%
10	Questions	80%	80%	92%	84.0%
11	Verbal periphrases	28%	48%	60%	45.3%
12	Plusperfect tense	100%	100%	100%	100.0%
13	Present continuous	100%	100%	100%	100.0%
14	Reflexive pronouns	44%	60%	28%	44.0%
15	Final clauses	92%	84%	96%	90.7%
16	Relative clauses	76%	68%	68%	70.7%
17	Hidden and explicit subject	56%	96%	100%	84.0%
18	Progressive tenses	100%	96%	100%	98.7%
19	Passive voice	88%	80%	96%	88.0%

5 Pedagogical Material: Analysis of Grammatical Distribution

Having evaluated the structures that SMILLE annotates by means of rules, we turned ourselves to the task of analyzing how courses of Portuguese as foreign language are organized in terms of grammar and how they present this information in the texts that exist in their pedagogical material. As a case study, we selected the material developed at the Universidade Federal de Juiz de Fora (UFJF) and used in its course for Brazilian Portuguese learners[8].

The corpus is composed of texts from handbooks used for teaching Brazilian Portuguese to foreigners at the UFJF and covers basic and intermediate levels. Since the rules of SMILLE covers those two levels of the CEFR, the non-existence of an advanced level was not a problem for our analysis. We also ignored the levels to which each of the SMILLE's structures are linked in the CEFR and considered only the curriculum of the specific course, as stated in the handbooks. The corpus contains texts used in reading activities, but we excluded texts with gaps used for exercises or texts that explained grammatical content of the language course.

[8] https://oportuguesdobrasil.wordpress.com/musicas-apresentadas-na-sala-de-aula/.

We also excluded lyrics and poetry, since these genres tend to use different punctuation and structure. The texts are skewed to the informative genre, with more emphasis to magazine and history articles, but there are also dialogues, literature, opinion, and general descriptions. The corpus contains 19,741 tokens (8,031 for basic and 11,710 for intermediate) distributed along 957 sentences (421 for basic and 536 for intermediate). It was annotated with all SMILLE structures for Portuguese and the frequencies were then standardized according to the number of sentences per document (in the case of syntactic structures) or tokens per document (for morphological or lexical structures).

For the analysis, we excluded grammatical structures that had less than 80% precision in the evaluation, but we added structures that are directly based on parser information (like simple verb tenses and word classes), since the parser has around 94% of accuracy for part-of-speech tagging and around 85% of accuracy for dependency parsing [21]. We analyzed the corpus in terms of distribution of grammatical structures at each level, looking at the most prominent structures in each of them in terms of significant differences in relative frequency between both levels. We then contrasted the different prominences with the presented grammatical structure of the handbooks to see if they are in consonance in terms of presentation of structures and exposure of learners to the structures.

By running a Mann-Whitney U test, we could see that some structures have a significantly different occurrence ($p < 0.05$) in both corpora. Some grammatical structures, like the present tense overall and, in specific, the present tense of the verb "ter", and the use of personal pronouns as subject, are significantly more prominent in the sentences of the basic level, as expected according to the pedagogical content of the handbook. Looking at the intermediate level, we have structures like the past future and the past imperfect tense as more prominent. According to the dispositions in the handbooks, both these tenses are taught at the end of the basic level and are reviewed with more emphasis during the intermediate level, so their predominance at the review level is understandable.

Probably due to the size of the corpus, some structures that have a different occurrence in both levels did not achieve a significant level of confidence ($p < 0.05$). So, for instance, the present forms of the verb "ser" and all forms of the present of the conjunctive occur more than double in texts from the basic level, where they are indeed emphasized according to the grammar content of the handbooks. On the other hand, structures associated to questions, like interrogative pronouns and simple questions in the present tense were over twice more frequent per sentence at the intermediate level, even though the formation of questions are emphasized as a topic at the basic level. Most of the observed structures did not show a significant difference and occur in a similar way in both basic and the intermediate levels, and this suggests that most of the texts used in the corpus are not there to emphasize grammatical aspects of the language.

This type of analysis can aid teachers and pedagogical coordinators in the task of preparing a Portuguese as foreign language course in a way that the texts can better reflect and emphasize the grammatical content that is being taught. In this analysis, our hypothesis that the texts would match their level were confirmed for the structures for which we had significant differences, but there are clues that some of the content from the intermediate level may be, in fact, too basic. A more in detail evaluation of each text would need to be performed to gauge the full extent to which the texts match their level, but this type of analysis is beyond the scope of this paper.

6 Final Remarks

In this paper, we presented SMILLE for Portuguese, a system that can recognize pedagogically relevant grammatical structures in raw texts. SMILLE covers structures from levels A1 to B2 following a CEFR categorization, corresponding to the basic and intermediate levels. It can be used not only to enhance texts that are to be used with language learners, but it can also be applied in the selection of these texts. As such, a teacher would have a help in selecting more appropriate texts based on their grammatical profile and the grammatical structures that need to be emphasized for the learners.

Since many of the recognized structures use rules on top of the parser annotation, we carried out a precision evaluation in three different genres: newspaper articles, literature and subtitles. Most of the structures scored as high as 100% of precision, such as the comparatives and the present continuous tense, but some of them, such as reflexive pronouns and verbal periphrases, scored much lower (respectively, 44% and 45.3%). Overall, the system achieved an average precision of 84% in the evaluated structures. For the structures that presented bad performance, we saw a mix of bad parsing performance and bad rules, so that we will be addressing these issues for the future versions of the system.

We also presented a case study of how SMILLE can be applied to observe the adequacy of texts used in Portuguese as foreign language courses. For that, we analyzed the texts presented in handbooks of the language course held at the Universidade Federal de Juiz de Fora. This analysis was performed to observe if the texts in each of the two available levels (basic and intermediate) actually emphasize the grammatical content that is described in the handbooks. From the structures that had significantly different use in both levels, we could observe that they do follow the description provided in the handbooks for the basic and the intermediate levels. Nevertheless, for some of the structures, there was a large difference (more than double) in terms of average relative frequency in the texts (but with no significant difference), pointing to a possible mismatch with the level's grammatical content and suggesting that further investigation would be needed to evaluate the adequacy of their distribution in the texts.

SMILLE can detect grammatical structures that are relevant for the learning of Portuguese as a foreign language and it can help analyze texts used in language courses, but it could also be applied to analyze a full profile of, for instance, how learners of Portuguese tend to write their texts in terms of grammatical organization. As future work, we are interested in expanding the corpus of handbook texts and include also learners' texts, to be able to compare how both these instances of language learning behave in terms of grammar.

References

1. Azab, M., Salama, A., Oflazer, K., Shima, H., Araki, J., Mitamura, T.: An english reading tool as a NLP showcase. In: The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations, pp. 5–8. Asian Federation of Natural Language Processing, Nagoya, Japan, October 2013. http://www.aclweb.org/anthology/I13-2002
2. Azab, M., Salama, A., Oflazer, K., Shima, H., Araki, J., Mitamura, T.: An NLP-based reading tool for aiding non-native english readers. Recent Advances in Natural Language Processing, p. 41 (2013)
3. Brown, J., Eskenazi, M.: Retrieval of authentic documents for reader-specific lexical practice. In: InSTIL/ICALL Symposium 2004 (2004)
4. Chinkina, M., Kannan, M., Meurers, D.: Online information retrieval for language learning. In: ACL 2016, p. 7 (2016)
5. Cross, J.: Noticing'in sla: Is it a valid concept. TESL-EJ **6**(3), 1–9 (2002)
6. Doughty, C.: Second language instruction does make a difference. Stud. Second Lang. Acquisition **13**(04), 431–469 (1991)
7. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60 (2014). http://www.aclweb.org/anthology/P/P14/P14-5010
8. Marujo, L., et al.: Porting reap to european portuguese. In: SLaTE, pp. 69–72 (2009)
9. Meurers, D., et al.: Enhancing authentic web pages for language learners. In: Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 10–18. Association for Computational Linguistics (2010)
10. Plonsky, L., Ziegler, N.: The CALL-SLA interface: Insights from a second-order synthesis (2016)
11. Reinders, H.: Towards a definition of intake in second language acquisition (2012)
12. Schmidt, R.: The role of consciousness in second language learning1. Appl. Linguistics **11**(2), 129–158 (1990)
13. Schmidt, R.: Attention, awareness, and individual differences in language learning. Perspect. Indiv. Characteristics Foreign Lang. Educ. **6**, 27 (2012)
14. Simard, D.: Differential effects of textual enhancement formats on intake. System **37**(1), 124–135 (2009)
15. Smith, M.S.: Input enhancement in instructed sla. Stud. Second Lang. Acquisition **15**(02), 165–179 (1993)
16. Smith, M.S., Truscott, J.: Explaining input enhancement: a mogul perspective. Int. Rev. Appl. Linguistics Lang. Teach. **52**(3), 253–281 (2014)
17. Tiedemann, J.: Finding alternative translations in a large corpus of movie subtitle. In: International Conference on Language Resources and Evaluation (2016)
18. Truscott, J.: Noticing in second language acquisition: a critical review. Second Lang. Res. **14**(2), 103–135 (1998)
19. Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., North, B.: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press, Cambridge (2009)
20. Zilio, L., Fairon, C.: Adaptive system for language learning. In: 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT), pp. 47–49. IEEE (2017)

21. Zilio, L., Wilkens, R., Fairon, C.: Passport: a dependency parsing model for portuguese
22. Zilio, L., Wilkens, R., Fairon, C.: Enhancing grammatical structures in web-based texts. In: Proceedings of the 25th EUROCALL, pp. 839–846, Accepted, 2017
23. Zilio, L., Wilkens, R., Fairon, C.: Using NLP for enhancing second language acquisition. In: Proceedings of Recent Advances in Natural Language Processing, pp. 839–846 (2017)

A Corpus Study of Verbal Multiword Expressions in Brazilian Portuguese

Carlos Ramisch[1](✉), Renata Ramisch[2], Leonardo Zilio[3], Aline Villavicencio[4], and Silvio Cordeiro[1]

[1] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
{carlos.ramisch,silvio.cordeiro}@lis-lab.fr
[2] Interinstitutional Center for Computational Linguistics, São Carlos, Brazil
renata.ramisch@gmail.com
[3] Université catholique de Louvain, Louvain-la-Neuve, Belgium
leonardo.zilio@uclouvain.be
[4] University of Essex, Colchester, UK
alinev@gmail.com

Abstract. Verbal multiword expressions (VMWEs) such as *to **make ends meet*** require special attention in NLP and linguistic research, and annotated corpora are valuable resources for studying them. Corpora annotated with VMWEs in several languages, including Brazilian Portuguese, were made freely available in the PARSEME shared task. The goal of this paper is to describe and analyze this corpus in terms of the characteristics of annotated VMWEs in Brazilian Portuguese. First, we summarize and exemplify the criteria used to annotate VMWEs. Then, we analyze their frequency, average length, discontinuities and variability. We further discuss challenging constructions and borderline cases. We believe that this analysis can improve the annotated corpus and its results can be used to develop systems for automatic VMWE identification.

Keywords: Multiword expressions · Annotation · Corpus linguistics

1 Introduction

Multiword expressions (MWEs) are groups of words presenting idiosyncratic characteristics at some level of linguistic processing [1]. Some MWEs function as verb phrases, and are thus referred to as verbal MWEs (VMWEs). Examples in Brazilian Portuguese (PT-BR) include verbal idioms (e.g. ***fazer das tripas coração*** 'make.INF of-the.FEM.PL tripes heart' ⇒ 'to do everything possible'), light-verb constructions (e.g. ***tomar** um **banho*** 'take.INF a shower') and inherently reflexive verbs (e.g. ***queixar-se*** 'complain.INF-self.3' ⇒ 'to complain').

VMWEs have been the focus of much attention, both in linguistics and in natural language processing [1,3,11,15]. From a linguistic point of view, they present restricted variability patterns, licensing phenomena such as passivization, pronominalization of components, reordering, and free PP-movement

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 24–34, 2018.
https://doi.org/10.1007/978-3-319-99722-3_3

depending on the VMWE category [8,10,14,17]. Moreover, verbs (and VMWEs) tend to have rich morphological inflection paradigms, and allow many (but not all) syntactic changes [5,6]. These are often unpredictable [11], making VMWEs challenging to represent in resources and to model in applications.

For the automatic identification of VMWEs, their variability and their potential for discontinuous realizations make them hard to model, especially when put together with non-compositionality and ambiguity [3]. Indeed, VMWEs were the focus of initiatives like the PARSEME shared task[1] [15], whose goal is to foster the development and evaluation of computational tools for VMWE identification. A by-product of this shared task was the release of freely available VMWE-annotated corpora in several languages, including PT-BR.

The goal of this paper is to study the characteristics of VMWEs in PT-BR using the PARSEME corpus. We describe their annotation and analyze their diversity and distribution. A deeper understanding of this complex phenomenon can inspire linguistic models and boost the development of systems to identify them automatically. In Sects. 2 and 3 we briefly discuss the criteria used to annotate VMWEs and the corpus. Our analyses are in Sects. 4 and 5, and we conclude in Sect. 6.

2 Annotation of Verbal Multiword Expressions

Our corpus was annotated according to the multilingual PARSEME guidelines v1.1, not restricted to PT-BR.[2] They define a MWE as a group of words that displays "some degree of orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is considered general grammar rules of a language" [15]. VMWEs are defined as "multiword expressions whose syntactic head in the prototypical form is a verb." VMWEs are annotated using flat annotations, where each token is tagged as being part of a VMWE or not, and where *lexicalized components* are explicitly marked, as these are the obligatory VMWE components. For instance, in *Maria* ***tomou*** *dois* ***banhos*** 'Maria took two showers', only the lexicalized components are shown in bold[3] as the determiner (*dois* 'two') can be replaced or omitted. Below, we summarize the criteria used to identify and categorize VMWEs, focusing on those that are relevant for PT-BR.

Verbal Idioms (VID) present some kind of semantic idiosyncrasy. Tests for semantic idiosyncrasies are hard to formulate, so we use flexibility tests[4] as a proxy to capture semantic idiosyncrasies. Success in *any* of these flexibility tests results in annotation as VID:

[1] Editions 1.0 (2017) and 1.1 (2018): http://multiword.sourceforge.net/sharedtask2018.

[2] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1.

[3] Boldface indicates lexicalized components for all examples throughout this paper.

[4] A *flexibility test* verifies to what extent a change usually allowed by a language's grammar also applies to the candidate to annotate.

1. CRAN: The expression contains a cranberry word[5] e.g. ***foi para as*** *cucuias* 'went to the.FEM.PL cucuias' ⇒ 'went wrong'.
2. LEX: Replacement of a component by related words (e.g. synonyms, hyponyms, hypernyms) leads to ungrammaticality or unexpected meaning change e.g. ***quebrou*** *um* ***galho***/ *#ramo* 'broke a branch/#twig' ⇒ 'helped'.
3. MORPH: At least one of the components of the VMWE presents restricted morphological inflection with respect to general morphology, e.g. ***bateram perna***/*#pernas* 'hit.PST.3PL leg.SG/#legs.PL' ⇒ 'they walked around'.
4. MSYNT: Morpho-syntactic changes lead to ungrammaticality or unexpected meaning change, e.g. *ela* ***eu perdi meu/#teu tempo*** 'I lost.PST.1SG my/#your time' ⇒ 'I wasted my time'.
5. SYNT: Syntactic changes are restricted, e.g. *eu* ***pisei na bola*** 'I stepped on-the ball' ⇒ 'I made a mistake' but not *#a bola na qual eu pisei* 'the ball on which I stepped'.

Light-Verb Constructions (LVC) are VMWEs composed of a light verb v and a noun n referring to an event or state. Their two sub-categories are LVC.full and LVC.cause. For LVCs, the following tests must be applied in the order specified below:

1. N-ABS: the noun n is abstract, e.g. *festa* 'party' and *prioridade* 'priority' in ***faremos*** *uma* **festa** 'we will throw a party' and *ele* ***dá prioridade*** *ao trabalho* 'he gives priority to his work'.
2. N-PRED: the noun n has at least one semantic argument, e.g. *visitas* 'visits' in ***fez visitas*** 'made visits', whose arguments are the visitor and the visitee.
3. N-SUBJ-N-ARG: v's subject is a semantic argument of n, e.g. *Maria* in *Maria* ***tomou banho*** 'Maria took a shower', which is the agent of *banho* 'shower'.
   - If test 3 passes, apply the two tests below:
     4. V-LIGHT: the verb v has light semantics, e.g. *prestar* 'to lend' in ***presta atenção*** 'lends attention' ⇒ 'pays attention'.
     5. V-REDUC: it is possible to omit v and refer to the same event/state, e.g. *o discurso da Maria* 'the speech by Maria' for *Maria* ***fez um discurso*** 'Maria gave a speech'. If this test passes, LVC.full is chosen.
   - If test 3 fails, apply V-SUBJ-N-CAUSE.
     6. V-SUBJ-N-CAUSE: v's subject is an external participant expressing the cause of n, e.g. *ratos* 'rats' in *ratos me* ***dão medo*** 'rats give me fear' ⇒ 'rats scare me'. If this test passes, LVC.cause is chosen.

Inherently Reflexive Verbs (IRV) are composed by a verb and a reflexive clitic, but the clitic does not fulfill one of its usual roles (reflexive, reciprocal, medium-passive, etc.). A verb-clitic combination is annotated as IRV only if one of the tests below passes:

1. INHERENT: the verb never occurs without the reflexive clitic, e.g. ***se queixam*** 'self.3 complain.PRS.3PL' ⇒ 'complain' but not **queixam* and ***me abstenho*** 'self.1SG abstain.PRS.1SG' ⇒ 'I abstain' but not **abstenho.*

[5] A word that does not co-occur with any other word outside the VMWE.

2. DIFF-SENSE: the reflexive and non-reflexive versions do not have the same sense, such as *ele* ***se encontra*** *na cadeia* 'he self.3 meet in prison' ⇒ 'he is in prison' but *#ele me encontra na cadeia* 'he meets me in prison'.
3. DIFF-SUBCAT: the reflexive and non-reflexive versions do not have the same subcategorization frame, e.g. *ela* ***se esqueceu*** de *Maria* 'she self.3 forgot of Maria' ⇒ 'she forgot Maria' but *ela esqueceu Maria* 'she forgot Maria'.

LVC.full VID
(1) Da mesma forma que a imprensa **tem** o **direito** de **tomar posição**, [...]

IRV
(2) [...] ao **se identificar** como policial ele teria dito 'você não sabe com quem

LVC.full
você mexeu', e **efetuou** os **disparos** na vítima [...]

Fig. 1. Two example sentences with highlighted VMWE annotations (UD-train-s7090 and UD-train-s8536). Category labels shown above, lexicalized components in bold.

Table 1. Overall corpus statistics: number of sentences, tokens, annotated VMWEs and categories in the training (train), development (dev) and test portions.

	Sentences	Tokens	VMWEs	VID	LVC.full	LVC.cause	IRV
train	22,017	506,773	4,430	882	2,775	84	689
dev	3,117	68,581	553	130	337	3	83
test	2,770	62,648	553	118	337	7	91
Total	27,904	638,002	5,536	1,130	3,449	94	863

3 VMWE-Annotated Corpus

The corpus used in this paper is freely available at the PARSEME v1.1 repository.[6] It contains texts from two sources: 19,040 sentences coming from the informal Brazilian newspaper *Diário Gaúcho* (DG) [2] and 9,664 sentences coming from the training set of the Universal Dependencies *UD_Portuguese-GSD* v2.1 treebank (UD) [7]. DG contains running text from full documents, whereas UD contains randomly shuffled sentences from the web.

In addition to manual VMWE annotations, the corpus includes lemmas, part-of-speech (POS) tags, morphological features, and syntactic dependencies using the Universal Dependencies tagsets [7]. On the DG part, POS tags and syntactic dependencies were predicted automatically. On both the UD and DG parts, lemmas and morphological features were also predicted automatically. Predictions were made using UDPipe [16] and the CoNLL-2017 shared task model [19].

[6] http://hdl.handle.net/11372/LRT-2842.

Table 2. Top-5 most frequent VMWEs per category, with frequency in parentheses.

LVC.full	VID	IRV	LVC.cause
marcar gol (47)	*fazer parte* (56)	*apresentar-se* (40)	*dar acesso* (7)
ter chance (43)	*ir ao ar* (48)	*tratar-se* (37)	*causar prejuízo* (6)
fazer gol (40)	*entrar em campo* (26)	*encontrar-se* (33)	*dar continuidade* (5)
ter direito (33)	*chamar atenção* (21)	*queixar-se* (31)	*gerar emprego* (4)
ter condição (29)	*ser a vez* (17)	*referir-se* (26)	*dar origem* (4)
correr risco (28)	*ter pela frente* (15)	*esquecer-se* (25)	*colocar em risco* (4)

Table 3. Average and histogram of VMWE length (L), i.e. nb. of lexicalized items, and of gap size (G), i.e. nb. of non-lexicalized items between first/last lexicalized ones.

	Length (L)				Gap size (G)			
	Avg(Stdev)	%L=2	%L=3	%L ≥ 4	Avg(Stdev)	%G = 0	%G = 1	%G ≥ 2
VID	2.90(±1.01)	42.04	34.16	23.81	0.42(±0.80)	66.64	28.32	5.04
LVC	2.05(±0.24)	95.06	4.54	0.40	1.09(±2.03)	40.76	41.18	18.06
IRV	2.00(±0.08)	99.30	0.58	0	0.13(±0.45)	87.72	12.05	0.23
All	2.22(±0.61)	84.90	9.97	5.11	0.80(±1.72)	53.36	34.01	12.63

Figure 1 shows two corpus excerpts. All categories are represented: LVC.full (e.g. ***tem*** *o* ***direito*** 'has the right'), VID (e.g. ***tomar posição*** 'to take position'), and IRV (e.g. ***se identificar*** 'self.3 identify' ⇒ 'to identify oneself (as)'). In the whole corpus, 1 sentence contains 5 VMWEs, 6 sentences contain 4 VMWEs, 42 sentences contain 3 VMWEs, 473 sentences contain 2 VMWEs, 4,435 sentences contain 1 VMWE and 22,947 sentences contain no VMWE annotation at all.

Table 1 contains a summary of the corpus statistics. It contains in total 27,904 sentences and 5,536 annotated VMWEs, yielding an average of about 1 VMWE every 5 sentences. The predominant category is LVC.full, which represents more than 60% of the annotations. Then, VID and IRV represent respectively around 20% and 15% of the annotations. The corpus contains only few instances of LVC.cause, representing less than 2% of the total number of VMWEs. Because of its use in a shared task, the corpus is split into 3 portions: a training set (train), a development set (dev) and a test set.

The annotation of VMWEs was performed by a team of six PT-BR native speakers, including the authors of this paper, using a dedicated annotation platform [18]. The reported inter-annotator agreement between two of the annotators on a sample of 2,000 sentences is $\kappa = 0.771$ for VMWE identification, and $\kappa = 0.964$ for categorization [15].

Table 2 shows the 5 most frequent annotated VMWEs in each category. To extract this list, we have used the lemmas of annotated VMWEs in their canonical order to neutralize alternations (e.g. passive voice, enclitic vs proclitic pronouns). Since the majority of sentences comes from the DG newspaper, many VMWEs are related to topics often published in this newspaper, such as football (e.g. *marcar/fazer gol* 'mark/make goal' ⇒ 'to score a goal') and television (e.g. *ir ao ar* 'go to-the air' ⇒ 'to go on air'). In the remainder of this paper, we analyze the annotated VMWEs in terms of their properties and challenging aspects [3].

4 Characterization of Annotated VMWEs

Length and Discontinuities. Table 3 summarizes the distribution of VMWE length and gap size. Most IRVs have exactly 2 lexicalized ($\mathrm{L}=2$) adjacent ($\mathrm{G}=0$) components. IRVs containing gaps ($\mathrm{G}=1$) simply correspond to the non annotated intervening hyphen in proclitic uses (e.g. ***chama*** - ***se*** 'calls - self') whereas those of length 3 ($\mathrm{L}=3$) or containing gaps larger than 1 ($\mathrm{G}\geq 2$) correspond to annotation or tokenization errors (e.g. ***se auto*** - ***proclamava*** 'self auto - proclaim'). Most LVCs also have exactly 2 lexicalized components ($\mathrm{L}=2$) but some include a lexicalized preposition (e.g. ***submetido a*** *um* ***tratamento*** 'subjected to a treatment'). The majority of LVCs have a gap ($\mathrm{G}=1$) corresponding to a determiner. The distance[7] between the first and last lexicalized components of LVCs ranges from 0 (1,349 cases out of 3,543 LVCs) to 36 (1 case), with 9.20% having a distance of 3 or more intervening tokens (e.g. ***teve*** *há três anos a* ***ideia*** 'had three years ago the idea'). VIDs tend to be longer, with 2.9 tokens in average. The longest annotated VID contains 10 words (***está com a faca e o queijo na mão*** 'is with the knife and the cheese in-the hand' ⇒ 'is in good conditions to carry something out'). Most VIDs are continuous ($\mathrm{G}=0$) but it is not uncommon to include a gap (e.g. ***cai*** *muito* ***bem*** 'falls very well' ⇒ 'comes in very handy').

Overlaps. Overlapping VMWEs are rare but complex to model. Out of the 12,166 tokens belonging to a VMWE, 112 (≈1%) belong to multiple VMWEs simultaneously (overlaps). Among them, 67 are verbs, 27 are nouns and 18 belong to other POS tags. Overlaps are often caused by coordination, e.g. when a light verb is factorized for several predicative nouns (***ter***$_{1,2,3,4}$ ***ensino***$_1$ ***médio***$_1$ *completo,* ***experiência***$_2$ *em vendas, boa* ***comunicação***$_3$ *e* ***disponibilidade***$_4$ 'have completed high school, experience in sales, good communication and availability'). Noun overlaps are often due to coordination (e.g. *se vamos* ***fazer***$_1$ *ou não vamos* ***fazer***$_2$ ***sacrifícios***$_{1,2}$ 'if we will make or we will not make sacrifices') or due to relative clases (e.g. ***cometer***$_1$ *os* ***erros***$_{1,2}$ *que vinha* ***cometendo***$_2$ 'make the errors that he has been making').

Variability. The 5,536 annotated VMWE tokens correspond to 2,126 unique normalized forms, with 1,244 (58.5%) of them occurring only once.[8] This raises

[7] In number of intervening tokens.

[8] The *normalized form* of a VMWE is its sequence of lemmatized lexicalized components in lexicographic order, whereas its *surface form* is the textual sequence [8].

Table 4. Proportion of VMWEs in dev/test corpora also present in the training corpus.

	Unseen	Seen-identical	Seen-variant
dev ⊆ train	144/553 = 26%	180/553 = 33%	229/553 = 41%
test ⊆ train	156/553 = 28%	164/553 = 30%	233/553 = 42%

concerns over the variability of the annotated VMWEs, which could impact the usability of this corpus when building machine learning models to automatically identify VMWEs from incomplete/insufficient annotated data. Table 4 shows the coverage of the dev and test corpora with respect to the training corpus. Around 26–28% of the VMWEs in the dev/test corpora are unseen in the training data. Therefore, models learned on the training corpus will struggle to overcome 70% recall and should probably recur to external VMWE lexicons [4,9]. Among the 72–74% of seen VMWEs, most of them are actually variants, characterized by a normalized form identical to one seen in the training corpus, but with a different surface form. Hence, it is crucial to take morphological and syntactic variability into account when modeling VMWEs, otherwise ≈2/3 of them might be missed.

Ambiguity. Human annotators and automatic VMWE identification systems need to distinguish true VMWE occurrences from literal uses and accidental co-occurrence [13]. Because of the polysemous uses of reflexive clitics in PT-BR, IRVs are quite ambiguous [12]. Examples include *dar-se* (IRV 'to happen' vs. 'to give-self'), and *formar-se* (IRV 'to graduate' vs. passive of 'to form'). This ambiguity is magnified by accidental co-occurrence due to POS-tagging errors, when the homonymous conjunction *se* 'if' is wrongly identified as a reflexive clitic. VIDs are generally less ambiguous, with some interesting examples of true ambiguity such as *fechou a porta, mas se esqueceu de trancá-la* 'closed the door, but forgot to lock it' vs. *duas escolas* ***fecharam as portas*** 'two schools have shut down'.

5 Challenging and Borderline Examples

Challenging LVCs. According to the guidelines, LVCs contain predicative nouns (expressing an event or state, Sect. 2). These nouns are defined as having semantic arguments, that is, the meaning of the noun is only fully specified in the presence of its arguments. During annotation, we have found some challenging predicative constructions such as ***fazer falta*** 'make lack/foul', because they are ambiguous, and it is hard to identify the arguments of the noun. In *Os dois jogadores* ***fazem falta*** *ao time* 'The two players are missed by the team', the event can be rephrased as *a falta dos jogadores ao time* 'the lack of-the players to-the team', indicating that *falta* 'lack' has 2 arguments here, so it is a LVC.full. However, in *O jogador [...]* ***fez*** *uma* ***falta*** *desnecessária* 'The player [...] made an unnecessary foul', the verbless paraphrase *a falta do jogador* 'the player's foul' indicates that *falta* 'foul' only has one argument. Nonetheless this construction is also annotated as LVC.full. To complicate things further, *falta* 'foul' may also be combined with non-light verbs such as *cobrar* 'charge' and *bater* 'hit', where *falta* refers to a *free kick*. Both are annotated as VIDs.

Causative LVCs. The guidelines distinguish full LVCs (LVC.full) from causative ones (LVC.cause). The corpus includes unexpected causative VMWEs, like ***trazer riscos*** 'bring risks' and ***levar à criação*** 'lead to-the creation'. Verbs like *trazer* are unexpected to form causative relations, but this is the fourth most frequent causative verb among the ones we annotated. One of the examples is *A ausência do sexo também* ***traz*** *uma forte* ***angústia*** 'Lack of sex also causes strong anguish' which we annotated as a LVC.cause. Since the LVC category is the most frequent one PT-BR, the specific tests in the guidelines and the mistakes found during pilot annotations helped the annotators to be consistent in annotating challenging cases like the ones exposed in this section.

Challenging IRVs. The guidelines emphasize the difference between true IRVs and free constructions formed by a full verb combined with a reflexive clitic (Sect. 2). While it is relatively easy to identify IRVs that do not exist without the clitic, IRVs that bear a different meaning without the clitic posed some challenges to the annotation team. In particular, verbs like *encontrar-se* 'find-self' can be fully ambiguous in isolated sentences. For instance, in *A banda se encontrava novamente em São Paulo.* 'The band found itself again in São Paulo.' ⇒ 'The band met/was again in São Paulo', it is impossible to know, without access to a larger context, if the members of the band met each other, or if the information is solely that they were there. Another difficult case is *adaptar-se* 'adapt-self' that should not be annotated as IRV according to the provided tests. While the construction **A mãe adapta o filho à escola.* 'The mother adapts the son to-the school' is ungrammatical, the following one is perfectly admissible: *O escritor adapta o livro ao público.* 'The writer adapts the book to-the public'. Since the guidelines do not mention the semantic attributes of the arguments (e.g. +human), this example does not fit the definition of IRVs, even if it could be interesting to annotate it.

Underrepresented Categories: MVC, VPC and IAV. Some VMWE categories described in the guidelines are underrepresented in PT-BR, namely verb-particle constructions (VPC), multi-verb constructions (MVC) and inherently adpositional verbs (IAV). The latter was optional and was not annotated in PT-BR. Only two possible cases of MVC were found in the corpus: *querer dizer* 'want know' ⇒ 'to mean' and *ouvir falar* 'hear talk' ⇒ 'to hear (about)'. Because they are extremely infrequent, both were annotated as VID, with the former being among the top-10 most frequent annotated VIDs. As for VPCs, there is only one (borderline) example of this category, namely *jogar fora* 'throw away' ⇒ 'throw away'. Since it is difficult to prove that *fora* 'away' ⇒ 'away' works as a particle in this case (as opposed to an adverb), and this is the only potential example of VPC in the corpus, it was annotated as VID.

Metaphors. The concept of metaphors was relevant in the context of the PARSEME shared task, due to the fact that verbal metaphors are not always VMWEs. The distinction between these two categories is, as defined in the guidelines, "a relatively unstudied and open question". The guidelines suggest marking debatable examples and discussing them within the community. Given the characteristics of the corpus (newspaper and web texts) metaphors are rare. One of the most remarkable examples is the following: *o consumidor automaticamente pisa no freio e reduz as compras* 'the consumer automatically steps on-the brake and reduces the purchases'. A closer look shows that it is perfectly acceptable to exchange between *freio* 'brake' and *acelerador* 'accelerator' and keep the idea of the metaphor by opposition. Therefore, this possibility of changing the noun indicates that the construction is a regular metaphor, and not a VMWE.

Collocations. The guidelines define collocations as "combinations of words whose idiosyncrasy is purely statistical". While this definition is debated by several authors, the annotated VMWEs follow the definition provided in the guidelines. For instance, *Renata [...] está quase realizando um sonho*. 'Renata [...] is almost fulfilling a dream' could be considered as a collocation or an LVC. The corpus provides evidence that it is only a collocation: the sentence *o presidente eleito [...] admitiu realizar um sonho de seu pai.* 'The president-elect admitted he is fulfilling his father's dream', shows the possibility of someone else fulfilling someone's dream. Furthermore, both verb and noun allow several other arguments, like *realizar um desejo/uma tentativa* 'to make a wish/attempt' and *ter/carregar um sonho* 'to have/carry a dream'. The distinction between collocations and VMWEs requires special attention and linguistic analysis, in order to restrict the annotation only to the target constructions.

6 Conclusions and Perspectives

In this paper we discussed the Brazilian Portuguese PARSEME corpus containing VMWE annotations. We described the annotation guidelines and process, and analyzed the corpus in terms of the diversity and distribution of the annotated expressions, along with their linguistic characterization. This analysis can be used as a basis for refining the annotation protocol to better tailor VMWEs. Moreover, this work can provide a foundation for NLP tasks and applications that target precise modeling of lexical, syntactic and semantic characteristics of these expressions. This includes their automatic identification in corpora, for which syntactic variation and discontinuities in their realization create challenges for current approaches. The application of our findings to enhance the quality of the annotated corpus and to aid the development of automatic VMWE identification methods is part of our goal for future work.

Acknowledgement. We would like to thank Helena Caseli for her participation as an annotator. We would also like to thank the PARSEME shared task organizers, especially Agata Savary and Veronika Vincze. This work was supported by the IC1207 PARSEME COST action (http://www.parseme.eu) and by the PARSEME-FR project (ANR-14-CERA-0001). (http://parsemefr.lif.univ-mrs.fr/)

References

1. Baldwin, T., Kim, S.N.: Multiword expressions. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, 2nd edn, pp. 267–292. CRC Press, Boca Raton (2010)
2. Bocorny Finatto, M.J., Scarton, C.E., Rocha, A., Aluísio, S.M.: Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In: VIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, pp. 30–39. Sociedade Brasileira de Computação, Cuiabá, MT, Brazil (2011)
3. Constant, M., et al.: Multiword expression processing: a survey. Comput. Linguistics **43**(4), 837–892 (2017). https://doi.org/10.1162/COLI_a_00302
4. Constant, M., Nivre, J.: A transition-based system for joint lexical and syntactic analysis. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 161–171. Association for Computational Linguistics, August 2016. http://www.aclweb.org/anthology/P16-1016
5. Fotopoulou, A., Markantonatou, S., Giouli, V.: Encoding MWEs in a conceptual lexicon. In: Proceedings of the 10th Workshop on Multiword Expressions, MWE 2014, pp. 43–47. Association for Computational Linguistics (2014)
6. Nissim, M., Zaninello, A.: Modeling the internal variability of multiword expressions through a pattern-based method. ACM TSLP Special Issue MWEs **10**(2) (2013)
7. Nivre, J., et al.: Universal dependencies v1: A multilingual treebank collection. In: Calzolari, N., et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016, pp. 1659–1666. European Language Resources Association (ELRA), May 2016
8. Pasquer, C.: Expressions polylexicales verbales: étude de la variabilité en corpus. In: Actes de la 18e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (TALN-RÉCITAL 2017) (2017)
9. Riedl, M., Biemann, C.: Impact of MWE resources on multiword recognition. In: Proceedings of the 12th Workshop on Multiword Expressions, MWE 2016, pp. 107–111. Association for Computational Linguistics (2016). http://anthology.aclweb.org/W16-1816
10. Rosén, V., et al.: A survey of multiword expressions in treebanks. In: Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories Conference, December 2015. https://hal.archives-ouvertes.fr/hal-01226001
11. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: a pain in the neck for NLP. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 1–15. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45715-1_1
12. Sanches Duran, M., Scarton, C.E., Aluísio, S.M., Ramisch, C.: Identifying Pronominal Verbs: Towards Automatic Disambiguation of the Clitic 'se' in Portuguese. In: Proceedings of the 9th Workshop on Multiword Expressions, pp. 93–100. Association for Computational Linguistics, Atlanta, June 2013. http://www.aclweb.org/anthology/W13-1014
13. Savary, A., Cordeiro, S.R.: Literal readings of multiword expressions: as scarce as hen's teeth. In: Proceedings of the 16th Workshop on Treebanks and Linguistic Theories (TLT 2016), Prague, Czech Republic (2018)

14. Savary, A., Jacquemin, C.: Reducing information variation in text. In: Renals, S., Grefenstette, G. (eds.) Text- and Speech-Triggered Information Access. LNCS (LNAI), vol. 2705, pp. 145–181. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45115-0_6
15. Savary, A., et al.: The PARSEME Shared Task on automatic identification of verbal multiword expressions. In: Proceedings of the 13th Workshop on Multiword Expressions, MWE 217, pp. 31–47. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/W17-1704, http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-1704.pdf
16. Straka, M., Straková, J.: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 88–99. Association for Computational Linguistics, Vancouver, August 2017
17. Tutin, A.: Comparing morphological and syntactic variations of support verb constructions and verbal full phrasemes in French: a corpus based study. In: PARSEME COST Action. Relieving the Pain in the Neck in Natural Language Processing: 7th Final General Meeting, Dubrovnik, Croatia (2016)
18. van Gompel, M., van der Sloot, K., Reynaert, M., van den Bosch, A.: FoLiA in practice: the infrastructure of a linguistic annotation format, pp. 71–81 (2017). https://doi.org/10.5334/bbi.6
19. Zeman, D., et al.: Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 1–19 Association for Computational Linguistics, Vancouver, Canada, August 2017. http://www.aclweb.org/anthology/K/K17/K17-3001.pdf

Information Extraction

Nominal Coreference Resolution Using Semantic Knowledge

Evandro Fonseca[1(✉)], Aline Vanin[2], and Renata Vieira[1]

[1] PUCRS, Porto Alegre, Brazil
evandro.fonseca@acad.pucrs.br, renata.vieira@pucrs.br
[2] UFCSPA, Porto Alegre, Brazil
aline.vanin@ymail.com

Abstract. Coreference Resolution is a challenging task, considering the required linguistic knowledge and the sophistication of language processing techniques involved. Several other Natural Language Processing tasks may benefit from it, such as named entities recognition, relation extraction between named entities, summarization, sentiment analysis, among others. We propose a process for nominal coreference resolution in Portuguese, based on syntactic-semantic linguistic rules. Such rule models have been efficiently applied in other languages, such as: English, Spanish and Galician. They are useful when we deal with less resourceful languages, since the lack of sample-rich corpora may prevent accurate learning. We combine different levels of linguistic processing, using semantic relations as support, in order to infer referential relations between mentions. The proposed approach is the first model for Portuguese coreference resolution which uses semantic knowledge.

Keywords: Coreference resolution · Information extraction
Semantics

1 Introduction

Coreference Resolution is a process that consists in identifying the different mentions made to a specific entity in a discourse. In the example: "France is resisting. The country is one of the first in the ranking (...)". The noun phrases [the country] and [France] are considered coreferent. By grouping the terms that refer to the same entity, we form coreference chains. The coreference resolution task has received a great deal of attention from the computational linguistics community. There is a variety of models that solve coreferences for one or for multiple languages.

Currently, many papers concentrate their efforts on machine-learning techniques. Among them, Soon et al. [12], one of the pioneers in this type of approach, proposes a model based on supervised learning for English. However, when we deal with machine-learning techniques, the results depend not just on an adequate set of features, but on the quality and quantity of training samples.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 37–45, 2018.
https://doi.org/10.1007/978-3-319-99722-3_4

Although the amount of resources for Portuguese has been increasing in recent times there is still lack of annotated corpora, rich in coreference samples, to train efficient models. In addition, when we consider semantics, this lack is even greater, since the amount of samples reduces drastically. As an example, we can cite the two main corpora for English and for Portuguese: respectively, there are 34290 chains in the Ontonotes corpus [9] and 3898 chains for the Corref-PT [2] corpus. Considering these facts, machine learning approaches may not be the best option to build coreference models when there is not enough training data. For this scenario, we argue that rule-based approaches can provide better results. However, most of rule-based approaches explore just lexical and syntactic knowledge. Such knowledge is certainly indispensable and widely used in the task, as in "Former president Barack Obama said the U.S. owes him a "debt of gratitude" for his leadership. Obama was the first...". Between [*Former president Barack Obama*] and [*Obama*] there is a referential relation that can be identified by lexical similarity. In addition, lexical and syntactic processing goes beyond matching patterns. Consider the follow sentence: "Today I will be at the University of São Paulo and tomorrow at the University of Paraná...". Note that there is a lexical similarity in the mentions, which is evoked by the term "University". However, the entities "University if São Paulo" and "University of Paraná" are distinct and, therefore, non-coreferent. For cases like this we can use juxtaposition techniques to identify modifiers, like Lee et al. [8] propose. However, there are cases in which correference may not apply even though there is lexical similarity: "*Adalberto Portugal has informed that it is possible. Portugal is the first...*". It could be the case that while the first refers to a person, the second refers to the country. In other cases the similarity is at the semantic level:"*How do bees make honey? The process begins when the insects go hunting ...*". Note that we have no string similarity evidence to establish a coreference relation between the noun phrases [*bees*] and [*the insects*]. Here it is possible to notice the importance of semantic knowledge for the coreference resolution task.

2 Related Work

Currently few papers deal with semantic approaches for coreference resolution. Rahman et al. [10] evaluated the utility of world knowledge using a mention-pair and cluster-ranking model. For world knowledge, the authors used two knowledge bases: Yago and FrameNet. Their strategy consists in identifying relations like "Means" and"IS-A". Each relation is represented in YAGO as a triple. (AlbertEinstein, IS-A, physicist), for instance, denotes the fact that Albert Einstein is a physicist. The relation "Means" provides different ways of expressing an entity, and therefore it allows dealing with synonymy and ambiguity, i.e. for the two triples: (Einstein, Means, AlbertEinstein), and (Einstein, Means, AlfredEinstein) denotes the fact that Einstein may refer to the physicist Albert Einstein or the musician Alfred Einstein. From FrameNet, the authors used semantic role related to verbs.

Hou et al. [7] propose a rule based system to solve anaphora and bridging. Different from our work, which tries to identify coreference (identity relation), bridging resolution consists in recognizing non identity links. An example is the meronymy relation ("Part_of") as in [the house] [the chimney]. To identify this type of relation, the authors used WordNet[1]. Garcia et al. [5] propose Link-People: a model for coreference resolution which is tailored to person entities (which may be considered as an initial semantic orientation). They considered three languages: Portuguese, Spanish and Galician. Their model combines the multi-pass architecture and a set of constraints and rules. The authors use some matching rules from [8]; in addition, they use a set of specific rules to dealing with pronouns, anaphora, cataphora for person entities. In an error analysis, the authors mention the problem of lack of rich semantic resources, showing that their model could be improved by detecting semantic relations like synonymy, hyponymy and hyperonymy: [the boy] and [the youngster]. For Portuguese coreference, Garcia et al. built their own corpus [6] considering only entities of type person.

For Brazilian Portuguese, Silva [11] proposes a coreference resolution system based in the same Harem[2] semantic categories, using an unsupervised learning algorithm. Regarding semantic processing, the author uses synonymy relation based on Tep2.0[3], a thesaurus containing synonymy and antonymy for Portuguese. Silva reports that the semantic knowledge did not show improvements in his experiments. However, he considered a small corpus, containing just nine texts, which may be considered a limitation.

We see that the previous work combines the use of named entity categories and semantic knowledge resources. However, the only Portuguese semantic resource considered for this task was Tep2.0, which contains 8.528 synonym and antonym relations. There are more comprehensive semantic databases currently available. Onto-PT contains 168.858 synonymy relations, 91.466 hyperonymy/hyponymy, 9.436 meronymy and 92.598 antonymy relations.

3 Proposed Model

In this section, we describe the process for automatic coreference resolution in Portuguese. An overview of the proposed process is illustrated in Fig. 1.

Initially, we extract the noun phrases and its respective attributes using the parser CoGrOO[4], followed by pre-processing, which remove noun phrases that start with numerical entities such as percentage, money, cardinal and quantifiers (9%, $10.000, ten, thousand, 100 meters). Although there is numeric correference, it has a low occurrence and it requires a different processing. Therefore, we chose not to treat it. After the two first steps, we apply our set of rules. To identify semantic relations (Synonymy and Hyponymy), we use Onto.PT[5] and, to

[1] https://wordnet.princeton.edu/.

[2] https://www.linguateca.pt/HAREM/.

[3] http://www.nilc.icmc.usp.br/tep2/.

[4] http://cogroo.sourceforge.net/.

[5] http://ontopt.dei.uc.pt/.

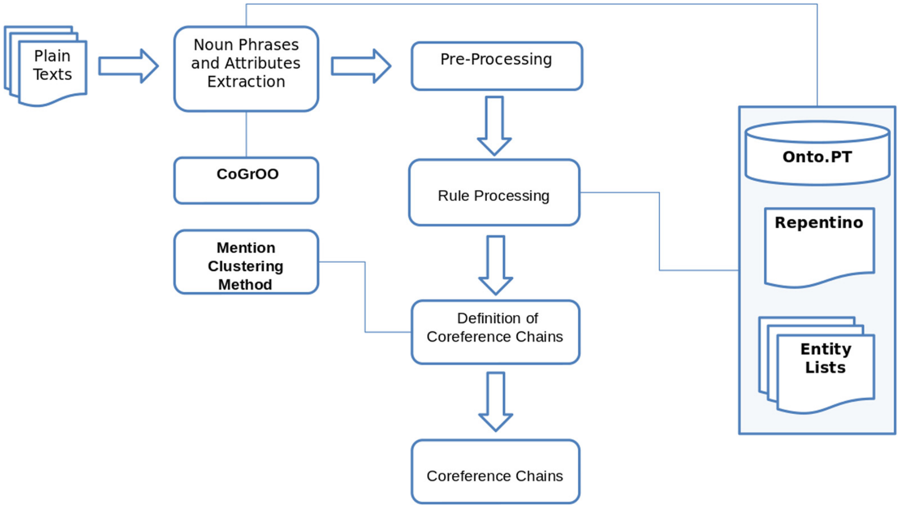

Plain Texts
Noun Phrases and Attributes Extraction
CoGrOO
Mention Clustering Method
Pre-Processing
Rule Processing
Definition of Coreference Chains
Coreference Chains
Onto.PT
Repentino
Entity Lists

Fig. 1. Proposed model

define the entity category of coreference chains we use Repentino[6] and entity lists [3], containing common and proper nouns, such as: lawyer, agronomist, street, avenue, João, Daiane, among others. To infer the chain category, we choose the most frequent semantic class, considering all noun phrases of the chain. In rule-based approaches, each step consists in applying some filter/rules, aiming to group two mentions m_x and m_y, if one or more rule is satisfied. In our approach, we use a graph-like structure to store rule processing among the mentions. Later, we use our clustering method, which aims to identify whether a mention is anaphoric or new in discourse.

Regarding our proposed rules, they are also found in works for English [8,10,12]. However, it was necessary to consider specific linguistic rules for Portuguese. Many of our rules have been adapted from the literature, considering the particular aspects of Portuguese and the limitations of the resources available for this language. However, few works, even for English, address the use of semantic rules/features for coreference resolution, such as hyponymy and synonymy. In Table 1 we show our set of rules. They are better described in [4].

4 CORP

A tool for coreference resolution in Portuguese, CORP, was built on the basis of the proposed model. The tool solves coreference for plain texts given as input. CORP is available in two versions: Desktop[7] and WebDemo[8]. The tool generates

[6] https://www.linguateca.pt/REPENTINO/.

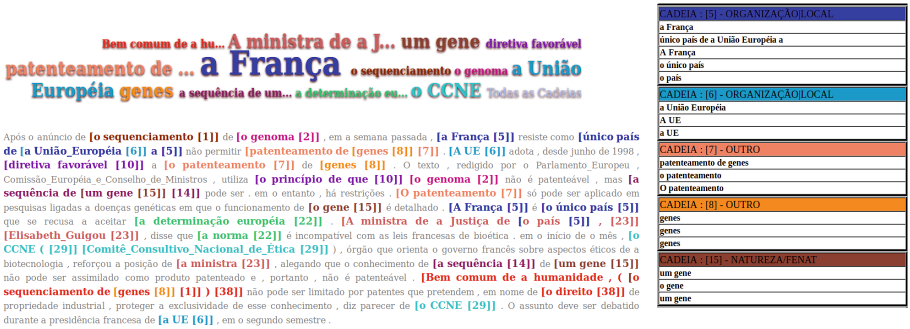
[7] http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/corp-coreference-resolution-for-portuguese/.

[8] http://ontolp.inf.pucrs.br/corref/.

Table 1. Lexical, syntactic and syntactic-semantic rules

Rule	Description	Example
CasamentoDe PadroesExato (R1)	Returns true if the mentions are equal	[Miguel Guerra] [Miguel Guerra]
CasamentoParcial PeloNucleo (R2)	Returns true if the strings up to their heads are equals	[o museu de Porto Alegre] [o museu]
Aposto Especificativo (R3)	Returns true if: for given two neighbor mentions, the current NP is a proper name without determinant AND the antecedent is a noun with a determinant	[o telescópio] [Gemini]
Aposto Explicativo (R4)	Returns true when two mentions are in appositive construction	[a ministra da justiça], [Elisabeth Guigou], ...
Acronimo (R5)	Returns true if a mention is acronym of the other	[a União Européia] [a UE]
PredicadoNominativo (R6)	Returns true when two mentions is in copulative subject-object relation	[A França] é [um país]
PronomeRelativo (R7)	Returns true when there are two adjacent mentions and the second NP is a relative pronoun	[A sonda WMAP] [cuja] missão
CasamentoRestrito PeloNucleo1 (R8)	Returns true if any of their head words match AND does not has modifier terms (nouns, proper nouns, verbs, adjectives and adverbs) AND the mentions are not embedded	[o Comitê Nacional de Ética] [o comitê]
CasamentoRestrito PeloNucleo2 (R9)	Returns true if any of their head words match AND does not has modifier terms (in this rule just nouns and adjectives are considered modifiers) AND the mentions are not embedded AND R8=false	[a estrada que ficará pronta] [a estrada que talvez fique pronta]
CasamentoEntre NomesProprios (R10)	Returns true if: both mentions contain proper nouns AND some of their words are equal AND are not embedded	[a Califórnia] [a adorável Califórnia]
CasamentoParcial EntreNomesProprios (R11)	Returns true if both mentions contains proper names AND at least one word of m_j is equal to a word of m_i AND there are no modifier terms (same clause of R8).	[a União Européia] [a União]
Hiponímia (R12)	Returns true if the head lemmas presents a Hyponymy relation	[as abelhas] [os insetos]
Sinonímia (R13)	Returns true if the head lemmas presents a Synonymy relation	[o menino] [o garoto]

an annotated version for the input text, in XML files. A visualization of the coreference chains that are generated is given, as shown in Fig. 2. We provide a table containing all coreference chains besides the text, in which the chain elements are highlighted in different colours. Note that there are some embedded mentions, such as "a União Européia" in"único país da União Européia". In those cases, we present the larger expression in the same colour and we use a different color only for the brackets and the ID of the inside mention. The chains can also be highlighted individually in the text by clicking in the corresponding item in the tag cloud. The tag clouds also work as a text summary.



Após o anúncio de [o sequenciamento [1]] de [o genoma [2]] , em a semana passada , [a França [5]] resiste como [único país de [a União_Européia [6]] a [5]] não permitir [patenteamento de [genes [8]] [7]] . [A UE [6]] adota , desde junho de 1998 , [diretiva favorável [10]] a [o patenteamento [7]] de [genes [8]] . O texto , redigido por o Parlamento_Europeu , Comissão_Européia_e_Conselho_de_Ministros , utiliza [o princípio de que [10]] [o genoma [2]] não é patenteável , mas [a sequência de [um gene [15]] [14]] pode ser . em o entanto , há restrições . [O patenteamento [7]] só pode ser aplicado em pesquisas ligadas a doenças genéticas em que o funcionamento de [o gene [15]] é detalhado . [A França [5]] é [o único país [5]] que se recusa a aceitar [a determinação européia [22]] . [A ministra de a Justiça de [o país [5]] , [23]] [Elisabeth_Guigou [23]] , disse que [a norma [22]] é incompatível com as leis francesas de bioética . em o início de o mês , [o CCNE ([29]] [Comitê_Consultivo_Nacional_de_Ética [29]]) , órgão que orienta o governo francês sobre aspectos éticos de a biotecnologia , reforçou a posição de [a ministra [23]] , alegando que o conhecimento de [a sequência [14]] de [um gene [15]] não pode ser assimilado como produto patenteado e , portanto , não é patenteável . [Bem comum de a humanidade , ([o sequenciamento de [genes [8]] [1]]) [38]] não pode ser limitado por patentes que pretendem , em nome de [o direito [38]] de propriedade industrial , proteger a exclusividade de esse conhecimento , diz parecer de [o CCNE [29]] . O assunto deve ser debatido durante a presidência francesa de [a UE [6]] , em o segundo semestre .

CADEIA : [5] - ORGANIZAÇÃO|LOCAL
a França
único país de a União Européia a
A França
o único país
o país
CADEIA : [6] - ORGANIZAÇÃO|LOCAL
a União Européia
A UE
a UE
CADEIA : [7] - OUTRO
patenteamento de genes
o patenteamento
O patenteamento
CADEIA : [8] - OUTRO
genes
genes
genes
CADEIA : [15] - NATUREZA/FENAT
um gene
o gene
um gene

Fig. 2. CORP - HTML output

5 Evaluation

We evaluate our model using three Portuguese corpora: Summ-it++[1][9], Corref-PT [2][10]. and Garcia et al.'s [6] corpus. In Table 2, we show results of the main related works, evaluated using the CoNLL metrics conference. Unfortunately, a comparison among these models is not possible, due to the distinct languages, corpora and scopes. It is possible to compare the evaluations performed considering our model, Summ-it++ and Corref-PT corpora and the employment of semantics. When we compare these results, it is possible to see that our precision decreases using semantics (13.7% for Summ-it++ and 9.3% for Corref-PT). However, there are a significant improvement in recall (7.7% for summ-it++ and 2.4% for Corref-PT, both cases considering MUC metric). If we analyze the F-measure and CoNLL metric, the model is worse, but we must consider that the use of semantic knowledge helps to identify new relations, such as: (fungus, small mushrooms), (scientists, researchers),(France, the country – as we can see in Fig. 2).

[9] Summ-it++ is available for download in http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/summ-it/.

[10] Corref-PT is available for download in http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/corref-pt/.

Table 2. Non-comparative results of our and main related works using CoNLL metrics

Model		Language	MUC			B^3			CEAF$_e$			CoNLL
			P	R	F	P	R	F	P	R	F	F
Martschat et al., 2015		EN	76.8	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Fernandes et al., 2014		EN	75.9	65.8	70.5	77.7	65.8	71.2	43.2	55.0	48.4	63.4
		CH	71.5	59.2	64.8	80.5	67.2	73.2	45.2	57.5	50.6	62.9
		AR	49.7	43.6	46.5	72.2	62.7	67.1	46.1	52.5	49.1	54.2
Lee et al., 2013		EN	60.9	59.6	60.3	73.3	68.6	70.9	46.2	47.5	46.9	59.4
Garcia et al., 2014		ES	94.1	84.1	88.8	84.8	62.9	72.2	71.0	83.4	76.7	79.2
		GL	94.6	89.0	91.7	88.4	72.9	79.9	76.6	87.6	81.7	84.4
		PT	92.7	82.7	87.4	84.5	65.8	74.0	67.9	84.4	75.2	78.9
Our model (Summ-it++)	without semantics	PT	58.8	44.4	50.6	59.3	41.7	49.0	53.7	54.2	54.0	51.2
	with semantics		45.1	52.1	48.3	43.8	49.5	46.5	45.7	57.4	50.9	48.6
Our model (Corref-PT)	without semantics		64.2	47.8	54.8	61.2	40.5	48.7	50.2	51.0	50.6	51.4
	with semantics		54.9	50.2	52.5	51.8	43.6	47.3	46.2	52.8	49.3	49.7

Even considering the scope differences of ours and Garcia et al.'s model, we performed a comparative evaluation (Table 3) involving two texts, written in Portuguese, belonging to Garcia et al.'s corpus [6], which annotates only person entities.

We believe that to perform a fair evaluation, we should involve Summ-it++ and Corref-PT. The Summ-it++ corpus has 5047 mentions (just 226 categorized as person); Corref-PT has 33514 mentions (just 3119 categorized as person). In few words, even though Garcia et al.'s model groups correctly all person entities, their model will cover just 4.5% of Summ-it++ and 9.3% of Corref-PT. Our model covers more than 50%. However, different from our model, which generates coreference from plain texts; Garcia et al.'s model requires a series of previously annotated information to obtain the coreference chains, such as: part-of-speech tagging, dependency parsing, named entity category and mention detection. Due to the complexity of this analysis, we consider it as future work.

In Table 3 we see that our model, as expected, presented a lower recall. This is due to the fact that we do not treat pronominal coreference, which is the focus of Garcia et al.'s model and corpus. Despite of that, our model has obtained high precision values for MUC and B-Cubed metrics (80% and 81.7%, respectively); for BLANC metric, our model outperformed Garcia et al.'s, due to the fact that BLANC considers the average of coreferential and non-coreferential links. In addition, our model has correctly identified two coreference chains that cannot be identified by Garcia et al.'s model (one chain containing four mentions "Org/Loc" and other containing two mentions type"Other"); these coreference chains were not annotated in Garcia's corpus, since these chains do not refer to a Person. Due to this scope these chains would count in our favor, but they are not considered in evaluation.

Table 3. Comparative analysis between our model and Garcia et al.

Model	MUC			B^3			CEAF$_m$			CEAF$_e$			BLANC			CONLL
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	F
Garcia et al., 2014	97.9	96.0	97.0	86.44	81.9	84.1	86.4	86.4	86.4	85.8	95.3	90.3	83.1	83.4	83.1	90.5
Our model	80.0	16.0	26.7	81.7	7.8	14.2	73.3	18.6	29.7	33.4	18.6	23.9	90.7	83.7	86.0	21.6

6 Conclusion and Future Work

This paper presented our study about coreference resolution and semantics. As a side contribution we provide a tool that runs over plain texts, generating correference annotation in XML and also a visualization of the generated chains. Even though the semantic model has presented lower precision and f-measure, it introduces gains in recall due to a new processing level. As future work, we aim to test our model using other semantic bases, like ConceptNet[11] and BabelNet[12]. We consider also that the semantic rules must be improved, for instance, by dealing with ambiguity, to increase our model precision.

Acknowledgments. The authors acknowledge the financial support of CNPq, CAPES and Fapergs.

References

1. Antonitsch, A., Figueira, A., Amaral, D., Fonseca, E., Vieira, R., Collovini, S.: Summ-it++: an enriched version of the summ-it corpus. In: Proceedings of 10th edition of the Language Resources and Evaluation Conference, Portorož, Slovenia (2016)
2. Fonseca, E., Sesti, V., Collovini, S., Vieira, R., Leal, A.L., Quaresma, P.: Collective elaboration of a coreference annotated corpus for portuguese texts. In: Proceedings of II Workshop on Evaluation of Human Language Technologies for Iberian Languages, vol. 1881, pp. 68–82, Murcia, Spain (2017)
3. Fonseca, E.B.: Resolução de correferências em língua portuguesa: pessoa, local e organização, Dissertação de Mestrado, Pontifícia Universidade Católica do Rio Grande do Sul (2014)
4. Fonseca, E.B., Sesti, V., Antonitsch, A., Vanin, A.A., Vieira, R.: Corp - uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências. Linguamatica **9**(1), 3–18 (2017)
5. Garcia, M., Gamallo, P.: An entity-centric coreference resolution system for person entities with rich linguistic information. In: Proceedings of 25th International Conference on Computational Linguistics, pp. 741–752, Dublin, Ireland (2014)
6. Garcia, M., Gamallo, P.: Multilingual corpora with coreferential annotation of person entities. In: Proceedings of the 9th Edition of the Language Resources and Evaluation Conference, pp. 3229–3233, Reykjavik, Iceland (2014)

[11] http://conceptnet.io/.
[12] https://babelnet.org/.

7. Hou, Y., Markert, K., Strube, M.: A rule-based system for unrestricted bridging resolution: recognizing bridging anaphora and finding links to antecedents. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 2082–2093, Doha, Qatar (2014)
8. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. Comput. Linguistics **39**(4), 885–916 (2013)
9. Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N.: Conll-2011 shared task: modeling unrestricted coreference in ontonotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–27, Portland, Oregon (2011)
10. Rahman, A., Ng, V.: Coreference resolution with world knowledge. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 814–824, Portland, Oregon, USA (2011)
11. da Silva, J.F.: Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado, Dissertação de Mestrado, Universidade de São Paulo (2011)
12. Soon, W.M., Ng, H.T., Lim, C.Y.: A machine learning approach to coreference resolution of noun phrases. Comput. Linguistics **27**(4), 521–544 (2001)

Pragmatic Information Extraction in Brazilian Portuguese Documents

Cleiton Fernando Lima Sena and Daniela Barreiro Claro(✉)

Formalisms and Semantic Applications Research Group (FORMAS),
LaSiD/DCC/IME – Federal University of Bahia (UFBA), Av. Adhemar de Barros,
s/n, Campus de Ondina, Salvador, Bahia, Brazil
cflsena2@gmail.com, dclaro@ufba.br

Abstract. The volume of published data in the Web has been increasing, and a great amount of those data is available in a natural language format. Manually analyzing each document is a time-consuming and tedious task. Thus, Open IE area emerges to help the extraction of semantic relationships in a large number of texts written in a natural language from different domains. Although a semantic analysis does not guarantee complete accuracy in extracting relations, a pragmatic analysis becomes important on Open EI to identify additional meanings (unsaid) that goes beyond semantics in a text. Our work developed a method for Open Information Extraction to extract relations from texts written in Portuguese in a first pragmatic level. We stated that a first pragmatic level deals with inferential, contextual and intentional aspects. We evaluate our approach, and our results outstand the most relevant related work on comparing accuracy and minimality measures.

Keywords: Open information extraction · Inference · Context
Intention · Pragmatic

1 Introduction

An increasing amount of digital data has been published over the world. Considering the Web, the rise and popularity of the Internet have spawned a vast collection of heterogeneous data. A significant amount of this data is published in a natural language format. A manual analysis of such amount of data to retrieve relevant information might be a time-consuming task. In this way, the Information Extraction (IE), also called traditional IE [14], emerged to identify useful patterns, automatically, in textual documents. This task had an initial focus on a small, homogeneous and well-known domain. However, some problems have been identified, such as the low extraction of facts in texts with different domains and the necessity of human intervention to extract new facts [1]. A new approach comes up called Open Information Extraction (Open IE) [8] to overcome some difficulties, especially those related to different domains. Open IE aims to extract

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 46–56, 2018.
https://doi.org/10.1007/978-3-319-99722-3_5

semantic information in texts from different domains in the form of verbs (or verbal phrases) and their arguments [11].

Open IE is characterized by not being limited to previously facts [8]. A fact through Open IE is composed of a relation between a pair of entities [1] defined as a triple $t = (arg_1, rel, arg_2)$. Where rel corresponds to the relation between the arguments (arg_1, arg_2). For example, in the sentence *The card is in the drawer.*, the triple (*The card, is in, the drawer*), would be extracted without the need to specify the relation *is in* or its arguments *The card* and *the drawer*. Three main strengths of Open IE can be regarded: (i) domain independence; (ii) greater coverage of facts; (iii) scalable for the Web [7]. On the other hand, Open IE has an important drawback: the low accuracy of extracted facts.

Open IE when compared to traditional IE produces a lot of invalid facts [2]. A fact is invalid when the information extracted is not consistent with the information described in the text [8]. For example, in the sentence *Peter, friend of Mary, traveled out of town* the fact (*Mary, traveled out of, town*) is considered invalid because the information that *Mary traveled* is not present in the sentence. The fact might be (*Peter, traveled out of, town*). This type of error occurs due to the difficulty of Open IE approaches such as TextRunner [1] and Reverb [8] on dealing with human written style.

It is also worth mentioning that it has been a challenge for Open IE approaches to extract facts that are implicit in texts. Aspects such as inference, context, and intention, which go beyond the meaning of the words, can influence the extraction of implicit facts or the sense of the fact, changing their meaning. For example, in the sentence *Norisring is a street circuit located in Nuremberg*, semantic methods without an inferential approach might not extract the information #1 (*Norisring, located in, Nuremberg*), by a transitive inference. Authors in [4] expose such kind of problem. They cope only English language and a limited version of transitive rule. As a consequence, the single fact #1 is extracted. The method proposed in [13], which is a Portuguese approach, used both types of inference: rules of transitivity and symmetry. Considering the same sentence, the method only extracts the same fact #1 due to a limitation: in a sentence, only a rule by time is extracted, e.g., a transitive or a symmetric fact. Moreover, the method proposed in [13] is restricted to few patterns of transitivity and symmetry.

Regarding contextual information, the method proposed by [11] was one of the few papers which deals with context in a sentence. For example, taking the sentence *Romulo will travel out of the country, when the value of the dollar falls*, their approach extracts the information *(Romulo, will travel out of, the country)*. This information is inconsistent as it is conditioned to the value of the dollar.

Considering the intentional information, to the best of our knowledge, no Open IE method has addressed so far the intentional paradigm. Taking the sentence as an example *Unfortunately, the exam score has not yet been published by the teacher*, a possible extraction by intention could be (*the exam score, should be published by, the teacher*).

Thus, we propose an Open IE method for the Portuguese language which copes with inferential, contextual and intentional aspects thus reaching a first pragmatic level. This method can improve both the quality and the number of valid facts extracted, due to a more careful analysis of the sentence, i.e. regarding particular cases of the Portuguese language. Our inferential layer was increased by the number of transitivity and symmetry rules and the use of general rules to infer new implicit facts. Both transitive and symmetric types of inference are extracted from a single sentence by applying rules that can generalize the transitive and symmetric patterns. Our hypothesis states that extracting implicit facts with our inferential method can increase the number of extractions in a single sentence. Our contextual layer enhanced the method proposed by [11]. We broaden their approach by the use of subordinate conjunctions, adverbs, prepositions, and adversative coordination sentences. Our hypothesis states that including some morphosyntactical aspects might improve the informativeness and the quality of extracted facts. Finally, our intentional layer can extract implicit facts in a sentence, through verbs in the Future of the Preterite (Portuguese Grammar) or Conditional Tense (English Grammar). We state that the use of a specific grammar case can improve the number of extracted facts. Thus, we evaluated our approach, and our results outstand the most relevant related work on comparing accuracy and minimality measures.

The remainder of this paper is organized as follows: Sect. 2 presents some related works and Sect. 3 describes the Pragmatic in Open IE area. Section 4 explain our method and Sect. 5 describes the experimental setup. In Sect. 6 we present our results and we conclude with some discussions in Sect. 9.

2 Related Works

The startup of Open IE area was made by the TextRunner [1] system, which uses a self-supervised approach to training its positive and negative examples in English. A classifier is trained using these samples to extract facts. New other systems emerged, such as WOE [14] that uses self-supervised learning to create their examples in the training phase.

The second generation of Open IE systems has left the learning stage of patterns to express relationships. ReVerb [8] is the first approach that manages syntactic and lexical constraints to extract arguments and relations expressed by verbs in English sentences.

A new wave generates new methods by the use of a Dependency Parser (DP) between the morphological classes of words and a set of rules. These approaches detect useful parts (clauses) in a sentence. One of the most cited work is ArgOE [9], which supports English, Spanish, Portuguese, and Galician languages.

From a Portuguese language perspective, there is an inferential method proposed by [13] (from now on called SGC_2017) which uses a morphosyntactic approach. Another method that copes with the Portuguese language is DependentIE [12] which uses a Dependency analyzer for Portuguese.

Despite the fact that Open IE is still producing a significant number of invalid facts from multi-domain and large-scale texts, there is a considerable loss of facts that are implicit in a document. Such kind of facts has not yet been extracted. The inferential, contextual and intentional approaches remain an open challenge for the Open IE area. We state that the combination of these approaches can achieve the first level of pragmatic information which can be extracted from a sentence. Performing these approaches might increase the number of useful extracted facts.

3 Pragmatic in Open IE

In some domains, i.e., speech and writing, the pragmatic can act in the understanding of discourses from extralinguistic factors, as well as in the interpretation of signs and their meanings. Pragmatic understands a language beyond the semantic level [10]. In this way, interpretation is the object of Pragmatics, and these meanings are "a special kind of intention recognized by the receiver", dealing with the Grice Implicatures Theory [10].

Grice Implicatures Theory aims to determine the meaning of what goes beyond what has been said [10]. Thus, it identifies a meaning beyond the semantic meaning. It is worth noting that Grice's Theory gets this additional meaning as an "unsaid" meaning, which can be inferred using a proposition with a semantic value [6]. In turn, Grice Implicatures Theory checks essential features so that an individual can convey implicit information in sentences [6]. For example, considering the sentence *Renato left the door open* can be interpreted as *Renato should close the door*. Grice's theory divide the implicatures into two basic types: conventional and conversational [10].

Conventional implicatures connect the semantic meaning of the words [6]. An extra unsaid meaning is transmitted, and it does not affect the meaning of what has been said [10]. For example, regarding the sentence *Carlos is a politician, but he does not tell lies.*, it can be inferred, through conventional implicatures, that politicians tell lies.

Conversational implicature is divided into two categories: generalized and particularized [10]. The generalized implicatures do not rely on a context to identify extra meanings. For example, taking the sentence *Theresa gave gifts to a baby yesterday.*, it can be inferred that this baby is not Teresa's son since there is no additional information that may contradict this implication [6]. Particularized implicatures are related to specific contexts, which the extra meaning transmitted depends on the condition of the information in a message [6]. For instance, regarding the sentence *Renato is happy because he is working now.*, it is possible to imply that Renato was unemployed or Renato was sad to be unemployed.

From the implications of Grice, we apply the concepts of conventional implicatures and conversational particularized implicatures in the inferential, contextual and intentional approaches, to achieve a first level of a pragmatic extraction.

4 Our PragmaticOIE Method

PragmaticOIE starts by preprocessing the sentences through POS tagger and NP chunker analyzers (Fig. 1).

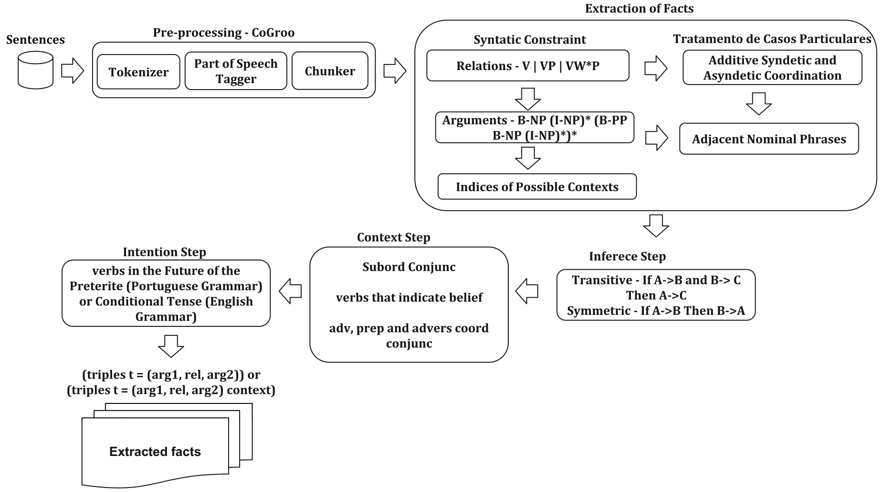

Sentences
Pre-processing - CoGroo
Tokenizer
Part of Speech Tagger
Chunker
Extraction of Facts
Syntatic Constraint
Relations - V | VP | VW*P
Arguments - B-NP (I-NP)* (B-PP B-NP (I-NP)*)*
Indices of Possible Contexts
Tratamento de Casos Particulares
Additive Syndetic and Asyndetic Coordination
Adjacent Nominal Phrases
Inferece Step
Transitive - If A->B and B-> C Then A->C
Symmetric - If A->B Then B->A
Context Step
Subord Conjunc
verbs that indicate belief
adv, prep and advers coord conjunc
Intention Step
verbs in the Future of the Preterite (Portuguese Grammar) or Conditional Tense (English Grammar)
(triples t = (arg1, rel, arg2)) or
(triples t = (arg1, rel, arg2) context)
Extracted facts

Fig. 1. PragmaticOIE architecture.

Relations are first identified through the verb-based syntactic constraints (adapted from ReVerb [8]), then particular treatments are applied to identify new relations mediated by syndetic additive coordination sentences and asyndetic coordination sentences. The arguments for each relationship are extracted based on a syntactic constraint proposed in [13]. A particular treatment is applied to extract more arguments that are both adjacent to the first argument and part of the same relation. Indexes of a context, which can be positioned either left or right of an argument, are calculated. Extracted facts (triples in the format $t = (arg1, rel, arg2)$) of a sentence and indexes of possible contexts related to each extracted fact are combined.

Afterward, the inference layer takes place. The extracted facts are submitted to transitive and symmetrical rules. New facts (through transitivity or symmetry) are inferred if any pattern is found. Holding the context layer, indexes are calculated, and some verbs, which give the sense of belief and adverbs, prepositions and adversative coordination are identified. To progress in the extraction, these extracted facts are treated by the intention layer. This final approach holds verbs in the Future of Preterite (Portuguese Grammar) or Conditional Tense (English Grammar). New facts are inferred, finishing the execution flow of our method (Fig. 1).

5 Experimental Setup

We evaluated our method PragmaticOIE into five datasets with random sentences: INFER-100 (inferential - 100 sentences), CONTEXT-100 (contextual - 100 sentences), INTENT-50 (intentional - 50 sentences), WIKI-200 (multi features - 200 sentences) e CETEN-200 (multi features - 200 sentences). These datasets were created within two data sources: Portuguese Wikipedia[1] and CETENFolha (Corpus of Electronic Texts Extracts NILC/Folha de S. Paulo) version 2008[2].

To validate our method we compared PragmaticOIE against three relevant methods of the state of the art on the Portuguese language: ArgOE [9], SGC_2017 [13] and DependentIE [12]. Two evaluation metrics have been adopted by those relevant methods [3,7] and were employed in this work: precision and minimality. Precision in this work comprises an extracted fact that is coherent with the information present in the sentence [8,11]. Minimality is an extracted fact that cannot be decomposed into new facts from its arguments [12].

It is noteworthy that in Open Information Extraction systems, *recall* is one of the most challenging measures to be calculated due to the open nature of these systems (MAUSAM, 2016). Frequently, a human can identify new facts to be extracted that the approaches are not able to.

Our evaluation process was performed by two experts (Brazilian Portuguese natives). The Kappa coefficient [5] was used to measure the degree of agreement between the assessments made by both experts.

6 Results and Discussions

Results from the three methods through the five datasets were evaluated with precision and minimality metrics. The performance of PragmaticOIE is depicted in Table 1. Our PragmaticOIE method was superior in almost all situations, although ArgOE [9] which obtained better results in minimality sometimes. This trade-off occurs because ArgOE extracts a small number of extractions when it is compared to PragmaticOIE. The best results are bolded.

Kappa coefficient [5] was calculated to verify the degree of agreement between the evaluations carried out by both specialists. Table 2 presents the results obtained. INFER-100, CONTEXT-100, and INTENT-50 datasets obtained a high degree of agreement between the evaluators. This behavior was expected, since the sentence structure does not have many variations in these datasets. WIKI-200 and CETEN-200 datasets obtained either a high degree of agreement, but, concordances were on average 82%. This behavior was expected because, unlike datasets with pragmatic characteristics, these datasets have sentences varying both characteristics and structures.

[1] Available: https://pt.wikipedia.org/. Accessed: 08/05/2018.

[2] Available: http://www.linguateca.pt/cetenfolha/. Accessed: 08/05/2018.

Table 1. Comparison of accuracy and minimality in all evaluated datasets.

Dataset		PragmaticOIE	DependentIE	SGC_2017	ArgOE
INFER-100	Precision	**90.51**%	67.37%	70.07%	81.58%
	Minimality	77.90%	75.00%	69.79%	**91.94**%
CONTEXT-100	Precision	50.00%	**51.11**%	21.98%	33.09%
	Minimality	**89.13**%	61.74%	70.42%	84.44%
INTENT-50	Precision	**69.42**%	65.22%	24.76%	61.40%
	Minimality	**88.11**%	68.00%	70.27%	85.71%
WIKI-200	Precision	**75.91**%	56.90%	50.45%	52.17%
	Minimality	77.12%	85.00%	32.30%	**86.90**%
CETEN-200	Precision	**54.37**%	52.37%	48.23%	46.67%
	Minimality	**80.29**%	69.68%	67.53%	64.29%

Table 2. Kappa coefficient calculated on all datasets.

Datasets					
	INFER-100	CONTEXT-100	INTENT-50	WIKI-200	CETEN-200
Precision	96.20%	94.40%	93.80%	87.60%	81.80%
Minimality	97.10%	94.90%	83.10%	88.40%	77.50%

7 Extracted Facts Analysis

Methods developed inside Open IE extract facts without previously determining the type of the relation. This freedom rises a problem: incoherent extractions. In this section, we discuss some of the extractions obtained by PragmaticOIE, DependentIE [12], SGC_2017 [13] and ArgOE [9] against the adopted datasets. We organize this discussion comparing some facts extracted by PragmaticOIE with DependentIE, SGC_2017 and ArgOE.

7.1 PragmaticOIE X Other Methods

Given the following sentence *O grau de adesão e respeito a essas normas está ligado aos atributos morais dos participantes* (The degree of adherence and respect to these standards is linked to the moral attributes of the participants), in the Table 3, a comparison is made between the extracted facts by PragmaticOIE method and by DependentIE method. Observing the set of extracted fact by the PragmaticOIE method, it is perceived that it is incoherent. This fact occurred because the method incorrectly identified the 1 argument by extracting the closest to the left of the relation. On the other hand, DependentIE method was able to extract the 1 argument correctly and, consequently, a valid fact.

Considering now the sentence *A Grande Praça, o coração da cidade que já foi centro administrativo e religioso, é formada por um conjunto de*

Table 3. Example which DependentIE was better than PragmaticOIE.

PragmaticOIE	(respeito a essas normas, está ligado aos, atributos morais dos participantes) (respect to these standards, is linked to, the moral attributes of the participants)
DependentIE	(O grau de adesão, está ligado, aos atributos morais dos participantes) (The degree of adherence, is linked to, the moral attributes of the participants)
Best Fact	**DependentIE**

templos, pirâmides e acrópoles (The Great Square, the heart of the city that once was administrative and religious center, is formed by a set of temples, pyramids, and acropolis) in Table 4, it shows the extracted facts. PragmaticOIE extracts coherent information through the treatment of Asyndetic Coordination while SGC_2017 does not. Information presented in the extracted fact by the SGC_2017 approach did not correspond to the information contained in the sentence because the triple argument 1 is incoherent.

Table 4. Example where PragmaticOIE is better than SGC_2017.

PragmaticOIE	(A Grande_Praça, é formada por, um conjunto de templos) (The Great Square, is formed by, a set of temples)
SGC_2017	(centro administrativo e religioso, é formada por, um conjunto de templos) (administrative and religious center, is formed by, a set of temples)
Best fact	**PragmaticOIE**

Finally, analyzing the sentence *Serviços de emergência como hospitais e prontos-socorros funcionarão normalmente nos três dias do feriado* (Emergency services such as hospitals and emergency rooms will normally operate during the

Table 5. Example that ArgOE and PragmaticOIE are equivalent.

PragmaticOIE	(prontos-socorros, funcionarão normalmente nos, três dias do feriado) (emergency rooms, will normally operate in, the three days of the holiday)
ArgOE	(prontos-socorros, funcionarão nos, três dias do feriado) (emergency rooms, will operate in, the three days of the holiday)
Best fact	**Both**

three days of holidays) in Table 5, it shows another comparison between extracted facts extracted. We consider both facts as valid. In this example, PragmaticOIE extracts more details when compared to ArgOE by adding information that the emergency room will function normally. However, we believe that both methods were successful in carrying this kind of task.

8 Threats of Validity

According to the results obtained in INFER-100 dataset (Table 1), PragmaticOIE method was superior in the precision measure, behind only ArgOE [9] method on minimality measure. It is worth noting that the accuracy of 90% of our PragmaticOIE method indicates that the use of the inference approach in Open Information Extraction task is feasible because the method extracted a high rate of coherent facts. Considering the CONTEXT-100 dataset, results show that PragmaticOIE method was also superior to both SGC_2017 and ArgOE and similar to DependentIE in the precision measure. On the other hand, in the minimality measure, PragmaticOIE was superior to all evaluated methods. Although in a first moment, the results indicate that the contextual approach reduces the amount of extracted facts, it is worth noting that this is an expected behavior in this dataset, since many facts extracted by other methods become a context information and no longer a fact is extracted within the PragmaticOIE method. Considering the INTENT-50 dataset, results show that the PragmaticOIE method was superior to the other methods evaluated.

In WIKI-200 dataset (Table 1), PragmaticOIE was superior with almost all the evaluated criteria, behind only ArgOE [9] in precision and minimality measure. However, this trade-off occurred because ArgOE obtained a much lower number of extractions when compared to ours. In CETEN-200 dataset, all methods obtained inferior results in comparison with WIKI-200, regarding the precision measure. This could be due to the construction of CETEN-200 dataset which is formed by journalistic texts, almost all sentences have a far-fetched language, even if it has several domains, which it is hard to extract the facts. In both datasets, we verified that PragmaticOIE was superior to other methods in all evaluated measures, with exception of ArgOE which was superior in minimality measure.

We verified that PragmaticOIE might fail, especially regarding the rules of transitive and symmetric inference, in which we consider relations of type $IS-A$ as general relations to infer new facts. It is worth considering that our treatment of Asyndetic Coordination is not able to cover all the possibilities of the Portuguese language, since the sentences can be written in different forms, making this treatment hard. Finally, our results confirmed that PragmaticOIE was superior both in precision and in the number of valid extractions due to our concerns about writing aspects in Portuguese language and the generalization of our inference approach.

9 Conclusions

Open IE has been receiving considerable attention. Main methods of the state of the art are still flawed to extract implicit information from documents. Our PragmaticOIE method takes into account particular aspects such as inference, context and intention. Our inferential approach was based on both methods [4,13], but different from them, PragmaticOIE can generalize transitive and symmetric facts and can extract both inferences in the same sentence. Our contextual approach was based on [11], but different from them, PragmaticOIE has added new contextual features that increase the consistency and informativeness of the extractions. Finally, the intentional approach proposed in this work extracts intentional information, and no work has undertaken it, as far as we know. The combination of these levels enables the achievement to extract the first level of pragmatic facts. As extracting new useful facts, our results outstand the other methods of the state in almost all datasets and measures.

Acknowledgement. Authors would like to thank FAPESB BOL3288/2015 for financial support.

References

1. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction for the web. IJCAI **7**, 2670–2676 (2007)
2. Banko, M., Etzioni, O., Center, T.: The tradeoffs between open and traditional relation extraction. In: ACL, vol. 8, pp. 28–36. Association for Computational Linguistics, Stroudsburg (2008)
3. Bast, H., Haussmann, E.: Open information extraction via contextual sentence decomposition. In: 2013 IEEE Seventh International Conference on Semantic Computing (ICSC), ICSC 2013, pp. 154–159. IEEE, Irvine (2013)
4. Bast, H., Haussmann, E.: More informative open information extraction via simple inference. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C.X., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 585–590. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06028-6_61
5. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. **22**(2), 249–254 (1996). http://dl.acm.org/citation.cfm?id=230386.230390
6. da Costa, J.C.: A teoria inferencial das implicaturas: descrição do modelo clássico de grice. Letras de Hoje **44**(3) (2009)
7. Del Corro, L., Gemulla, R.: Clausie: Clause-based open information extraction. In: Proceedings of the 22nd International Conference on World Wide Web, WWW 2013, pp. 355–366. ACM, New York (2013). https://doi.org/10.1145/2488388.2488420
8. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, pp. 1535–1545. Association for Computational Linguistics, Stroudsburg (2011). http://dl.acm.org/citation.cfm?id=2145432.2145596

9. Gamallo, P., Garcia, M.: Multilingual open information extraction. In: Pereira, F., Machado, P., Costa, E., Cardoso, A. (eds.) EPIA 2015. LNCS (LNAI), vol. 9273, pp. 711–722. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23485-4_72
10. Grice, H.P.: Studies in the Way of Words. Harvard University Press (1989)
11. Mausam, M.S., Bart, R., Soderland, S., Etzioni, O.: Open language learning for information extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, pp. 523–534. Association for Computational Linguistics, Stroudsburg (2012). http://dl.acm.org/citation.cfm?id=2390948.2391009
12. de Oliveira, L.S., Glauber, R., Claro, D.B.: Dependentie: an open information extraction system on portuguese by a dependence analysis. Encontro Nacional de Inteligência Artificial e Computacional (2017)
13. Sena, C.F.L., Glauber, R., Claro, D.B.: Inference approach to enhance a portuguese open information extraction. In: Proceedings of the 19th International Conference on Enterprise Information Systems, ICEIS, vol. 1, pp. 442–451. INSTICC, ScitePress, Porto, Portugal (2017). https://doi.org/10.5220/0006338204420451
14. Wu, F., Weld, D.S.: Open information extraction using wikipedia. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, pp. 118–127. Association for Computational Linguistics, Stroudsburg (2010). http://dl.acm.org/citation.cfm?id=1858681.1858694

Concordance Comparison as a Means of Assembling Local Grammars

Juliana P. C. Pirovani[1](✉), Elias de Oliveira[1], and Eric Laporte[2]

[1] Universidade Federal do Espírito Santo - UFES, Av. Fernando Ferrari, 514,
Vitória, ES 29075-910, Brazil
jupcampos@gmail.com, elias@lcad.inf.ufes.br

[2] Université Paris-Est, LIGM, UPEM/CNRS/ENPC/ESIEE,
77420 Champs-sur-Marne, France
eric.laporte@univ-paris-est.fr

Abstract. Named Entity Recognition for person names is an important but non-trivial task in information extraction. This article uses a tool that compares the concordances obtained from two local grammars (LG) and highlights the differences. We used the results as an aid to select the best of a set of LGs. By analyzing the comparisons, we observed relationships of inclusion, intersection and disjunction within each pair of LGs, which helped us to assemble those that yielded the best results. This approach was used in a case study on extraction of person names from texts written in Portuguese. We applied the enhanced grammar to the Gold Collection of the Second HAREM. The F-Measure obtained was 76.86, representing a gain of 6 points in relation to the state-of-the-art for Portuguese.

Keywords: Concordance · Local grammar
Named Entity Recognition

1 Introduction

Named Entity Recognition (NER) involves automatically identifying names of entities such as persons, places and organizations. Person names are a fundamental source of information. Many applications seek information on individuals and their relationships, e.g. in the context of social networks. However, extracting this type of Named Entity (NE) is challenging: person names are an open word class, which includes many words and grows every day [8].

"*A good portion of NER research is devoted to the study of English, due to its significance as a dominant language that is used internationally*" [15, page 470]. An influential impetus to the development of systems for this purpose in Portuguese came with the HAREM [9,14] events, a joint assessment of the area organized by Linguateca [7]. The annotated corpora used in the first and second HAREM, known as the Golden Collection (GC), serve as a reference for recent works on Portuguese NER.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 57–65, 2018.
https://doi.org/10.1007/978-3-319-99722-3_6

The main approaches used to develop NER systems involve (i) machine learning, whereby systems learn to identify and classify NEs from a training corpus, (ii) the linguistic approach, which involves manual description of rules in which NEs can appear, and (iii) a hybrid approach that combines both previous methods.

"*Local grammars (LG) are finite-state grammars or finite-state automata that represent sets of utterances of a natural language*" [6, page 1]. They were introduced by Maurice Gross [5] and serve as a way to group phrases with common characteristics (usually syntactic or semantic). Describing rules in the form of LGs for the construction of Information Extraction (IE) systems requires human expertise and training in linguistics; little computational aid for this task is available.

A method for constructing LGs around a keyword or semantic unit is presented by [6]. LGs for extracting person names from Portuguese texts were presented in [3,11]. In the Second HAREM [9], the Rembrandt system, which uses grammar rules and Wikipedia as sources of knowledge [4], ranked best for the 'person' category. A comparison between four tools to recognize NEs in Portuguese texts [2] suggested that the rule-based approach is the most effective for person names. Recently, LGs have been successfully integrated in a hybrid approach to Portuguese NER [12].

This paper describes how to use the Unitex concordance comparison tool [1] as an aid to constructing an LG. Our point of departure was a set of LGs to identify person names in Portuguese texts. By comparing concordances obtained from them, we found some relationships between them in terms of set theory. Taking into account these relationships, we picked the best LGs and combined them in order to achieve better performance.

This article is organized as follows. Section 2 presents the methodology used in this work. The results of the study are presented in Sect. 3, and Sect. 4 presents conclusions and avenues for future research.

2 The Methodology

The input to our experiment was a repository of small LGs to recognize person names. Some were obtained from the literature (e.g. those presented in [3]) and we created others.

All of these LGs were created and processed with Unitex [1], an open-source system initially developed at University of Paris-Est Marne-La-Vallée in France. A local grammar is represented as a set of one or more graphs referred to as Local Grammar Graphs (LGG). Unitex allows for creating LGGs, preprocessing texts, applying dictionaries to texts, applying LGs to extract information, generating concordances and comparing concordances.

The LGG shown in Fig. 1 recognizes honorific titles such as *Sr.*, *Sra.* and *Dr.* ("Mr.", "Mrs.", "Dr.") followed by words with the first letter capitalized, as identified by the code `<PRE>` in Unitex dictionaries. The `<<..>>` after `<PRE>` denotes the application of a morphological filter to words with the first letter capitalized, indicating that they must include at least two characters. This prevents the recognition of definite articles at the beginning of sentences, for example. Between the capitalized words, prepositions or abbreviations may occur and are recognized by two graphs, *Preposicao.grf* and *Abreviacoes.grf*, which have been created separately and are included as subgraphs. Examples of phrases recognized by the graph (occurrences) include *Sra. Joana da Silva* and *Dr. Antônio de Oliveira Salazar*. A list of occurrences accompanied with one line of context is referred to as a concordance.

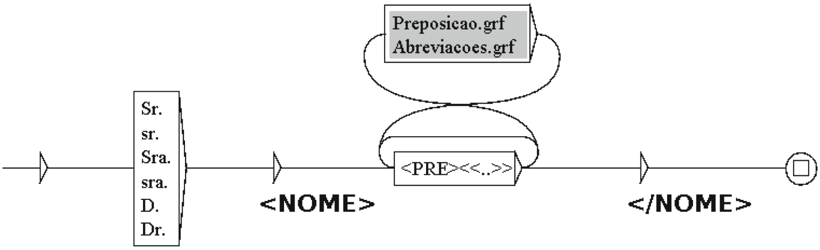


Fig. 1. LGG G_1 (ReconheceFormasDeTratamento.grf)

Unitex allows for attaching outputs to graph boxes. Outputs are displayed in bold under boxes. In Fig. 1, `<NOME>` (“name”) and `</NOME>` shown under the arrows represent such outputs. Unitex inserts them into the concordance when a graph is applied in the “MERGE with input text” mode. Thus, the identified names appear enclosed in these XML tags in the concordance file.

The LGs of the repository are small but can be combined to compose a larger grammar to identify person names.

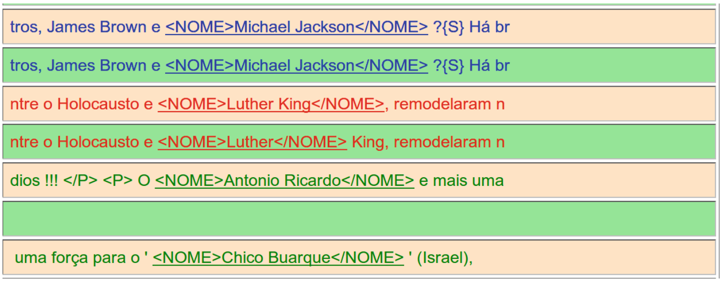
We applied the LGs of the repository to the Golden Collection (GC) of the Second HAREM, producing a concordance file for each LG. We used Portuguese and English dictionaries because several English names appear in GC texts.

The GC of the Second HAREM [9] is a subset of 129 annotated texts. These texts have different textual genres and are written in European or Brazilian Portuguese. The HAREM classifies ten categories of NEs: abstraction, event, thing, place, work, organization, person, time, value, and other. Person names, the focus of this work, are classified as a subtype within the ‘person’ category and are represented by the code PERSON (INDIVIDUAL). In the GC of the Second HAREM, 1,609 NEs are annotated with this code.

2.1 Concordance Comparison

We compared all the concordances pairwise (every pair of files) using the ConcorDiff concordance comparison tool provided by Unitex. This tool can be applied to any pair of concordance files, provided they are in the Unitex format, which is publicly documented in the manual [10].

The Unitex ConcorDiff program compares two concordance files line by line and shows their differences. The result is an HTML page that presents alternate lines of both concordances and that leaves an empty line when an occurrence appears in only one of them. An example is presented in Fig. 2. The lines with a pink background shading (lines 1, 3, 5 and 7) are from the first concordance (the first parameter to ConcorDiff), and those with a green background shading (lines 2, 4 and 6) are from the other concordance (the second parameter to ConcorDiff).



tros, James Brown e <NOME>Michael Jackson</NOME> ?{S} Há br

tros, James Brown e <NOME>Michael Jackson</NOME> ?{S} Há br

ntre o Holocausto e <NOME>Luther King</NOME>, remodelaram n

ntre o Holocausto e <NOME>Luther</NOME> King, remodelaram n

dios !!! </P> <P> O <NOME>Antonio Ricardo</NOME> e mais uma

uma força para o ' <NOME>Chico Buarque</NOME> ' (Israel),

Fig. 2. Part of a concordance comparison file (Color figure online)

Lines in blue characters (lines 1 and 2) are the occurrences common to the two concordances. In the example shown in Fig. 2, this means that both LGs recognized *Michael Jackson*. Lines in red characters (lines 3 and 4) correspond to occurrences that overlap only partially, which is the case, for instance, when an occurrence in a concordance is part of an occurrence in the other. In the example, an LG recognized *Luther King*, and the other recognized *Luther*. Lines in green characters (lines 5 and 7) are the occurrences that appear in only one of the two concordances. *Antonio Ricardo* and *Chico Buarque* were recognized only by the first LG. Lines in purple characters indicate identical occurrences with different outputs inserted, which does not happen in this example.

We then analyzed the files generated by ConcorDiff.

2.2 Composition of LG from Concordance Comparisons

Let G_X and G_Y be two LGs, and let C_X and C_Y the respective concordance files obtained by applying them to the same corpus. Thus, C_X is the set of occurrences identified by G_X, and C_Y is the set of occurrences identified by G_Y. Let $C_X \times C_Y$ be the file that shows the differences between concordances C_X and C_Y and is obtained through the ConcorDiff program of Unitex. In $C_X \times C_Y$,

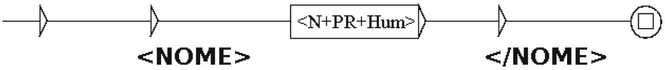
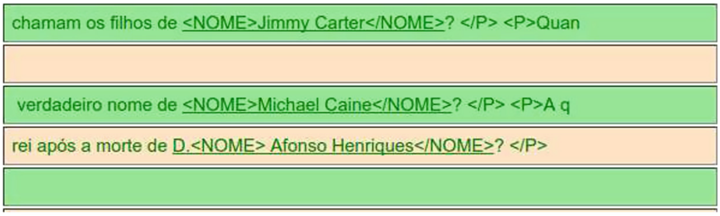


Fig. 3. LG G_2 (ReconheceNomesCompostos.grf)



chamam os filhos de <NOME>Jimmy Carter</NOME>? </P> <P>Quan

verdadeiro nome de <NOME>Michael Caine</NOME>? </P> <P>A q

rei após a morte de D.<NOME> Afonso Henriques</NOME>? </P>

Fig. 4. Part of the concordance comparison $C_1 \times C_2$ (Color figure online)

Table 1. Main relationships observed through concordance comparison

Relation	Situation	Character color	Consequence
Inclusion	$C_X \subset C_Y$	Blue and green (on green background)	Keep G_Y
	$C_Y \subset C_X$	Blue and green (on pink background)	Keep G_X
Intersection	$C_X = C_Y$	Blue	Keep or G_X or G_Y
	$C_X = C_Y$ with different outputs	Violet	Analyze ambiguity
	$C_X \cap C_Y \neq \emptyset$	Blue and green (on different backgrounds)	Keep G_X and G_Y
Disjunction	$C_X \cap C_Y = \emptyset$, with $C_X = \emptyset$	Green (on green background)	Keep G_Y
	$C_X \cap C_Y = \emptyset$, with $C_Y = \emptyset$	Green (on pink background)	Keep G_X
	$C_X \cap C_Y = \emptyset$	Green (on different backgrounds)	Keep G_X and G_Y
Disjunction with partial overlapping of occurrences	$C_X \cap C_Y = \emptyset$, with $C_X \sim C_Y$[a]	Red	Keep G_X if $\forall i\ \lvert x_i \rvert > \lvert y_i \rvert$, keep G_Y if $\forall i\ \lvert x_i \rvert < \lvert y_i \rvert$
	$C_X \cap C_Y = \emptyset$, with $\exists i\ \exists j\ \ x_i$ overlaps y_j	Red and green (on identical background)	Keep G_X and G_Y if the occurrences in green characters are relevant. If not, keep only the LG that matches larger occurrences

[a] $C_X \sim C_Y \Leftrightarrow (n = m$ and $\forall i\ \ x_i$ overlaps $y_i)$.

the elements x_1, x_2, ..., x_n of C_X are displayed on a pink background, while the elements y_1, y_2, ..., y_m of C_Y are displayed on a green background. It may exist between C_X and C_Y some relationships of the set theory, such as inclusion, intersection or disjunction, and these relationships can be observed by analyzing $C_X \times C_Y$.

Consider, for example, LGs G_1 (Fig. 1) and G_2 (Fig. 3). G_2 recognizes person names stored in dictionaries, through dictionary codes `N+PR` for proper names and `Hum` for nouns referring to human beings. Multiword person names such as *Marilyn Monroe, Cameron Diaz* and *Albert Einstein* are recognized by this LG after applying the English dictionary to the input text.

Figure 4 shows part of the concordance comparison $C_1 \times C_2$. The first line, y_1, includes the name *Jimmy Carter* recognized by G_2. The first line displayed on a pink background, x_1, includes the name *Afonso Henriques* occurring after *D.* and recognized by G_1. Since lines in green characters are occurrences identified by only one of the two graphs, the first two occurrences were identified by G_2 only, and the last one by G_1 only. If all the lines of the comparison are in green characters and distributed between the two background colors, C_1 and C_2 are disjoint sets: thus, both LGs G_1 and G_2 are worth retaining as subgraphs of a grammar because they recognize different names.

Table 1 summarizes the main set-theoretic relationships identified. Each situation has a consequence in terms of priority between LGs, for example: G_X can be discarded if G_Y is retained. After analysing relationships between all pairs of LGs, we selected a subset of LGs and combined them into a larger LG (30 LGGs) by invoking them in a main graph.

3 Results and Discussion

We could not compare the performance of the obtained LG to the initial set of small LGs, since this set does not make up a single annotator together. Instead, we simply evaluated two annotators, one based on the obtained LG and another on an enhanced version of it, and we compared the results to those of Rembrandt, as a widely known reference.

We applied the obtained LG to the HAREM corpus and generated an XML file with the identified NEs, annotated according to directives of the Second HAREM. Parts of the person names identified by LG that appear isolated in the text are also annotated.

This file was submitted to SAHARA [13] for performance evaluation. SAHARA is an online system for automatic evaluation for HAREM, which computes the precision, recall and F-measure of an NER system after the user configures the evaluation and submits XML-annotated files.

The results obtained by applying the LG to the GC of the Second HAREM were 59.06% for precision, 55.22% for recall and 57.07 for F-measure.

Then, we employed manual strategies to improve the performance of the LG. In the Second HAREM, some words in lowercase letters should form part of NE[1]. For example, the honorific titles recognized by LGG in Fig. 1 and the person's social position that appears before the name. In an example provided by HAREM,[2] *A rainha Isabel II surpreendeu a Inglaterra* "Queen Elizabeth

[1] http://www.linguateca.pt/aval_conjunta/HAREM/minusculas.html.

[2] http://www.linguateca.pt/aval_conjunta/HAREM/ExemplarioSegundoHAREM.pdf.

II surprised England", not only the name *Isabel*, but the whole phrase *rainha Isabel II* "Queen Elizabeth II" should be labeled as a person name.

We adapted the LGG ReconheceFormasDeTratamento.grf to address this issue by simply shifting the tag (`<NOME>`) before the honorific title in the graph, so that the title belongs to the tagged NE. Furthermore, we also used these words in lowercase letters to recognize the 'position' subcategory of the 'person' category, represented by PERSON(POSITION), and to recognize person names with a noun of social position in the left context.

The results obtained by the final LG are presented in Table 2. They were obtained with SAHARA by selecting the custom setting PERSON(INDIVIDUAL). This table also shows measures computed by SAHARA for Rembrandt, the system with the best performance for the 'person' category of the Second HAREM.

Table 2. Results considering PERSON(INDIVIDUAL): Rembrandt vs. final LG

System	Precision (%)	Recall (%)	F-Measure (%)
Rembrandt	79	64.08	70.76
LG	79.75	74.18	76.86

The LG outperfoms Rembrandt. The recall of the LG is approximately 10 % points above that of Rembrandt.

Although our LG recognizes only the 'individual' and 'position' subtypes of the 'person' category, its evaluation was also carried out using SAHARA for all types of categories by selecting the PERSON(*) setting. A comparison of the obtained results with the results of the four tools presented in [2] for the 'person' category is shown in Table 3.

Table 3. Results considering PERSON(*): Systems in [2] vs. final LG

Systems	Precision (%)	Recall (%)	F-Measure (%)
NERP-CRF	57	51	54
Freeling	55	61	58
Language-Tasks	63	62	62
PALAVRAS	61	65	63
LG	81	60	69

The LG has a better precision. However, as expected, it has a lower recall as it identifies fewer types of NEs: only two subtypes of the 'person' category ('individual' and 'position') are recognized, whereas the other systems recognize eight subtypes. We believe that with the addition of rules to the LG in order to recognize other subtypes of the ‘person’ category, the recall could be further increased, improving the LG approach even more as compared to other tools.

4 Conclusions

This paper presented the use of the Unitex concordance comparison tool as a computational aid in manual composition of LGs. We used this tool for the composition of an LG to identify person names in texts written in Portuguese. The same methodology can be applied to the construction of LGs for other purposes.

Table 1 was created by listing the main set-theoretic relationships (inclusion, intersection and disjunction) that we could observe when analyzing concordance-comparison files generated by Unitex. Taking into account these relationships, we could produce a more compact and easily understandable grammar. We could also observe that a concordance offers an overview of what a LG recognizes in a specific corpus, allowing ambiguities and false positives to be identified.

The results of out final LG show its potential for NE extraction. It performed better (gain of 6 points) than Rembrandt, the system with the best performance for the ‘person’ category in the Second HAREM, when evaluating the ‘person’ category, ‘individual’ subtype, for which it was created.

As avenues for future work, we plan to apply the LG approach to other corpora of texts written in Portuguese, and to assess performance with a corpus not used in the construction of the LG. Moreover, we may add rules for recognizing other types of NEs. We also intend to study the feasibility of building elementary LGGs automatically or semi-automatically from examples, with the goal of minimizing human effort during construction. The concordance comparison tool presented in this article might facilitate the automation of decision-making for this purpose.

References

1. Unitex (2018). http://unitexgramlab.org/. acesso em: 02 March 2018
2. Amaral, D.O., Fonseca, E.B., Lopes, L., Vieira, R.: Comparative Analysis of Portuguese Named Entities Recognition Tools. In: Chair, N.C.C., et al. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 2554–2558. European Language Resources Association (ELRA), Reykjavik, Iceland, May 2014
3. Baptista, J.: A Local Grammar of Proper Nouns. In: Seminários de Linguística, vol. 2, pp. 21–37. Universidade do Algarve, Faro (1998)
4. Cardoso, N.: REMBRANDT-Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In. In Cristina Mota and Diana Santos (eds.). Desafios na Avaliaação Conjunta do Reconhecimento de Entidades Mencionadas, vol. 1, pp. 195–211. Linguateca (2008)
5. Gross, M.: The Construction of Local Grammars. In ROCHE, E.; SCHABÈS, Y. (eds.). Finite-state language processing, Language, Speech, and Communication, Cambridge, Mass, pp. 329–354 (1997)

6. Gross, M.: A Bootstrap Method for Constructing Local Grammars. In: Bokan, N. (ed.) Proceedings of the Symposium on Contemporary Mathematics, pp. 229–250. University of Belgrad (1999)
7. Linguateca: (2018), http://www.linguateca.pt. acesso em: 02 March 2018
8. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
9. Mota, C., Santos, D.: Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM. Linguateca (2008). https://www.linguateca.pt/LivroSegundoHAREM/
10. Paumier, S.: Unitex 3.1 User Manual (2016). http://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf
11. Pirovani, J.P.C., de Oliveira, E.: Extração de Nomes de Pessoas em Textos em Português: uma Abordagem Usando Gramáticas Locais. In: Computer on the Beach 2015, pp. 1–10. SBC, Florianópolis, SC, March 2015
12. Pirovani, J.P.C., de Oliveira, E.: CRF+LG: A hybrid approach for the portuguese named entity recognition. In: International Conference on Intelligent Systems Design and Applications (ISDA 2017), Delhi, India (2017)
13. SAHARA: (2018). http://www.linguateca.pt/SAHARA/. acesso em: 02 March 2018
14. Santos, D., Cardoso, N.: Reconhecimento de Entidades Mencionadas em Português: Documentação e Actas do HAREM, a Primeira Avaliação Conjunta na Área. Linguateca (2007). http://www.linguateca.pt/aval_conjunta/LivroHAREM/Livro-SantosCardoso2007.pdf
15. Shaalan, K.: A Survey of Arabic Named Entity Recognition and Classification. Comput. Linguistics **40**(2), 469–510 (2014). https://doi.org/10.1162/COLI_a_00178

Challenges of an Annotation Task for Open Information Extraction in Portuguese

Rafael Glauber, Leandro Souza de Oliveira, Cleiton Fernando Lima Sena,
Daniela Barreiro Claro(✉), and Marlo Souza

Formalisms and Semantic Applications Research Group (FORMAS)
LASiD/DCC/IME, Federal University of Bahia, Salvador, Bahia, Brazil
rglauber@dcc.ufba.br, leo.053993@gmail.com, cflsena2@gmail.com,
{dclaro,msouza1}@ufba.br
http://formas.ufba.br/

Abstract. Open information extraction (Open IE) is a task of extracting facts from a plain text without limiting the analysis to a predefined set of relationships. Although a significant number of studies have focused on this problem in the last years, there is a lack of available linguistic resources for languages other than English. An essential resource for the evaluation of Open IE methods is notably an annotated corpus. In this work, we present the challenges involved in the creation of a golden set corpus for the Open IE task in the Portuguese language. We describe our methodology, an annotation tool to support the task and our results on performing this annotation task in a small validation corpus.

Keywords: Open information extraction · Portuguese · Corpora
Annotation

1 Introduction

While the quantity and diversity of textual contents on the Web are continually growing, traditional Information Extraction (IE) tools are designed to identify a fixed set of information types, thus having low coverage regarding all possible information obtained and processed from the Web. To solve this problem, Banko et al. [4] proposed the Open IE task, which aims to extract facts from sentences without predefining a set of target relationships to be analyzed.

Although Open IE has undoubtedly gained importance in the area in the last decade, most systems and methods available in the literature are still focused on the English language [25]. Considering those systems focused on Open IE in Portuguese language, only a few of them have been proposed in the last five years.

Supported by CAPES process grant number 1573345.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 66–76, 2018.
https://doi.org/10.1007/978-3-319-99722-3_7

The creation of an annotated corpora is a crucial step for fostering the development of new methods and the evaluation of existing ones in Natural Language Processing (NLP). Thus, we believe that the construction of such a resource for Open IE in Portuguese can have a considerable impact in the development of systems and methods available for this language.

As such, this work describes the process of building a reference corpus for Open IE in the Portuguese language, explaining the methodology and open challenges. We present our results and discuss them, looking towards the creation of a golden dataset for Open IE. Some of our contributions are described as follows:

- Systematic mapping on Open IE for the Portuguese language;
- Definition of an annotation guide for Open IE for the Portuguese language;
- Development of a support tool for the annotation task (OpenIEAnn);
- Analysis of our results in a small test corpus for the annotation task;

This paper is organized into sections as follows. Section 2 presents our systematic mapping for the Portuguese language. Section 3 describes our annotation guide. Section 4 presents the experimental setup and Sect. 5 presents and discusses our results. Finally, Sect. 6 concludes our work.

2 Systematic Mapping

The available resources, such as annotation tools and linguistic resources, for Open IE in Portuguese are insufficient when compared to those for the English language. Aiming to identify the studies and available resources in Portuguese for this task, we conducted a systematic mapping study (SMS). Our SMS follows Petersen's work [21] recommendations and the Systematic Mapping Study on Open Information Extraction [15]. In the planning step, we establish the main research question as follows: "What are the studies conducted in Portuguese Open IE area?". The search method used to find the primary studies were carried out by an automatic search in electronic databases. To recover primary studies on Portuguese, we used two keywords: *"open information extraction" + portuguese* or *"open relation extraction" + portuguese*[1]. Two databases was adopted: *Google Scholar*[2] and *dblp*[3].

Our inclusion criteria retain all studies on Open IE area focusing on the Portuguese language. We looked for studies, which contain keywords at least in title, summary, and keywords fields. *Exclusion criteria* (F–filters) for primary studies are:

- **F1:** Remove studies which have some "Open IE" terms, but are not studies on the topic (−103 entries removed).

[1] Queries was performed on March 2018.

[2] http://scholar.google.com Query 1: 172 entries and Query 2: 35 entries.

[3] https://dblp.uni-trier.de Query 1: 2 entries and Query 2: no matches.

- **F2:** Remove studies not published in journals or conferences (−33 entries removed).
- **F3:** Remove surveys or review papers (−10 entries removed).
- **F4:** Remove studies which do not extract facts from texts written in Portuguese (−11 entries removed).
- **Duplicated:** Remove one of the duplicate occurrences (−36 entries removed).

Table 1 presents the summary of the studies published in Portuguese Open IE area. As far as we know, only three studies made the datasets public during their research. To this point, we consider as public, the dataset indexed by some URL available on the Web and presented in the paper. The authors in [11] published a single dataset of sentences for Open IE evaluation systems to Portuguese[4].

Table 1. Summary of the studies published in Portuguese Open IE area. The sources were conferences and journals, for the last one, we used the italic font. R.Group is the research group or institute. Input indicates the NLP tasks combined with the proposed method. The approach indicates whether it is rule-based (Rules), machine learning (Data) or both (Mixed) and machine translate. ML indicates whether the system is multilingual and PD stands for Public Dataset.

Study	System	Year	Source	R.Group	Type	Input	Approach	ML	PD
[13]	DepOE	2012	ROBUS-UNSUP	CITIUS	Proposal	DP	Rules	✓	
[10]		2013	ENIAC	UFC/UNIFOR	Proposal	POS	Rules		
[14]	DepOE+	2014	*SEPLN*	CITIUS	Proposal	DP, Coreference	Rules	✓	
[26]		2014	*Linguamática*	FORMAS	Proposal	POS, Chunker	Mixed		
[7]		2014	IBERAMIA	PUC-RS	Proposal	POS, Parser	Data		
[11]	ArgOE	2015	EPIA	CITIUS	Proposal	DP	Rules	✓	✓
[9]		2015	HLT-NAACL	CMU/GOOGLE	Proposal	OLLIE [24]	Translate	✓	✓
[20]	Report	2015	STIL	UNIFOR	Proposal	POS, Chunker	Rules		
[6]		2016	PROPOR	PUC-RS	Proposal	POS, Parser	Rules		
[23]	RAPPORT	2016	PROPOR	CISUC	Application				
[12]	LinguaKit	2017	*Linguamática*	CITIUS	Application				
[1]		2017	*Knowledge Organization*	PUC-RS	Proposal	POS, Parser	Data		
[5]		2017	STIL	FORMAS	Proposal	POS, Chunker	Data		
[25]		2017	ICEIS	FORMAS	Proposal	POS, Chunker	Mixed		
[18]	DependentIE	2017	ENIAC	FORMAS	Proposal	DP	Rules		
[27]	SGS	2018	*J.UCS*	FORMAS	Proposal	POS, Chunker	Mixed		✓

[4] Download at http://gramatica.usc.es/~gamallo/prototypes/ArgOE-beta.tar.gz.

We believe that the resource is limited in size (103 sentences) and it is not domain independent (texts on ecological issues). The second study was presented by the authors in [9][5] whose dataset has not been revised by humans. Finally, the authors in [27] published their datasets in PostgreSQL's *dump* format[6]. This last dataset was manually annotated and it is composed of 582 facts extracted from sentences from the *CETENFolha* corpus[7]. We are unable to find the methodology applied in the annotation task, and thus it is hard to judge the quality of their result.

3 Annotation Guide

Our annotation guide is strongly based on the guidelines proposed by Hovy and Lavid [16]. We performed the task in five steps as shown in Fig. 1. The first step is the definition of the task that is based on the definition proposed by the authors in [28]: "An open information extractor is a function from a document, d, to a set of triples,$\{\langle arg_1, rel, arg_2\rangle\}$, where the *args* are noun phrases and *rel* is a textual fragment indicating an implicit, semantic relation between the two noun phrases.".

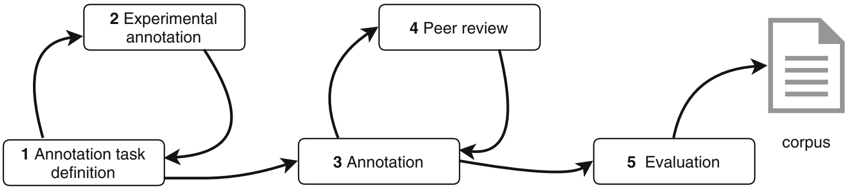


Fig. 1. Our flow to Portuguese Open IE annotation task.

The proposed definition by the authors in [28] is general, and it can lead to many differences among the annotators. In an interactive process between steps 1 and 2, we define a set of constraints to be applied to such general definition. This set of constraints does not indicate all possible restrictions but enables the proposed annotation task to be feasible. Therefore, the **first challenge** of this annotation is to set a threshold for an open-domain task. There is a trade-off between the feasibility of performing an evaluation of the outcome of the task and limiting the set of possible relationships from being extracted into a sentence. Our constraints are based on X-bar theory definitions published in "Novo manual de sintaxe" [17] and the set of constraints (C) for this study is as follows:

[5] Download at https://console.cloud.google.com/storage/browser/wikipedia_multilingual_relations_v1/multilingual_relations_data/auto/extractions/.

[6] Download at http://formas.ufba.br/page/downloads.

[7] http://www.linguateca.pt/cetenfolha/.

C1. When there is a word chain through a preposition forming a noun phrase (NP), we first select the fragment that is composed of a noun, proper noun or pronoun, its respective determinants and direct modifiers (articles, numerals, adjectives and some pronouns). For example:

- Adjectives: HIGH players/NEW students
- Articles: THE boy/A girl
- Numerals: TWO hamburgers
- Pronouns: MY shoes/SOME people

C2. When a sentence has an transitive verb with preposition (indirect mode), the preposition will be attached to the fragment *rel*. For example, given the sentence "David travels to another country." one fact could be {*David, travels to, another country*}.

C3. We call minimal fact (minimal) any extracted fact having as arguments NPs composed only of a noun, proper noun or pronoun with its determinants and direct modifiers. For example, in the sentence "Senator Barack Obama of Illinois was elected president of the United States over Senator John McCain of Arizona.", one minimal fact could be {*Senator Barack Obama, was elected president of, the United States*}, but {*Senator Barack Obama of Illinois, was elected president of, the United States*} is not minimal. It is, however, considered as a valid extraction.

C4. If there are fragments with a noun function (preposition chain) that modify arguments in minimal facts, new facts (not minimal) must be added by the annotator (see C3 second triple example).

C5. A fact must only be extracted from a sentence if it contains a proper noun or pronoun in, at least, one of the arguments.

C6. For n–ary facts, if there is no significant loss of information, the annotator must extract multiple binary facts. In the example presented by the authors in [2] "Elvis moved to Memphis in 1948.", two extracted facts {*Elvis, moved to, Memphis*} and {*Elvis, moved in, 1948*} are valid and minimal.

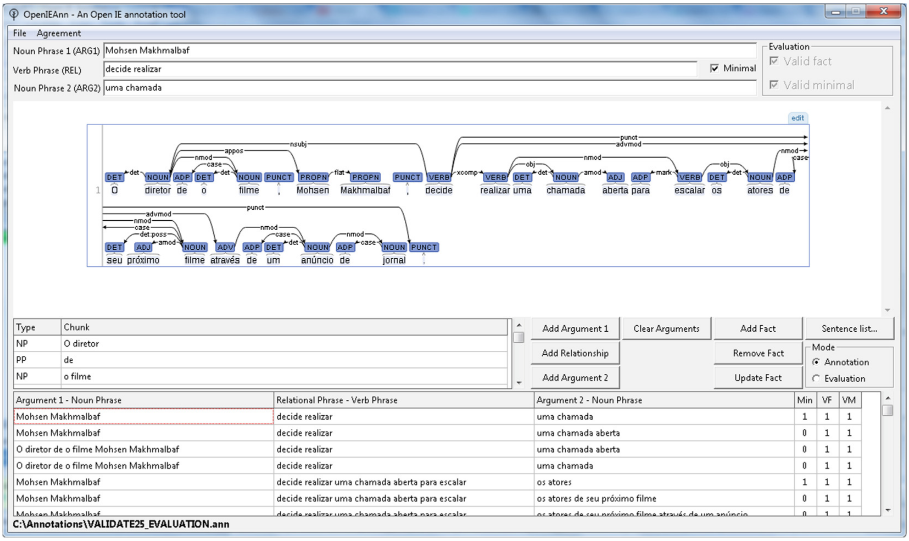
C7. The coordinating conjunctions with additive function can generate multiple extracted facts and also a fact with the coordinated conjunction. In the example "The newspaper is published in London and Madrid." there are at least three facts {*The newspaper, is published in, London*}, {*The newspaper, is published in, Madrid*} and {*The newspaper, is published in, London and Madrid*}.

C8. Relations and arguments in the extracted facts must agree in number. For example, in the sentence "Two of the world's main cities are London and Madrid.", the subject and the verb of the sentence are plurals. Thus the only possible extraction is {*Two of the world's main cities, are, London and Madrid*}, despite the coordinating conjunction.

The third step in this guide is the annotation task, and each annotator performed the task individually. This step is interactive with a fourth step evaluation. All annotators present their questions and then perform a new round of annotation to increase the agreement among participants. The last step is to evaluate all extracted facts among all annotators. Annotators evaluate all facts carried out among all annotators. The final version of the corpus is the set of all extracted facts with the evaluation by each annotator.

3.1 Proposed Tool

OpenIEAnn tool was developed to support the proposed annotation task. Figure 2 presents the main form of this tool. Two primary functions of this tool are to support the user in identifying and extracting facts in sentences and calculating the agreement among the raters of the annotation task. The tool was built using *brat rapid annotation tool*[8] version 1.3, *CoreNLP* version 3.9.1[9] for POS tagger and DP, *CoGrOO*[10] version 4.0 for Chunker, DKPro Statistics[11] version 2.1.0 for agreements, and *Universal Dependencies*[12] version 2.0 for CoreNLP models[13]. The tool, as well as all the models and resources are available in review version link[14]. Other functions available in OpenIEAnn are: (i) import raw text file with sentences to annotation format and (ii) export only sentences with the extracted facts.


OpenIEAnn - An Open IE annotation tool
File
Agreement
edit
DET NOUN ADP DET NOUN PUNCT PROPN PROPN PUNCT VERB VERB DET NOUN ADJ ADP VERB DET NOUN ADP
O diretor de o filme , Mohsen Makhmalbaf , decide realizar uma chamada aberta para escalar os atores de
DET ADJ NOUN ADV ADP DET NOUN ADP NOUN PUNCT
seu próximo filme através de um anúncio de jornal .
det appos nmod case nsubj flat xcomp obj amod mark punct advmod det:poss

Noun Phrase 1 (ARG1) Mohsen Makhmalbaf

Verb Phrase (REL) decide realizar ☑ Minimal

Noun Phrase 2 (ARG2) uma chamada

Evaluation
☑ Valid fact
☑ Valid minimal

Type	Chunk
NP	O diretor
PP	de
NP	o filme

Add Argument 1
Clear Arguments
Add Fact
Sentence list...
Add Relationship
Remove Fact
Add Argument 2
Update Fact

Mode
● Annotation
○ Evaluation

Argument 1 - Noun Phrase	Relational Phrase - Verb Phrase	Argument 2 - Noun Phrase	Min	VF	VM
Mohsen Makhmalbaf	decide realizar	uma chamada	1	1	1
Mohsen Makhmalbaf	decide realizar	uma chamada aberta	0	1	1
O diretor de o filme Mohsen Makhmalbaf	decide realizar	uma chamada aberta	0	1	1
O diretor de o filme Mohsen Makhmalbaf	decide realizar	uma chamada	0	1	1
Mohsen Makhmalbaf	decide realizar uma chamada aberta para escalar	os atores	1	1	1
Mohsen Makhmalbaf	decide realizar uma chamada aberta para escalar	os atores de seu próximo filme	0	1	1
Mohsen Makhmalbaf	decide realizar uma chamada aberta para escalar	os atores de seu próximo filme através de um anúncio	0	1	1

C:\Annotations\VALIDATE25_EVALUATION.ann

Fig. 2. Main form of the OpenIEAnn annotation tool.

[8] http://brat.nlplab.org/.
[9] https://stanfordnlp.github.io/CoreNLP/.
[10] http://cogroo.sourceforge.net/download/current.html.
[11] https://dkpro.github.io/dkpro-statistics/.
[12] http://universaldependencies.org/.
[13] The Brazilian Portuguese Universal Dependencies is converted from the Google Universal Dependency Treebanks version 2.0.
[14] http://formas.ufba.br/.

4 Experimental Setup

We carried out the annotation task on a small corpus. We randomly selected sentences from five different sources and domains. From each source, we recovered five sentences and built a corpus with 25 sentences as follows:

– 5 Wikipedia sentences - source in Portuguese Wikipedia version https://pt.wikipedia.org/wiki/
– 5 CETENFolha sentences - source by CETENFolha corpus https://www.linguateca.pt/cetenfolha
– 5 WEB sentences - source by Bing API
– 5 Adoro Cinema sentences - source by crawler in website http://www.adorocinema.com/
– 5 Europarl sentences - source by Europarl corpus v7.0 http://www.statmt.org/europarl/

Five Brazilian Portuguese natives participated in this experiment identified as rater 1, 2, 3, 4 and 5. Each rater was invited to perform two tasks. The first one was the Open IE task performed within the set of sentences considering our constraints. The second task was performed after extracting all the facts from the five raters. All those extracted facts were unified and the second task was to classify those extracted facts manually as valid or invalid.

Free-marginal multi-rater kappa (Randolph's kappa [22]) was set to calculate the agreement among the raters. The agreement of the second task is trivial. All raters evaluated the same extracted facts in a binary classification. For the first task, the divergence starts when each rater performs Open IE extractions different from other raters. The label is nominal, and each extracted fact must have a label given by a rater and if other raters have also performed the same extraction, thus the same label is assigned among them. Otherwise, random and different labels from other raters are given. The comparison of the extracted facts among raters was done in three ways: (i) full – that compares the arguments, relationship, and minimal property separately, (ii) partial – that does not evaluate the minimal property, and (iii) text – that concatenates the arguments with the relationship forming a single string.

5 Results

There are two rounds between the third and fourth step of our annotation task (Fig. 1). The degree of agreement among the raters in the first round was presented in Table 2. Generally, the agreement is low when we remember the small set of sentences in this step. The **second challenge** is to unify the understanding about the task performed. We believe that constraints should be followed by a relevant example set to fix the task rules.

In the first task, we performed two rounds to evaluate the behavior of raters between steps 3 and 4. In Table 3, we present the results of agreement for the second round of the first task. After an alignment meeting about the rules of

Table 2. Degree of agreement among raters in the 1st round of manual annotation.

Measure	Mode	1–2	1–3	1–4	1–5	2–3	2–4	2–5	3–4	3–5	4–5	All raters
Kappa	Full	0.0874	0.1795	0.1306	0.1547	0.0715	0.0911	0.0937	0.1313	0.0916	0.1212	0.0570
	Partial	0.1745	0.2164	0.1722	0.2294	0.1050	0.1183	0.1238	0.1760	0.1288	0.1517	0.0805
	Text	0.2142	0.2571	0.2007	0.2741	0.1321	0.1488	0.1796	0.1807	0.1577	0.1960	0.1013
#Fact	Full	189	198	226	165	231	247	187	263	213	227	435
	Partial	175	192	218	155	224	241	182	253	206	221	406
	Text	166	185	212	148	216	232	170	252	200	212	376
#Exact fact	Full	17	36	30	26	17	23	18	35	20	28	5
	Partial	31	42	38	36	24	29	23	45	27	34	12
	Text	36	48	43	41	29	35	31	46	32	42	17

the task, the agreement increased. The high agreement between raters 1–4 and 4–5 in both rounds contributed to achieving our results. However, there was a low agreement between raters 1–2. The **third challenge** is to solve the trade-off between the dedicated time to the task and the result of agreement expected for the generate corpus. As it is expected, a high amount of raters can decrease the agreement and require a high amount of rounds for the task. One suggestion is to eliminate the worst rater as done by the authors in [19].

Table 3. The degree of agreement among raters in the 2nd round of manual annotation.

Measure	Mode	1–2	1–3	1–4	1–5	2–3	2–4	2–5	3–4	3–5	4–5	All raters
Kappa	Full	0.0821	0.2315	0.2799	0.1130	0.0640	0.1001	0.0781	0.1662	0.1091	0.2233	0.0791
	Partial	0.1556	0.2480	0.3288	0.1630	0.0952	0.1240	0.1109	0.1870	0.1397	0.2615	0.1018
	Text	0.2081	0.2837	0.3607	0.1967	0.1360	0.1676	0.1545	0.2093	0.1818	0.2776	0.1252
#Fact	Full	189	227	245	279	211	235	237	286	298	298	471
	Partial	177	224	236	267	205	230	230	281	290	289	441
	Text	166	217	229	257	195	218	217	275	278	283	411
#Exact fact	Full	16	53	69	32	14	24	19	48	33	67	8
	Partial	28	56	78	44	20	29	26	53	41	76	14
	Text	35	62	83	51	27	37	34	58	51	79	22

In the second task before generating the corpus, all raters were invited to evaluate all extracted facts from the twenty-five sentences. In this task, we observed in Table 4 a higher agreement between the raters, thus making explicit the worst rater (or the most divergent).

Table 4. Degree of agreement among raters in the evaluation of 442 extracted facts.

Measure	1–2	1–3	1–4	1–5	2–3	2–4	2–5	3–4	3–5	4–5	All raters
Kappa	0.1176	0.7285	0.4705	0.3619	0.1719	0.1945	0.2036	0.6244	0.6063	0.7466	0.4226
#Exact fact	247	382	325	301	259	264	266	359	355	386	176

Low values for the agreement in the first task hide other challenges during the task. To identify these challenges, we consult the raters on the sources of disagreement. The list was extracted from the "inter–rater–agreement–tutorial" at dkpro.github website[15]. In Table 5 the results of the survey are presented. Two questions were unanimity and alert us to two challenges. The **fourth challenge** is to solve the"Hard or debatable cases" which is the first unanimity. Nevertheless, the second unanimity is the **fifth challenge** that introduces "Personal opinions or values". We believe that difficult cases can greatly increase the bias and use of personal values. When we increase the number of constraints to solve difficult cases, we are limiting the extraction of our relationships. On the other hand, if we do not do it, difficult/hard cases are even more biased. How the problem will be handled depends on the cause. However, an important issue related to these problems is the agreement measure.

Table 5. Survey results to identify the difficulties during the tasks.

Sources of disagreement	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**	**Rater 5**
Insecurity in deciding on a category		✓	✓		
Hard or debatable cases	✓	✓	✓	✓	✓
Carelessness				✓	✓
Difficulties or differences in comprehending instructions		✓		✓	
Openness for distractions					✓
Tendency to relax performance standard when tired	✓	✓		✓	✓
Personal opinions or values	✓	✓	✓	✓	✓

Studies such as the one proposed by the authors in [8] discuss the problems of bias and prevalence for kappa measures that are widely used. The authors in [3] suggest that in cases of detection of these problems coefficients like α and π are performed. We opted for a variation of kappa that solves these problems. While careful with the choice of agreement measure, we believe that this has not determined the low agreement values. The sentence set is small, but more than 400 facts have been extracted from all the raters. There is a difficulty in standardizing the triple arguments which can generate much duplicate information. Simple example such "David is a PhD student in Computer Science" can generates triples such {*"David", is, a PhD student in Computer Science*} and {*"David", is a PhD student in, Computer Science*}. Although the two facts contain the same information, we recognize it as relations between different concepts.

[15] https://dkpro.github.io/dkpro-statistics/dkpro-agreement-tutorial.pdf.

6 Conclusions

In this work, we draw up a set of constraints and conduct an annotation task for Open IE using a small set of sentences from Portuguese. Although the small set (hard to generalize), we consider that some of the main challenges of this task have been experienced. A large number of extracted facts in comparison to the initial set of sentences indicates a great difficulty in standardizing the task. This fact leads us to the most significant result of this study which is the low agreement between the rates. The Open IE task proved challenging to define a standard concept, and annotator bias an ever-present variable. The experience of performing the task in a small corpus enables the improvement of OpenIEAnn annotation tool, thus identifying some challenges and proposing some insights to mitigate these challenges.

The next steps of this research are (i) add more sentences into the corpus, (ii) evaluate the annotator bias through more sentences and (iii) add support for different languages in the OpenIEAnn tool.

References

1. de Abreu, S.C., Vieira, R.: Relp: Portuguese open relation extraction. Knowl. Org. **44**(3), 163–177 (2017)
2. Akbik, A., Löser, A.: Kraken: N-ary facts in open information extraction. In: Proceedings of AKBC-WEKEX, pp. 52–56. ACL (2012)
3. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008)
4. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. Proc. IJCAI **7**, 2670–2676 (2007)
5. Barbosa, G., Claro, D.B.: Utilizando features linguísticas genéricas para classificação de triplas relacionais em português (generic linguistic features for relational triples classification in portuguese)[in portuguese]. In: Proceedings of STIL, pp. 132–141 (2017)
6. Collovini, S., Machado, G., Vieira, R.: Extracting and Structuring Open Relations from Portuguese Text. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 153–164. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_16
7. Collovini, S., Pugens, L., Vanin, A.A., Vieira, R.: Extraction of relation descriptors for Portuguese using conditional random fields. In: Bazzan, A.L.C., Pichara, K. (eds.) IBERAMIA 2014. LNCS (LNAI), vol. 8864, pp. 108–119. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12027-0_9
8. Eugenio, B.D., Glass, M.: The kappa statistic: a second look. Comput. Linguist. **30**(1), 95–101 (2004)
9. Faruqui, M., Kumar, S.: Multilingual open relation extraction using cross-lingual projection. In: Proceedings of NAACL HLT, pp. 1351–1356 (2015)
10. Franco, W., Pinheiro, V., Pequeno, M., Furtado, V.: Aquisição de relações semânticas a partir de textos da wikipédia. In: Proceedings of ENIAC, p. 52. Unifor (2013)

11. Gamallo, P., Garcia, M.: Multilingual open information extraction. In: Pereira, F., Machado, P., Costa, E., Cardoso, A. (eds.) EPIA 2015. LNCS (LNAI), vol. 9273, pp. 711–722. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23485-4_72
12. Gamallo, P., Garcia, M.: Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação. Linguamática **9**(1), 19–28 (2017)
13. Gamallo, P., Garcia, M., Fernández-Lanza, S.: Dependency-based open information extraction. In: Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, pp. 10–18. ROBUS-UNSUP '12, ACL (2012)
14. Garcia, M., Gamallo, P.: Entity-centric coreference resolution of person entities for open information extraction. Procesamiento del Lenguaje Natural **53**, 25–32 (2014)
15. Glauber, R., Claro, D.B.: A systematic mapping study on open information extraction. Expert Systems with Applications (2018). https://doi.org/10.1016/j.eswa.2018.06.046, http://www.sciencedirect.com/science/article/pii/S0957417418303932
16. Hovy, E., Lavid, J.: Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. Int. J. Transl. **22**(1), 13–36 (2010)
17. Mioto, C., Silva, M.C.F., Lopes, R.E.V.: Novo manual de sintaxe. Insular (2005). https://books.google.com.br/books?id=GPCpSAAACAAJ
18. de Oliveira, L.S., Glauber, R., Claro, D.B.: Dependentie: an open information extraction system on Portuguese by a dependence analysis. In: Proceedings of ENIAC, pp. 271–282. FC-UFU (2017)
19. Passonneau, R., Habash, N., Rambow, O.: Inter-annotator agreement on a multilingual semantic annotation task. In: Proceedings of LREC, pp. 1951–1956 (2006)
20. Pereira, V., Pinheiro, V.: Report-um sistema de extração de informações aberta para língua portuguesa (report-an open information extraction system for portuguese language). In: Proceedings of STIL, pp. 191–200 (2015)
21. Petersen, K., Vakkalanka, S., Kuzniarz, L.: Guidelines for conducting systematic mapping studies in software engineering: an update. Inf. Softw. Technol. **64**, 1–18 (2015)
22. Randolph, J.J.: Free-marginal multirater kappa (multirater k [free]): an alternative to fleiss' fixed-marginal multirater kappa. Online submission (2005)
23. Rodrigues, R., Gomes, P.: Improving question-answering for Portuguese using triples extracted from corpora. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 25–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_3
24. Schmitz, M., Bart, R., Soderland, S., Etzioni, O., et al.: Open language learning for information extraction. In: Proceedings of EMNLP, pp. 523–534. ACL (2012)
25. Sena, C.F.L., Glauber, R., Claro, D.B.: Inference approach to enhance a Portuguese open information extraction. In: Proceedings of the 19th International Conference on Enterprise Information Systems ICEIS, vol. 1, pp. 442–451. INSTICC, ScitePress (2017)
26. Souza, E.N.P., Claro, D.B.: Extração de relações utilizando features diferenciadas para português. Linguamática **6**(2), 57–65 (2014)
27. Souza, E.N.P., Claro, D.B., Glauber, R.: A similarity grammatical structures based method for improving open information systems. J. Univers. Comput. Sci. **24**(1), 43–69 (2018)
28. Wu, F., Weld, D.S.: Open information extraction using wikipedia. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, pp. 118–127. ACL (2010)

Task-Oriented Evaluation of Dependency Parsing with Open Information Extraction

Pablo Gamallo[1(✉)] and Marcos Garcia[2]

[1] Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela, Galiza, Spain
pablo.gamallo@usc.es
[2] LyS Group, Universidade da Coruña, Galiza, Spain
marcos.garcia.gonzalez@udc.gal

Abstract. This paper presents a comparative evaluation of several Portuguese parsers. Our objective is to use dependency parsers in a specific information extraction task, namely Open Information Extraction (OIE), and measure the impact of each parser in this task. The experiments show that the scores obtained by the evaluated parsers are quite similar even though they allow to extract different (and then complementary) itens of information.

Keywords: Dependency parsing · Open Information Extraction
Task-oriented evaluation

1 Introduction

The most popular method for dependency parser comparison involves the direct measurement of the parser output accuracy in terms of metrics such as labeled attachment score (LAS) and unlabeled attachment score (UAS). This assumes the existence of a gold-standard test corpus developed with the use of a specific tagset and a list of dependency names by following some specific syntactic criteria. Such an evaluation procedure makes it difficult to evaluate parsing systems developed with different syntactic criteria from those used in the gold-standard test. Direct evaluation has been thought to compare strategies based on different algorithms but trained on the same treebanks and using the same tokenization. In fact, the strict requirements derived from direct evaluation prevents us from making fair comparisons among systems based on very different frameworks.

In this paper, we present a task-oriented evaluation of different dependency syntactic analyzers for Portuguese using the specific task of Open Information Extraction (OIE). This evaluation allows us to compare under the same conditions very different systems, more precisely, parsers trained on treebanks with different linguistic criteria, or even data-driven and rule-based parsers. Other task-oriented evaluation work focused on measure parsing accuracy through its

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 77–82, 2018.
https://doi.org/10.1007/978-3-319-99722-3_8

influence in the performance of different types of NLP systems, such as sentiment analysis [11].

OIE is an information extraction task that consists of extracting basic propositions from sentences [2]. There are many OIE systems for English language, including those based on shallow syntactic information, e.g. TextRunner [2] and ReVerb [6], and those using syntactic dependencies: e.g. OLLIE [14] or ClauseIE [4]. There are also some proposals for Portuguese language: DepOE [10], Report [17], ArgOE [8], DependentIE [13], and the extractor of open relations between named entities reported in [3]. In order to use OIE systems to evaluate dependency parsers for Portuguese, we need an OIE system for Portuguese taking as input dependency trees. For the purpose of our indirect evaluation, we will use the open source system described in [8], which takes as input dependency trees in CoNLL-X format.

2 The Role of Dependency Parsing in OIE

We consider that it is possible to indirectly evaluate a parser by measuring the performance of the OIE system in which the parser is integrated as many errors made by the OIE system come from the parsing step. Let us take for example one of the sentences of our evaluation dataset (and described in the next section):

> `A regulação desses processos depende de várias interações de indivíduos com os seus ambientes`
> *The regulation of these processes depends on several interactions of individuals with their environments*

One of the evaluated systems extracts the following two basic propositions (to simplify we show just the English translation):

("the regulation of these processes", "depends on", "several interactions of individuals"),
**("the regulation of these processes", "depends with", "their environments")*

The second proposition is not correct since it has been extracted from an odd dependency, such as shown in Fig. 1. The dependency between "environments" and "depends" (red arc below the sentence) is incorrect since "environments" is actually dependent on the noun "interactions".[1] In sum, any odd dependency given by the parser makes the OIE system incorrectly extract, at least, one odd triple.

Furthermore, the resulting triples extracted by an OIE system are also an excellent way of visualizing the type of errors made by the depedency parser and, thereby, dependency-based OIE systems can be seen as useful linguistic tools to carry out error analysis on the parsing step.

[1] In this analysis, we use labels and syntactic criteria based on Universal Dependencies, e.g. prepositions are case-marking elements that are dependents of the noun or clause they attach to or introduce.

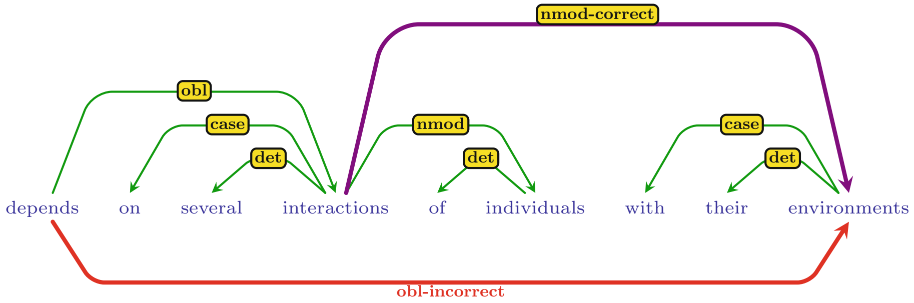


Fig. 1. Dependency analysis with Universal Dependencies. The head of "environments" is the noun "interactions" via *nmod* dependency, and not the verb "depends". (Color figure online)

3 Experiments

Our objective is to evaluate and compare diferent Portuguese dependency parsers which can be easily integrated into an open-source OIE system. For this purpose, we use the OIE module of LinguaKit, described in [8], which takes as input any text parsed in CoNLL-X format. We were able to integrate five Portuguese parsers into the OIE module: two rule-based parsers and three data-driven parsers. The rule-based systems are two different versions of DepPattern [7,9]:

The parser used by ArgOE [8], and that available in LinguaKit.[2] The three data-driven parsers were trained using MaltParser 1.7.1[3] and two different algorithms: Nivre eager [15], based on arc-eager algorithm, and 2-planar [12]. They were trained with two versions of Floresta Sintá(c)tica treebank: Portuguese treebank Bosque 8.0 [1] and Universal Dependencies Portuguese treebank (UD_Portuguese) [16], which aims at full compatibility with CoNLL UD specifications.

In order to adapt the parsers to be used by the OIE system, we implemented some shallow conversion rules to align the tagset and dependency names of Bosque 8.0 and UD_Portuguese to the PoS tags and dependency names used by the OIE system. This is not a full and deep conversion since the OIE system only uses a small list of PoS tags and dependencies. So, before training a parser on the Portuguese treebank, first we must identify the specific PoS tags and dependencies used by the extraction module, and second, we have to change them by the corresponding labels. For UD_Portuguese, we also have to change the syntactic criteria on preposition dependencies. Concerning the rule-based parsers, no adaptation is required since the OIE system is based on the dependency labels of DepPattern. A priori, this could benefit systems that did not have to be adapted, but we have no way of measuring it.

[2] http://github.org/citiususc/linguakit.

[3] http://www.maltparser.org/.

To evaluate the results of the OIE system with the parsers defined above, five systems were configured, each one with a different parser. OIE evaluation is inspired by that reported in [4,8]. The dataset consists of 103 sentences from a domain-specific corpus, called *CorpusEco* [18], containing texts on ecological issues. These sentences were processed by the 5 extractors, given rise to 862 triples. Then, each extracted triple was annotated as correct (1) or incorrect (0) according to some evaluation criteria: triples are not correct if they denote incoherent and uninformative propositions, or if they are constituted by over-specified relations, i.e., relations containing numbers, pronouns, or excessively long phrases. We follow similar criteria to those defined in previous OIE evaluations [4,5]. Annotation was made on the whole set of extracted triples without identifying the system from which each triple had been generated.

The results are summarized in Table 1. *Precision* is defined as the number of correct extractions divided by the number of returned extractions. *Recall* is estimated by identifying a pool of relevant extractions which is the total number of different correct extractions made by all the systems (this pool is our gold-standard). So, *recall* is the number of correct extractions made by the system divided by the total number of correct expressions in the pool (346 correct triples in total).[4]

Table 1. Evaluation of five OIE systems configured with five dependency parsers

Systems	Precision	Recall	Fscore
deppattern-ArgOE	.440	.265	.330
deppattern-Linguakit	.612	.361	.454
maltparser-nivrearc	.581	.248	.347
maltparser-2planar	.516	.228	.316
maltparser-nivrearc-ud	.616	.236	.341

The results show that there is no clear difference among the evaluated systems except in the case of *deppattern-Linguakit*, which relies on a rule-based parser. However, a deeper anaysis allows us to observe that rule-based and data-driven parsers might be complementary parsers as they merely share about 25% of the correct triples. More precisely, the number of correct extractions made by *deppattern-Linguakit* reaches 125 triples, but only 30 of them are also extracted by *maltparser-nivrearc*. This means that a voting OIE system consisting of the two best rule-based and data-driver parsers would improve recall in a very significant way without losing precision.

[4] Labeled extractions along with the gold standard are available at https://gramatica.usc.es/~gamallo/datasets/OIE_Dataset-pt.tgz.

4 Conclusions

In this article, we showed that it is possible to use OIE systems to easily compare parsers developed with different strategies, by making use of a coarse-grained and shallow adaptation of tagsets and syntactic criteria. By contrast, comparing very different parsers by means of direct evaluation is a much harder task since it requires carrying out deep changes on the training corpus (golden treebank). These changes involve adapting tagsets before training, reconsidering syntactic criteria at all analysis level and yielding the same tokenization as the golden treebank. Moreover, the proposed task-oriented evaluation might help linguists make deep error analysis of the parsers since the extraction of basic propositions allows humans to visualize and interpret linguistic mistakes in an easier way than obscure syntactic outputs.

Acknowledgments. Pablo Gamallo has received financial support from a 2016 BBVA Foundation Grant for Researchers and Cultural Creators, TelePares (MINECO, ref:FFI2014-51978-C2-1-R), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016–2019, ED431G/08) and the European Regional Development Fund (ERDF). Marcos Garcia has been funded by the Spanish Ministry of Economy, Industry and Competitiveness through the project with reference FFI2016-78299-P, by a Juan de la Cierva grant (IJCI-2016-29598), and by a 2017 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation.

References

1. Afonso, S., Bick, E., Haber, R., Santos, D.: Floresta sintá(c)tica: a treebank for Portuguese. In: The Third International Conference on Language Resources and Evaluation, LREC 2002, pp. 1698–1703, Las Palmas de Gran Canaria, Spain (2002)
2. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: International Joint Conference on Artificial Intelligence, pp. 2670–2676 (2007)
3. Collovini, S., Machado, G., Vieira, R.: Extracting and structuring open relations from Portuguese text. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 153–164. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_16
4. Del Corro, L., Gemulla, R.: ClausIE: clause-based open information extraction. In: Proceedings of the World Wide Web Conference (WWW-2013), pp. 355–366, Rio de Janeiro, Brazil (2013)
5. Etzioni, O., Fader, A., Christensen, J., Soderland, S., Mausam, M.: Open information extraction: the second generation. In: International Joint Conference on Artificial Intelligence, pp. 3–10. AAAI Press (2011)
6. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545. ACL (2011)
7. Gamallo, P.: Dependency parsing with compression rules. In: Proceedings of the 14th International Workshop on Parsing Technology (IWPT 2015), Bilbao, Spain, pp. 107–117. Association for Computational Linguistics (2015)

8. Gamallo, P., Garcia, M.: Multilingual open information extraction. In: Pereira, F., Machado, P., Costa, E., Cardoso, A. (eds.) EPIA 2015. LNCS (LNAI), vol. 9273, pp. 711–722. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23485-4_72
9. Gamallo, P., Garcia, M.: Dependency parsing with finite state transducers and compression rules. Inf. Process. Manag. (2018). Accessed 5 June 2018
10. Gamallo, P., Garcia, M., Fernández-Lanza, S.: Dependency-based open information extraction. In: ROBUS-UNSUP 2012: Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), Avignon, France, pp. 10–18 (2012)
11. Gómez-Rodríguez, C., Alonso-Alonso, I., Vilares, D.: How important is syntactic parsing accuracy? An empirical evaluation on sentiment analysis. Artif. Intell. Rev. 1–17 (2017, forthcoming). https://doi.org/10.1007/s10462-017-9584-0
12. Gómez-Rodríguez, C., Nivre, J.: A transition-based parser for 2-planar dependency structures. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, Stroudsburg, PA, USA, pp. 1492–1501 (2010)
13. Claro, D.B., de Oliveira, L.S., Glauber, R.: Dependentie: an open information extraction system on Portuguese by a dependence analysis. In: Proceedings of XIV Encontro Nacional de Inteligência Artificial e Computacional, pp. 271–282 (2017)
14. Mausam, M., Schmitz, M., Soderland, S., Bart, R., Etzioni, O.: Open language learning for information extraction. In: Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 523–534 (2012)
15. Nivre, J., et al.: Maltparser: a language-independent system for data-driven dependency parsing. Nat. Lang. Eng. **13**(2), 115–135 (2007)
16. Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., de Paiva, V.: Universal dependencies for Portuguese. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling), pp. 197–206, Pisa, Italy, September 2017
17. Santos, V., Pinheiro, V.: Report: um sistema de extração de informações aberta para a língua Portuguesa. In: Proceedings of the X Brazilian Symposium in Information and Human Language Technology (STIL), Natal, RN, Brazil, pp. 191–200 (2015)
18. Zavaglia, C.: O papel do léxico na elaboração de ontologias computacionais: do seu resgate à sua disponibilização. In: Martins, E.S., Cano, W.M., Filho, W.B.M. (eds.) Lingüística IN FOCUS - Léxico e morfofonologia: perspectivas e análises, pages 233–274. EDUFU, Uberlândia (2006)

Portuguese Named Entity Recognition Using LSTM-CRF

Pedro Vitor Quinta de Castro(✉), Nádia Félix Felipe da Silva,
and Anderson da Silva Soares

Universidade Federal de Goiás, Goiânia, GO 74690-900, Brazil
{I.pedrovitorquinta,II.nadia,III.anderson}@inf.ufg.br

Abstract. Named Entity Recognition is a challenging Natural Language Processing task for a language as rich as Portuguese. For this task, a Deep Learning architecture based on bidirectional Long Short-Term Memory with Conditional Random Fields has shown state-of-the-art performance for English, Spanish, Dutch and German languages. In this work, we evaluate this architecture and perform the tuning of hyper-parameters for Portuguese corpora. The results achieve state-of-the-art performance using the optimal values for them, improving the results obtained for Portuguese language to up to 5 points in the F1 score.

Keywords: Natural Language Processing · Named Entity Recognition · Deep learning · Neural networks · Portuguese language

1 Introduction

Hundreds of millions of unstructured textual information are exchanged every minute [1]. Named Entity Recognition (NER) is an important Natural Language Processing (NLP) task which focus on extracting and classifying named entities from this unstructured textual information, making them interpretable and accessible to different communication channels. The NER task can be approached either by using a rule/pattern based system, or by a machine learning method [2].

As far as we know, few works have focused on neural network architectures with evaluations performed in Portuguese language [3], while several studies have been done for English Language [4–8]. In this paper, we study the LSTM-CRF neural architecture proposed by [4] in the Portuguese language context. The architecture combines a character-based word representation model with word embeddings. This combination is fed into a bidirectional Long Short-Term

Thanks to Data-H Data Science and Artificial Intelligence (www.datah.com.br) and Aviso Urgente (https://avisourgente.com.br) for the financial support, and to Cicero Nogueira dos Santos for kindly sharing insights regarding the HAREM corpora.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 83–92, 2018.
https://doi.org/10.1007/978-3-319-99722-3_9

Memory (LSTM) network, which is finally connected to a Conditional Random Fields (CRF) layer to perform sequential classification.

As main contributions of this paper, we point out: (i) Since Portuguese is such a morphologically rich language, we intend not only to evaluate how LSTM-CRF performs in Portuguese corpora, but also to perform the hyperparameters tuning in order to achieve the best results for the language in study. (ii) We present the first comparative study about the word embeddings with LSTM based methods for Portuguese NER. We experimented with four different pre-trained word embeddings from [9]—FastText [10,11], Glove [12], Wang2Vec [13] and Word2Vec [14].

2 Related Work

Classical approaches for NER are dependent on handcrafted features, which have language-specific values and are cumbersome to maintain. For instance, [15] created a model based on CRF and used 17 different features for each word in the training corpus, such as part-of-speech (POS) tags, capitalization and the word itself, considering a context window of size 2. Deep learning approaches provide an alternative to these classical approaches.

One of the main advantages of using a deep learning approach that uses word and character level embeddings as input for model training is the independence of language specific features, since the features used are the ones that are automatically learned by using these two types of embeddings. Hence it is possible to use the same network to train models for different languages, as long as it is provided an annotated corpus for each language, as well as the pre-trained word embeddings. [3] used the same network to train models for Portuguese and Spanish, and [4] trained models for English, Dutch, German and Spanish.

Word-level embeddings are multidimensional vectors that represent features automatically learned by unsupervised training. These features represent morphological, syntactic and semantic information about the words. The unsupervised learning of these features is tipically performed on massive corpora, such as Wikipedia[1] and news archives. Using such a large amount of text allows the understanding of contexts on which certain types of words tend to occur [14]. The use of character-level embeddings is important because they allow to capture orthographic features, such as prefixes, suffixes and letter case, the latter being essential for identifying proper names in a text. These orthographic features also represent the importance of the characters used in the language in study. In Portuguese, for example, characters such as “ç” and accented vowels are quite usual. In addition, character-level embeddings are specially important for morphologically rich languages, such as Portuguese, because they provide additional intra-word and shape information to the features learned.

Dos Santos and Guimarães [3], one of the few works to apply neural networks to Portuguese NER, introduced the CharWNN architecture, which uses Convolutional Neural Networks (CNN) to learn character-level features, combined with

[1] https://www.wikipedia.org/.

pre-trained word-level embeddings to perform sequential classification. Lample et al. [4] used a deep learning approach and outperformed several methods that used handcrafted features and external resources, in four different languages. Because of that, more attention will be paid on their architecture in Sect. 3.

The approaches to Portuguese NER are still in a level way lower than languages such as English or Spanish. While the best result for Enghish corpus present 90.94% for the F1 score, the best reported result for Portuguese has only 71.23% in the same score. It is difficult to make comparisons between Portuguese NER due to the absence of standardized benchmarks [16]. Table 1 shows different settings by the authors to achieve their results.

Table 1. Reported results for different languages.

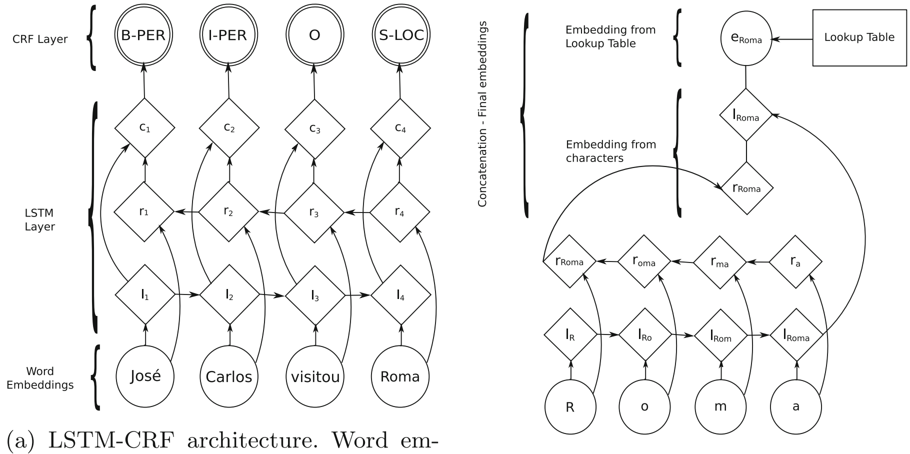
Author	Language	Corpora	Evaluation Script	Precision	Recall	F1
Amaral et al. [15]	Portuguese	train: HAREM I	SAHARA	83.48%*	44.35%*	57.92%*
		test: HAREM II				
Santos et al. [3]	Portuguese	train: HAREM I	CoNLL	73.98%**	68.68%**	71.23%**
		test: miniHAREM		67.16%*	63.74%*	65.41%*
	Spanish	SPA CoNLL-2002 Corpus		82.21%	82.21%	82.21%
Lample et al. [4]	English	CoNLL-2003 Corpus	CoNLL	*not shown*	*not shown*	90.94%
	Spanish	SPA CoNLL-2002 Corpus		*not shown*	*not shown*	85.75%

* Indicates the results for predicting all 10 categories from HAREM.

** Indicates the results for predicting 5 selected categories from HAREM.

3 LSTM-CRF Architecture

The LSTM-CRF architecture proposed by [4], as depicted in Fig. 1, is based on two intuitions: (i) Assigning tags for tokens in a text is based on contextual information, i.e., depends on other words and how they are related; (ii) In order to determine if a token is a name, it is important to consider both orthographic and distributional evidences. Orthographic evidences would be related to the shape of the word (the features that determine the appearance of the word), and distributional evidences would be related to the location in which the word



(a) LSTM-CRF architecture. Word embeddings are fed into to a bidirectional LSTM. l*i* represents the word *i* and its left context, r*i* represents the word *i* and its right context. The two vectors are concatenated, yielding a representation of the word *i* in its context, c*i*.

(b) The character embeddings of the word "Roma" are fed into a bidirectional LSTM. Their last outputs are concatenated to an embedding from a lookup table to obtain a representation for this word.

Fig. 1. Word and character level embeddings in the LSTM-CRF architecture. Adapted from [4].

tends to occur (the features that are related to neighboring words in sentences and in the corpus).

Recurrent neural networks (RNNs) represent a class of deep neural networks which are more suitable to handle sequential data, such as texts. Plain feed-forward networks, such as multi-layer perceptrons (MLP), and even CNNs, are limited in the sense that they take a fixed-size vector as input and produce a fixed-sized vector as output. RNNs, on the other hand, support sequences of vectors, being able to take inputs of variable sizes, also producing outputs of variable sizes. In theory, RNNs were conceived to capture long-term dependencies in large sequences, but, in practice, this was not possible due to the occurrence of vanishing and exploding gradient issues [17]. In order to overcome this limitation, [18] proposed the LSTM network, a type of RNN network in which the hidden units are enhanced with three multiplicative gates that control how the information is forgotten and propagated while flowing through each time step. These three gates are: update gate, forget gate and output gate. Equations (1) to (6) show the formulas used to update an LSTM unit at time t.

$$\mathbf{i}_t = \sigma(\mathbf{W}_i\mathbf{h}_{t-1} + \mathbf{U}_i\mathbf{x}_t + \mathbf{b}_i) \tag{1}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{h}_{t-1} + \mathbf{U}_f\mathbf{x}_t + \mathbf{b}_f) \tag{2}$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c\mathbf{h}_{t-1} + \mathbf{U}_c\mathbf{x}_t + \mathbf{b}_c) \tag{3}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \tag{4}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \mathbf{h}_{t-1} + \mathbf{U}_o \mathbf{x}_t + \mathbf{b}_o) \tag{5}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{6}$$

where $\mathbf{i}_t$ represents the update gate, $\mathbf{f_t}$ represents the forget gate and $\mathbf{o}_t$ represents the output gate, all three in a given time t. $\mathbf{c}_t$ and $\tilde{\mathbf{c}}_t$ represent the cell state and the candidate cell state of the LSTM unit, in a given time t. $\mathbf{W}$ stands for the weight matrices of the hidden state $\mathbf{h}$, $\mathbf{U}$ stands for the weight matrices of the input $\mathbf{x}$ and $\mathbf{b}$ stands for the bias vectors. σ represents element-wise sigmoid function and $\odot$ represents element-wise product.

Considering an input sentence represented by $\{\mathrm{x}_{\mathit{1}}, \mathrm{x}_{\mathit{2}}, \mathrm{x}_{\mathit{3}}, ..., \mathrm{x}_n\}$, with n words encoded as a d-dimensional vector, the bidirectional LSTM unit would calculate a hidden state $\overrightarrow{\mathbf{h}_t}$ and a hidden state $\overleftarrow{\mathbf{h}_t}$ for the left and right contexts of the sentence, at every word i, as depicted by Fig. 1a. The bidirectional aspect of the LSTM can be implemented by using a second LSTM unit that computes the right context by reading the same sentence in reverse. When the two hidden states are computed, they are concatenated into a single representation, $h_t = [\overrightarrow{\mathbf{h}_t};\overleftarrow{\mathbf{h}_t}]$.

Figure 1b shows how word embeddings are generated in the LSTM-CRF architecture. The character lookup table is initialized using a random uniform distribution, providing an embedding for every character found in the corpus. For each word in each input sentence, every character from the word is processed in direct and reverse order, using the embedding of the character from the lookup table and feeding it into a bidirectional LSTM unit. For each word, the character level embedding is resulting from the concatenation of the forward and backward representations from the LSTM unit, and this final character embedding is then concatenated with the word level embedding, obtained from the word embeddings used for training.

As for the sequential classification of the named entities, CRF is the algorithm used to predict the sequence of labels. It is a type of statistical modeling method which is often applied in pattern recognition and machine learning. When labeling sequences of words with CRF, the model provides a correlation understanding between words and labels which occur close to each other, i.e., it uses the label from surrounding words in order to determine the label of a given target word. As an example of NER labeling using the IOB2 tagging scheme [19], a word labeled with I-PESSOA could not follow a word labeled with O[2].

[2] This is because ***I*** indicates an internal token in the named entity, and ***O*** indicates a non-entity token, which means that anything after it would be the starting token of an entity or another non-entity token. Since the first token of a named entity starts with ***B***, according to the IOB scheme, it is not possible that an internal entity token follows a non-entity token.

4 Experimental Evaluation

4.1 Datasets

HAREM [20] is considered to be the main reference of corpora for the Portuguese NER task. It is a joint evaluation in the area of NER in Portuguese, intended to regulate and evaluate the success in the identification and classification of proper names in the Portuguese language. HAREM had two editions: HAREM I and HAREM II. MiniHAREM was an intermediate event that repeated the same evaluation as HAREM I. Each of these events produced a gold standard collection for the evaluation of NER systems. The corpora are annotated with ten named entity categories: Person (PESSOA), Organization (ORGANIZACAO), Location (LOCAL), Value (VALOR), Time (TEMPO), Abstraction (ABSTRACCAO), Title (OBRA), Event (ACONTECIMENTO), Thing (COISA) and Other (OUTRO). [3] experimented in two scenarios: total and selective. For the *total* one, all ten categories of HAREM are considered, while only five are considered in the *selective* scenario: Person, Organization, Location, Time and Value. We compare our results with the ones obtained by [3], for both total and selective scenarios. As for model evaluation, we use the same CoNLL script from [3,4]. We use the same HAREM corpora for training and testing datasets that was used by [3]. The gold standard collection from HAREM I was used as the training set, and the gold standard collection from MiniHAREM was used as the test set.

4.2 Parameterization, Training and Experimental Setup

For tagging schemes, we experimented with two different IOB tagging schemes: IOB2 [19] and IOBES. IOBES differs from IOB2 because it labels single-token entities with the ***S*** prefix, and also labels the final tokens from multi-token entities with the ***E*** prefix. Regarding word embeddings, we experimented with four different pre-trained word embeddings from [9]: FastText [10,11], Glove [12], Wang2Vec [13] and Word2Vec [14], all of them with dimension 100. As mentioned in [3,4], we picked the FastText, Wang2Vec and Word2Vec embeddings that were trained with the skip-gram model. The training process of these word-embeddings are described in [21].

Besides the character and word level embeddings, we also experimented with a capitalization feature. This feature is a representation of the word capitalization: 0 if all characters are lowercase, 1 if all characters are uppercase, 2 if the first letter is uppercase and 3 if a letter besides the first is uppercase. In addition to the capitalization feature, we also experimented normalizing the words before producing the dictionaries that are used to perform the word-embedding lookup. This normalization is nothing more than converting the word to its lowercase form, and does not affect the data structures used to learn character-level features.

This architecture uses two bidirectional LSTMs, one for learning the features from the character-level embeddings, and one for the word-level representations. Despite [4] observing that the increase of the number of hidden units for each

LSTM did not have a significant impact on the model's performance, we have experimented with two different dimensions for each. [4] used 25 hidden units for each character LSTM, the forward and the backward, and we have experimented with 25 and 50 hidden units. For the word LSTMs, [4] used 100 hidden units, and we have experimented with 100 and 200 units.

The model is trained using the backpropagation algorithm, and optimization is done using stochastic gradient descent (SGD), with a learning rate of 0.01 and a gradient clipping of 5.0. [4] experienced with other optimization methods, but none performed better than SGD with gradient clipping. In order to determine the best set of parameter values to be used in our experiment, from the ones mentioned in this section, first we picked 6 parameters: tagging scheme, word embedding, capitalization feature, word normalization and dimension of character and word LSTM units. From all the possible values for each of these parameters, there are 128 different combinations to be evaluated. We ran each of the combinations 10 times, in the selective scenario, with only 5 epochs, which we considered as sufficient for determining the best set of parameters. Once we determined the best set of parameters, we trained the model for 100 epochs, using the parameter values obtained from the previous step. We also trained with these parameters 10 times, in order to estimate an average of the model's performance.

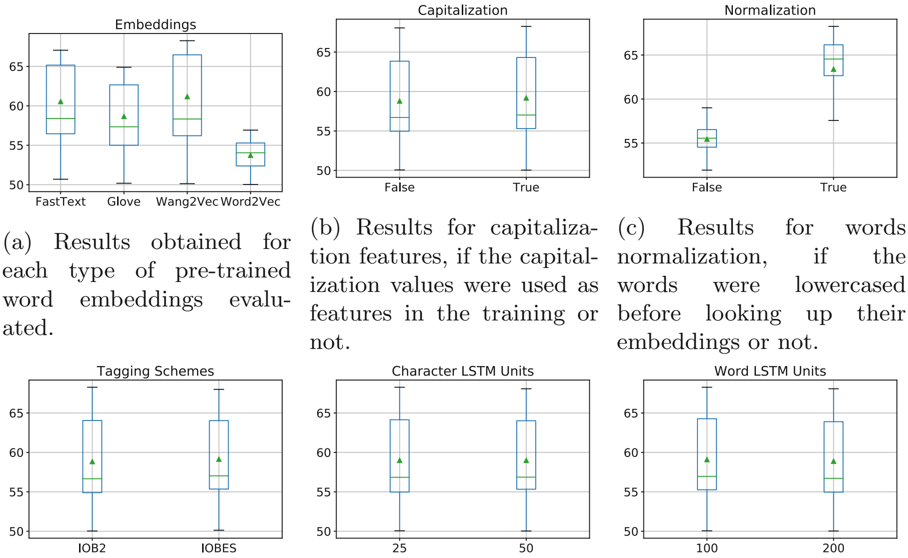
4.3 Results

Figure 2 contains boxplots with the comparisons between each set of parameter values assessed in our trainings. We realized that the ones that had the greatest impact in our training were embedding type and word normalization, while the different values assessed for tagging scheme, capitalization and hidden units dimensions did not have a considerable impact in the results. Figure 2a indicates that Wang2Vec embeddings outperformed Word2Vec, Glove and FastText, with a mean F1 score of 61.17. FastText, which is ranked second, had a mean F1 score of 60.54, while Glove scored 58.65 and Word2Vec scored 53.72.

The normalization of words was the most significant parameter that we experimented with. Keeping the words as they are, without normalization, provided a mean score of 55.78, while normalizing the words to lower case form gave a mean score of 64.47. We realized that the normalization had such a great impact because of the pre-trained word embeddings used for NER training. All words contained in the embeddings were only presented in their lowercase form, so whenever a lookup was performed in the embeddings table, if the word started with an uppercase letter, it would not be found, and a random vector would be initialized to it. So, performing the normalization prior to the lookup enforces the use of the proper word vectors for NER training.

From the results obtained after running all combinations of parameters values, we verified that the optimal combination for training a final model would be: Wang2Vec as pre-trained word embeddings, IOBES tagging scheme, normalization of words, use of capitalization features, 25 hidden units for the character

LSTM and 100 hidden units for word LSTM. Despite not having a considerable difference between the results obtained for the different values of tagging schemes, capitalization features and hidden units dimension for both character and word LSTMs, the choice of values for these parameters were due to their results being slightly higher than the other evaluated options. Table 2 displays the comparison between the results from [3] and the ones obtained in our final training, using the tuned hyperparameters. LSTM-CRF outperformed CharWNN in both total and selective scenarios, improving the F1 score in both total and selective scenarios by 5 points.



(a) Results obtained for each type of pre-trained word embeddings evaluated.

(b) Results for capitalization features, if the capitalization values were used as features in the training or not.

(c) Results for words normalization, if the words were lowercased before looking up their embeddings or not.

(d) Results obtained for IOB2 and IOBES tagging schemes.

(e) Results obtained for 25 and 50 units for character LSTM units.

(f) Results obtained for 100 and 200 units for word LSTM units.

Fig. 2. Results obtained for each set of parameters values evaluated. Each boxplot depicts the data related to the F1 score obtained for each of the 1280 executions, grouped by the set of parameter values displayed in each of them. The green triangles represent the arithmetic means of the F1 scores obtained. (Color figure online)

Table 2. Comparison with the state-of-the-art for the HAREM I corpus

Architecture	Total scenario			Selective scenario		
	Precision	Recall	F1	Precision	Recall	F1
CharWNN	67.16%	63.74%	65.41%	73.98%	68.68%	71.23%
LSTM-CRF	**72.78%**	**68.03%**	**70.33%**	**78.26%**	**74.39%**	**76.27%**

5 Conclusions

In this paper, we experimented different scenarios of Named Entity Recognition with Portuguese corpora, using a deep neural network architecture based on bidirectional LSTM and Conditional Random Fields. We evaluated different combinations of hyperparameters for training, and verified the optimal values for the parameters that had a greatest impact in the performance of the model: word embeddings model and word normalization. We achieve state-of-the-art performance for Portuguese NER task using the optimal values for these parameters.

The word embedding model that had the best performance in our experiments was Wang2Vec, which is a method derived from modifications in Word2Vec. The purpose of these modifications was to improve the capture of the syntactic behavior of words, taking into consideration the order in which they appear in the texts. We verify that this improves the performance of a sequence labeling task such as NER. We also verify that normalizing words before looking up their embeddings greatly improves the performance of the model, opposed to looking up the embeddings according to the letter case they are in the text.

For future work, we will experiment on the effects of applying this NER model in texts belonging to a specific domain, instead of a general purpose corpora such as the ones based on news and wikipedia articles.

References

1. How Much Data is Created on the Internet Each Day? https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/. Accessed 19 Mar 2018
2. Maynard, D., Bontcheva, K., Augenstein, I.: Natural Language Processing for the Semantic Web, 1st edn. Morgan and Claypool, San Rafael (2017)
3. dos Santos, C., Guimarães, V.: Boosting named entity recognition with neural character embeddings. arXiv preprint arXiv:1505.05008 (2015)
4. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
5. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. arXiv preprint arxiv:1103.0398 (2011)
6. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from Wikipedia. In: Artificial Intelligence, vol. 194, pp. 151–175. Elsevier Science Publishers Ltd., Essex (2013). https://doi.org/10.1016/j.artint.2012.03.006
7. Chiu, J., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. arXiv preprint arXiv:1511.08308 (2015)
8. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. arXiv preprint arXiv:1603.01354 (2016)
9. Repositório de Word Embeddings do NILC. http://www.nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc. Accessed 30 Mar 2018
10. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)

11. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
12. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014), vol. 12, pp. 1532–1543 (2014)
13. Ling, W., Dyer, C., Black, A., Trancoso, I.: Two/too simple adaptations of word2vec for syntax problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2015)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arxiv:1301.3781 (2013)
15. Amaral, D., Vieira, R.: NERP-CRF: a tool for the named entity recognition using conditional random fields. In: Linguamática, vol. 6, pp. 41–49 (2014)
16. Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J.: Named entity recognition: fallacies, challenges and opportunities. Comput. Stand. Interfaces **35**, 482–489 (2013)
17. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**, 157–166 (1994). https://doi.org/10.1109/72.279181
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
19. Sang, E., Veenstra, J.: Representing text chunks. arXiv preprint arxiv:cs/9907006 (1999)
20. HAREM: Reconhecimento de entidades mencionadas em português. https://www.linguateca.pt/HAREM/. Accessed 21 Mar 2018
21. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: evaluating on word analogies and natural language tasks. arXiv preprint arXiv:1708.06025 (2017)

Disambiguating Open IE: Identifying Semantic Similarity in Relation Extraction by Word Embeddings

Leandro M. P. Sanches(✉), Victor S. Cardel, Larissa S. Machado,
Marlo Souza, and Lais N. Salvador

Federal University of Bahia, Salvador, Bahia, Brazil
leandrompsanches@gmail.com, vscardel@gmail.com, machad.lari@gmail.com,
{msouza1,laisns}@ufba.br

Abstract. Open Information Extraction (Open IE) methods enable the extraction of structured relations from domain-independent unstructured sources. However, due to lexical variation and polysemy, we argue it is necessary to understand the meaning of an extracted relation, rather than just extracting its textual structure. In the present work, we investigate different methods for associating relations extracted by Open IE systems with the semantic relations they describe by using word embedding models. The results presented in our experiments indicate that the methods are ill-suited for this problem and show that there is still a lot to research on the Relation Disambiguation in Portuguese.

Keywords: Relation Disambiguation · Open Information Extraction
Semantic relations

1 Introduction

Information Extraction (IE) is the area that studies methods for obtaining structured information from unstructured textual sources. According to Etzioni [3], the problem of IE "is the task of recognizing the assertion of a particular relationship between two or more entities in text". Information Extraction is useful to methods of question answering [3] and many other applications.

IE methods can be divided into Traditional IE, which focus on extracting instances of a set of predefined relations in a domain, and Open IE methods, which aims to identify relations in a domain-independent scalable manner [3,11,25,26].

While the Open IE paradigm attacked scalability and domain-independence problems in IE methods, it has introduced problems of its own. Firstly, there is the problem of lexical variation, i.e. the same relation can be expressed by several different textual descriptions. Also, the same superficial structure can be used to express different meanings, a problem known as *polysemy*.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 93–103, 2018.
https://doi.org/10.1007/978-3-319-99722-3_10

Due to these problems, it is arguable that the Open IE paradigm has achieved its promises of automating IE. Particularly, it is not clear whether Open IE methods can be applied in the context of Web as a Corpus, as proposed by Banko et al. [3].

In this sense, for these methods to be applicable for IE, we argue it is necessary to understand the meaning of an extracted relation, rather than just presenting its textual structure. In other words, we need to associate relation instances extracted by Open IE systems with the semantic relations they describe, when one such relation is known - a problem we call Relation Disambiguation.

More formally, given a set of semantic relations and a relational fact $r_1 = (e_1, rel, e_2)$, we call Relation Disambiguation the process of identifying which semantic relation expresses the same information as expressed in the fact r_1, i.e. for which relation fact r_1 is an instance.

In this work, we will investigate methods for Relation Disambiguation for Portuguese. To achieve this, we will use word embedding models to identify semantic similarity between relation descriptors obtained using an Open IE system and semantic relations described in a domain ontology.

This work is structured as follows: Sect. 2 describes the problem of lexical variation in Open IE and the task of Relation Disambiguation, which we tackle in this work; Sect. 3 presents the related work; Sect. 4 describes the methods studied in this work and Sect. 5 presents the empirical evaluation of theses methods, while Sect. 6 presents our discussions on these results. Finally, we present our final considerations.

2 IE and Open IE

Information Extraction is the area that studies methods for obtaining structured information from unstructured sources. Traditional IE methods rely on manually crafted identification rules and/or supervised machine learning algorithms to identify relation instances in the text. These methods require extensive human involvement in crafting such rules and annotating large amounts of data. As such, these methods are not easily adaptable to other domains nor are scalable to large amounts of data. To overcome such difficulties, unsupervised methods for information extraction were devised, aiming to identify relation instances within text without previous knowledge of the domain [3,11,26].

While many methods focus on discovering shallow or syntactic patterns to extract relations from text, specially in those works following the so-called Open IE paradigm, few have been concerned with identifying the semantics of such extractions. This approach, however, have some important limitations, in our opinion.

First, the same information may be expressed in multiple ways, making it hard for one to be able to retrieve all desired information when querying a base constructed over such extractions. For example, consider the following text fragment:

> Sócrates (470 a.C-399 a.C.) foi um importante filósofo grego do segundo período da filosofia. Nasceu em Atenas (...) Tales de Mileto (624 a.C.-548 a.C.), nascido na cidade de Mileto (...) Anaximandro de Mileto (610 a.C.-547 a.C.) foi um discípulo de Tales original de Mileto (...) [21][1]

From this fragment, we can identify three different textual descriptions for the same information of one philosopher being born in a city: "nasceu" (was born in), "nascido na cidade de" (born in the city of), "original de" (originated from).

Also, the same superficial structure can be used to express different meanings, e.g., *"A plays B"* means different things whether A refers to a musician, to an athlete, or to an actor - a problem known as *polysemy*.

Since most Open IE systems perform no semantic analysis, the semantic similarity between the extracted relations cannot be detected by them. Consequently, the actual result of most of Open IE extraction differs very little from that of a shallow parser that identifies Subject-Verb-Object structures in sentences, reproducing the surface structure of the sentence, not its semantic content.

In this work, we argue that recognizing semantic similarity can be of great use for relation extraction, allowing for the extraction of semantic information from an unstructured source.

3 Related Work

To the best of our knowledge, no methods have been proposed in the literature for the task of Relation Disambiguation for Portuguese. For other languages, works such as that of Lassen et al. [14] and Jacquemin et al. [12] aim to perform semantic annotation of texts by object properties described in given ontology.

Lassen et al. [14] is an example of Traditional IE method and applies machine learning to identify semantic relations within text. While Jacquemin et al. [12] propose using word sense disambiguation, domain information and lexical semantic resources to perform semantic disambiguation. The first method requires manual annotation of the types of relations identified in the text and, as such, presents difficulties regarding domain independence and scalability. The second method, on the other hand, relies on the semantic relations present in lexical ontologies to tackle lexical variation in Open IE. Since lexical ontologies are not available for many languages and for all domains, we believe its applicability to Open IE Relation Disambiguation is rather limited.

Works on ontology alignment, automatic ontology enriching or relational discovery are also related to our problem, since they aim to identify semantic correspondences in relations described in different ways. Regarding ontology

[1] Socrates (470 B.C.E. - 399 B.C.E.) was an important greek philosopher of Philosophy's second period. He was born in Athens (...) Thales of Miletus (624 B.C.E. - 548 B.C.E.) was born in the city of Miletus (...) Anaximander of Miletus (610 B.C.E. - 547 B.C.E.) was a disciple of Thales originated from Miletus.

alignment, methods based on both intensional and extensional correspondence are applicable to our problem. For example, the work of Van Diggelen et al. [23] propose a protocol to compute logical correspondences of relations between two terminologies, based on their extension, while works such as [9] propose the use of lexical resources to identify similarities between the elements of different ontologies. Notice that while these methods are relevant to our problem, semantic resources such as ontologies are much more "well-behaved" then extractions from unstructured sources and, as such, these methods can be very difficult to adapt to our problem.

Works on discovering new ontological relations using information extracted from textual sources [4,17,20] are also related to our problem. While their focus is identifying new relations to enrich an ontology, we believe they can also be used for Relation Disambiguation and to reduce *lexical variance* in Open IE systems.

In relation to the *Open IE* systems, to the best of out knowledge, there are several available systems proposed in the literature, but only a few of which proposed to process texts in the portuguese language. Among those, we highlight the systems RePort [19], ArgOE [7], DepOE [8] and DependentIE [18].

Notice that the literature in word sense disambiguation is also interesting to our study, but it is too numerous to analyse here and only tangentially related. Even though it is true that word sense disambiguation can be used to improve the performance of the Relation Disambiguation, the problems related to each task are not the same.

4 Relation Disambiguation

In this work, given a set of relation extractions and a set of previously known semantic relations, we study how to identify which relation extractions are instances of each known semantic relation. To do this, we will explore the use of word embedding models - namely Word2Vec [15] models - to identify semantic similarity between relation descriptors. In this work, we study three methods of Relation Disambiguation.

The first method studied by us, our baseline, consists of directly computing the semantic similarity between each relation descriptor rel in an extraction $t = (e_1, rel, e_2)$ and (the name of) each semantic relation (object and data properties [24]) in the ontology. The tuple t is disambiguated as an instance of the relation r for which it has the highest semantic similarity, as long as this measure is above some empirically defined confidence threshold, called disambiguation threshold.

As a metric of semantic similarity, we use cosine similarity between these two sets, defined by Eq. 1 [20], in which T_i is a set of words, $|T_i|$ is the number of words in T_i and $maxSim(w, T_i)$ is the maximum similarity value between a word w in T_i and some word in $T_{i'}$, based on a Word2Vec model.

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \in T_1}(maxSim(w, T_2))}{|T_1|} + \frac{\sum_{w \in T_2}(maxSim(w, T_1))}{|T_2|}\right) \quad (1)$$

This baseline is a simplistic proposal, since it assumes that the surface structure of a relation descriptor describing some semantic relation must be more related (in meaning) to this relation than to any other. Moreover, since the semantic similarity metric is computed based purely on the words that compose a descriptor and not its structure, and due to the substantial variation in the form in which an information can be expressed via text, we believe this strategy may not be able to identify all instances of a relation.

To overcome this problem, based on the work of Subhashree and Kumar [20] on discovering new semantic relations, we propose the second method, which consists of grouping the relation descriptors into clusters based on semantic similarity and disambiguating each cluster to the known semantic relations. Performing such clustering of semantic descriptors, we believe, may reduce the impact of descriptor variability in the disambiguation process, creating a more accurate representation of the meaning of relation expressed by that cluster.

To perform the clustering, the relation descriptors are filtered according to their adequacy of the relation descriptor to the domain of discourse - as described by the ontology. This adequacy is established by calculating the word-to-word similarity between the relation descriptor and the main concepts of the domain, using the Word2Vec model and cosine similarity. Any relation descriptor with adequacy below a significance threshold is discarded from the set.

Subsequently, the clustering algorithm uses the semantic similarity measure described in Eq. 1 to decide in which cluster a given relation descriptor should be included. If the similarity between a given relation descriptor and every cluster is below a given clustering threshold, a new cluster is created with this relation. The disambiguation of the entire cluster as a semantic relation is, then, performed as before by taking the maximal similarity between any descriptor in the cluster and the known semantic relations.

Notice that our second method uses only intensional information for relation clustering and disambiguation. However, semantic relations also possess extensional information which could be used in the disambiguation process. As such, we developed our third method which relies on extensional information for clustering and, as it has been done before, intensional information for disambiguation. So in this approach, relation descriptors are clustered based on the frequency of arguments co-occurrence between the two descriptors using the K-Means algorithm. Our third method is an adaptation of Mohamed et al.'s [16] method for discovering new relations.

5 Evaluation

In this section, we describe the empirical validation of our methods, the data used and the results achieved, considering different values for the disambiguation and clustering thresholds. Our experiments were performed on the domain of contemporary art using the Contemporary Art Ontology [22]. The Contemporary Art Ontology is a domain ontology constructed in the portuguese language composed of 149 classes, 18 object properties and 14 data properties [22]. We present some examples of the object properties of the ontology (in translated form) in Table 1.

Table 1. Examples of object properties of the contemporary art ontology

Property	Domain	Range
act_as	Person	Profession
born_in	Person	Geographic_Object
produced_by	Event or Work_of_Art or Registry	Organization or Person
author_acting_as	Collection or Event or Work_of_Art	Profession
conceived_at	Work_of_Art	Geographic_Object

The choice of this ontology was due to the fact that it was previously applied for semantic annotation of multimedia contents [22]. Since we believe automatic semantic annotation is a natural application for Open IE methods, we believe that such an ontology - and thus domain - would provide an interesting case of study to our methods.

5.1 The Data

In our experiments, we used as input for the Open IE tool a corpus consisted of 370 Wikipedia articles in Portuguese from the domain of contemporary art. To construct this corpus, we retrieved articles within 34 Wikipedia categories manually selected for the domain, including information about artists, painters, writers, artworks and contemporary architecture. The articles were further cleaned removing hyperlinks, tables, lists and Wikipedia structure using the WikiExtractor tool [2].

This input corpus was subsequently processed by Gamallo and Garcia's multilingual Open IE tool ArgOE [7]. The choice of ArgOE was due to the fact that it was the only tool for the portuguese language for which the source code was readily available, as far we know. From the relations extracted by ArgOE, we discarded all extractions missing arguments, resulting in 8370 extracted relation triples.

For the evaluation, 110 triples out of the 8370 triples were randomly selected and distributed to four human annotators, who performed manual disambiguation of the relations according to the target ontology. Each relation triple could be classified as an instance of a relation (object or data property) in the ontology or not having any equivalent representation in it.

Each annotator received a set of 35 triples to annotate, with 10 triples in common between all annotators. The 10 common triples were used compute inter-annotator agreement, while the remaining 100 were used to evaluate other methods proposed in Sect. 4. In order to establish a consensus among the annotators about the annotation process, we performed a training of the human annotators in which we discussed the phenomenon to be analyzed, explained the structure of the ontology and discussed how to perform the disambiguation.

As previously mentioned, to identify semantic similarity between relation descriptors we explored the use of Word2Vec [15] models. Our Word2Vec model was trained over the corpus of Brazilian Portuguese journalistic texts CETENFOLHA[2], which has more than 25 thousand words.

We also trained a second Word2Vec model over the CETENFOLHA corpus and our Wikipedia articles corpus, but the obtained model performed much poorly than the one trained exclusively on the CETENFOLHA corpus. We believe that differences in the tokenization model and lexicon between the sources introduced noise in the model, explaining thus its poorer performance. As such, in this work, we choose to use the model trained only on the CETENFOLHA corpus.

5.2 Methodology

The three methods described in Sect. 4 were implemented in Java, using Apache Jena [1] to access the ontology, the Weka workbench [6] implementation of the K-Means clustering and Deeplearning4j library [5] implementation of the Word2Vec training algorithm with Skipgram and 200 dimensions [15].

The evaluation was performed using accuracy (A), precision (P), recall (R) and F-measure (F1) metrics calculated by comparing relations disambiguated by each method with the ones disambiguated by human experts. In this evaluation, we varied the disambiguation and clustering threshold values in order to better understand how these values impact the quality of the result of each method.

In this work, we consider accuracy as the ratio of agreement between the predictions by the method and the human annotators among all annotations, i.e. including those relations which have not been disambiguated as a semantic relation of the ontology. On the other hand, precision is computed as the ratio between the relation descriptors which have been correctly disambiguated as a semantic relation and all the triples which have been disambiguated by the system. Finally, recall is computed by the ratio between the relations which have been correctly disambiguated by the system and the amount of the relations disambiguated as an instance of a semantic relation in the ontology by the human annotators.

5.3 Results

Regarding the manual disambiguation of triples performed by the human judges, just a total of 11 triples out of the set of 100 randomly selected relations were successful associated to a relation in the ontology by human annotators. From the 10 extractions in common between the 4 annotators, we obtained a value for Fleiss' Kappa coefficient [13] of 0.52, indicating an overall moderate inter-annotator agreement.

We evaluated all three methods varying the value of the disambiguation threshold from the values 0.1 to 0.9. The results are shown in the Table 2 where a clustering threshold of 0.6 and a domain adequacy threshold of 0.35 were used for the second method, while the number of clusters was set as $K = 40$ for the third method, a number superior to that of relations in the ontology.

[2] https://www.linguateca.pt/cetenfolha/.

Table 2. Results of the evaluation for the three methods

T	First method				Second method				Third method			
	P	R	F1	A	P	R	F1	A	P	R	F1	A
0.1	0.0004	**0.2727**	0.0007	0.09	0	0	0	0.84	0.0002	0.0909	0.0003	0.02
0.2	0.0004	**0.2727**	0.0008	0.24	0	0	0	0.84	0.0002	0.0909	0.0003	0.02
0.3	0.0007	**0.2727**	0.0013	0.48	0	0	0	0.84	0.0002	0.0909	0.0003	0.02
0.4	0.0010	**0.2727**	0.002	0.62	0	0	0	0.85	0.0002	0.0909	0.0003	0.02
0.5	**0.017**	**0.2727**	**0.034**	0.71	0	0	0	0.87	0.0002	0.0909	0.0003	0.02
0.6	0.0041	0.1818	0.008	0.85	0	0	0	0.88	0.0002	0.0909	0.0003	0.02
0.7	0.0137	0.1818	0.0255	0.87	0	0	0	**0.89**	0.0002	0.0909	0.0003	0.03
0.8	0	0	0	0.88	0	0	0	**0.89**	0.0002	0.0909	0.0003	0.03
0.9	0	0	0	**0.89**	0	0	0	**0.89**	0.0002	0.4909	0.0003	0.03

As shown in Table 2, the first method achieved better F1-score. However, with the thresholds above 0.7, the threshold becomes very restrictive and the method cannot disambiguate the relations any longer.

In both the first and the second methods, the accuracy increases with higher disambiguation thresholds. This happens due to the fact that 89% of the triples in the evaluation dataset were not manually associated to any ontology property. As such, higher values for the disambiguation threshold means less disambiguated triples, which reduces the chances of errors by the system, thus elevating its accuracy.

We also evaluated the impact of the clustering threshold in the second method. The experiments, however, indicated no impact in either Precision and Recall, and only limited impact on Accuracy - obtaining a value of 0.85 for a clustering threshold of 0.10 and a value of 0.89 for a clustering threshold of 0.9.

6 Discussion

The results presented in this work indicate that the methods may be ill-suited for the problem of Relation Disambiguation. Notice that, while some of the methods achieved high accuracy in our experiments, this result is due to the extreme imbalance on the experimental data - which is expected from the fact that Open IE aims at extracting all possible semantic relations from text, not only a limited number restricted to a domain.

One possible reason for the poor performance of the methods lie in the word embedding model used in this work. Notice that our model was trained using Word2Vec algorithm using skipgram with only 200 dimensions over a relatively small corpus. We chose a relatively small dimensionality for the representation space due to limits in our computational resources and due to the high cost of processing a larger model.

A common problem also concerns the extractions made by the preprocessing tool. The ArgOE extraction performs poorly in the chosen domain, mainly due to problem in the syntactic analysis of the texts, e.g the triple: *("a provocação", "passa as nossas reações pela", "exigência")*[3] was extracted from the sentence *"A provocação nas obras de Graça Martins passa pela exigência de tornarmos consequentes as nossas reações"*[4].

Different to the first two methods, the third method's poor results was also probably caused by the fact that, due to the arguments have not been also disambiguated into entities, few or no co-occurrences were found between the majority of the relations - which prevents the clustering algorithm of finding any useful information to create suitable clusters.

7 Final Considerations

In this work, we tackle the problem of Relation Disambiguation for Open IE systems for the portuguese language. To do this, we implemented three methods of identifying when a relational descriptor in a set of Open IE extractions possess the same semantic information as a previously known semantic relation.

Our methods consider different approaches to accomplish this disambiguation. The first and second methods have intensional approaches applying cosine similarity calculation through a Word2Vec model and the third method makes use of an extensional approach constructing descriptors co-occurrence matrices.

Another difference is related to the way to carry out the disambiguation, the second and third method use clusters to group similar descriptors with posteriori descriptors disambiguation. In contrast, the first method performs the disambiguation of relations individually.

As shown in the experiments, all these methods did not present satisfactory results according to F-measure, showing that little or no information has been disambiguated correctly. The poor results show that Relation Disambiguation is not an easy problem to solve and there is still a lot to research to be done on this topic. Particularly, for the portuguese language, for which there is a limited number of good IE and Open IE systems available, obtaining good quality Relation Disambiguation methods seems essential to improve results in areas such as question answering.

As an immediate future work, we aim to evaluate the effect of different high-quality word embedding models for the Portuguese language in the studied disambiguation methods, as studied by Hartman et al. [10].

Acknowledgements. This study was partially funded by *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)* and by *Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB).*

[3] ("The provocation", "goes through our reactions to the", "demand").

[4] *"The provocation in Grace Martins's works goes through the demand to make our reactions consistent"*, in English.

References

1. Apache Software Foundation: Apache jena: a free and open source java framework for building semantic web and linked data applications (2018). https://jena.apache.org/
2. Attardi, G.: Wikipedia extractor (2016). http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
3. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence, Hyderabad, pp. 2670–2676 (2007)
4. Barchi, P.H., Hruschka, E.R.: Never-ending ontology extension through machine reading. In: Hybrid Intelligent Systems (HIS), pp. 266–272. IEEE (2014). https://doi.org/10.1109/HIS.2014.7086210
5. Deeplearning4j DT: Deeplearning4j: open-source distributed deep learning for the JVM (2017). http://deeplearning4j.org
6. Frank, E., Hall, M.A., Witten, I.H.: The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, 4th edn. Morgan Kaufmann, Online (2016). https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf
7. Gamallo, P., Garcia, M.: Multilingual open information extraction. In: Pereira, F., Machado, P., Costa, E., Cardoso, A. (eds.) EPIA 2015. LNCS (LNAI), vol. 9273, pp. 711–722. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23485-4_72
8. Gamallo, P., Garcia, M., Fernandez-Lanza, S.: Dependency-based open information extraction. In: Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, pp. 10–18. Association for Computational Linguistics, Stroudsburg (2012)
9. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: S-match: an algorithm and an implementation of semantic matching. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 61–75. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-25956-5_5
10. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., Aluísio, S.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology, pp. 122–131 (2017)
11. Hasegawa, T., Sekine, S., Grishman, R.: Discovering relations among name dentities from large corpora. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 415. Association for Computational Linguistics (2004)
12. Jacquemin, B., Brun, C., Roux, C.: Enriching a text by semantic disambiguation for information extraction. In: Proceeding of the Workshop on Using Semantics for Information Retrieval and Filtering: State of the Art and Future Research (LREC 2002), Las Palmas, Canary Islands, Spain, pp. 45–51 (2002). http://arxiv.org/abs/cs/0506048
13. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977). http://www.jstor.org/stable/2529310
14. Lassen, T., Terney, T.V.: An ontology-based approach to disambiguation of semantic relations. In: Proceedings of the Workshop on Learning Structured Information in Natural Language Applications (2006)

15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
16. Mohamed, T.P., Hruschka Jr., E.R., Mitchell, T.M.: Discovering relations between noun categories. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, pp. 1447–1455. Association for Computational Linguistics, Stroudsburg (2011)
17. Nimishakavi, M., Singh, U.S., Talukdar, P.: Relation schema induction using tensor factorization with side information. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 414–423. Association for Computational Linguistics, Austin (2016). https://aclweb.org/anthology/D16-1040
18. Oliveira, L.S., Glauber, R., Claro, D.B.: Dependentie: An open information extraction system on portuguese by a dependence analysis. In: Encontro Nacional de Inteligência Artificial e Inteligência Computacional. Sociedade Brasileira de Computação (SBC), Uberlandia (2017)
19. Pereira, V., Pinheiro, V.: Report-um sistema de extração de informações aberta para língua portuguesa. In: Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology, pp. 191–200. Sociedade Brasileira de Computação, Natal (2015)
20. Subhashree, S., Kumar, P.S.: Enriching linked datasets with new object properties. CoRR abs/1606.07572 (2016). http://arxiv.org/abs/1606.07572
21. TODA MATÉRIA: Filósofos pré-socráticos.https://www.todamateria.com.br/filosofos-pre-socraticos/. Accessed 6 Apr 2018
22. Trillo, C.D.P.: Recuperação de vídeos indexados por conceitos. Master's thesis, Universidade de São Paulo, São Paulo, March 2005. https://www.ime.usp.br/~rmcobe/onair/files/christian_thesis.pdf
23. Van Diggelen, J., Beun, R.J., Dignum, F., Van Eijk, R.M., Meyer, J.J.: Ontology negotiation: goals, requirements and implementation. Int. J. Agent-Oriented Softw. Eng. **1**(1), 63–90 (2007)
24. W3C: Owl web ontology language overview (2004). http://www.w3.org/TR/2004/REC-owl-features-20040210/
25. Wu, F., Weld, D.S.: Open information extraction using wikipedia. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 118–127. Association for Computational Linguistics, Stroudsburg (2010). http://dl.acm.org/citation.cfm?id=1858681.1858694
26. Xavier, C.C., Lima, V.L.S., Souza, M.: Open information extraction based on lexical semantics. J. Braz. Comput. Soc. **21**(4) (2015). https://doi.org/10.1186/s13173-015-0023-2

Natural Language Processing Applications

Personality Recognition from Facebook Text

Barbara Barbosa Claudino da Silva and Ivandré Paraboni(✉)

School of Arts, Sciences and Humanities,
University of São Paulo, São Paulo, Brazil
{barbara.barbosa.silva,ivandre}@usp.br

Abstract. This work concerns a study in the Natural Language Processing field aiming to recognise personality traits in Portuguese written text. To this end, we first built a corpus of Facebook status updates labelled with the personality traits of their authors, from which we trained a number of computational models of personality recognition. The models include a range of alternatives ranging from a standard approach relying on lexical knowledge from the LIWC dictionary and others, to purely text-based methods such as bag of words, word embeddings and others. Results suggest that word embedding models slightly outperform the alternatives under consideration, with the advantage of not requiring any language-specific lexical resources.

Keywords: Big Five · Personality recognition

1 Introduction

The increasing complexity of computer systems has been accompanied by the development of ever more sophisticated human-machine communication methods. Current systems are capable of interpreting and reproducing a wide range of human behaviour, including emotions and feelings. These temporary manifestations of character are however heavily influenced by a stable set of patterns of human behaviour that are largely foreseeable. These patterns - or traits - constitutes what is generally understood as human personality [1].

The computational recognition of human personality is at the heart of the design of the so-called intelligent systems, and will be the focus of the present work as well. Fundamental personality traits may be recognised through a range of methods proposed in the Psychology field. Among these, the most popular are those based on the lexical hypothesis, which establishes that personality traits are observable in the words that we use to communicate. This approach has been refined from an initial survey of 4,500 traits identified in the 1930s to produce, independently and simultaneously in several studies, a stable framework known as the Big Five model of human personality [2].

The Big Five model comprises five key dimensions (or traits) - Openness to experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism -

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 107–114, 2018.
https://doi.org/10.1007/978-3-319-99722-3_11

which are widely accepted as an adequate basis for the representation of human personality [3]. From a computational perspective, the Big Five model is central to human-computer interaction studies in general and, given its linguistic motivation, it makes also a suitable theoretical basis for Natural Language Processing (NLP) research. Knowing an individual's personality traits (e.g., from their social network updates) enables the production of personalised content in many ways, including the presentation of more appealing website or the generation of targeted advertisement, among many other possible applications.

Based on these observations, this paper presents a study of automatic Big Five personality recognition for the Brazilian Portuguese language. More specifically, we built a corpus of Facebook text labelled with personality information, and designed a number of experiments involving alternative text representations and supervised machine learning methods to recognise each of the Big Five traits. In doing so, our goal was to determine which representation and method would provide best results for our target language and domain.

The reminder of this article is organised as follows. Section 2 introduces a number of basics concepts related to the Big Five model and personality inventories, and briefly discusses the related work on Big Five recognition from text. Section 3 presents our current work, comprising the corpus construction and the experiments that were conducted. Section 4 presents our results, and Sect. 5 draws a number of conclusions and discusses further studies.

2 Related Work

The Big Five personality model [2] comprises five fundamental dimensions of human personality: Openness to experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism. Each of the five dimensions is modelled as a scalar value representing the degree to which an individual expresses a given personality trait or not. Thus, for example, a high value for Extraversion indicates an extrovert individual, whereas a low value for this dimension indicates an introvert.

Big Five personality dimensions may be estimated by many well-known methods proposed in the Psychology field, being the most popular the use of inventories of personality such as the 44-item Big Five inventory - or BFI - that has become popular in Computer Science studies as well. The BFI was originally developed for the English language, but it was subsequently replicated in dozens of other languages, including Brazilian Portuguese. In particular, the study in [3] validated a Brazilian Portuguese version of the BFI called IGFP-5 by presenting a factorial analysis involving a sample of 5,089 respondents from the five regions of Brazil. This inventory will also be adopted in the present study, as discussed in Sect. 3.

The computational recognition of personality from text tends to follow a traditional methodology of supervised [4] or semi-supervised [5] machine learning. The task may be modelled as a classification problem (e.g., deciding whether an individual is a introvert or not), as a regression problem (e.g., determining the scalar value of a dimension of personality), or as a ranking problem (e.g., ordering a set of individuals according to a dimension of interest.)

One of the first large-scale initiatives to recognise personality traits for the English language was the work in [6]. This consisted of an experiment involving 2,263 essays written by 1,200 students who completed a personality inventory, but it was limited to the lower and upper ends of the scale for Extraversion and Neuroticism. Essay words were grouped into four categories of well-defined psychological meaning: function (articles, prepositions, etc.), cohesion (demonstratives etc.), evaluation (terms that evaluate the validity, likelihood, acceptance, etc.) and judgement (terms that express the author's attitude in relation to content.) The texts were represented by the frequencies of each category, and the binary classes Extraversion and Neuroticism were classified using SVMs, with a maximum accuracy of 58%.

In [7], an extended version of the same set of texts and inventories from [6] was considered. In this approach, learning features consisted of 88 word categories provided by the psycholinguistic dictionary LIWC (Linguistic Inquiry and Word Count) [8] and 26 attributes provided by the MRC (Medical Research Council) database [9] composed of 150,837 lexical items. An experiment was carried out to discriminate between the upper and lower ends of the Big Five dimensions, with maximum accuracy ranging from 50%–62% when using SVMs.

Studies as in [6,7] are based on word counts provided by psycholinguistic lexical resources. By contrast, studies as in [4,10] rely solely on the text itself by making use of n-grams models. In these studies, the objective was also to discriminate between individuals scores for four of the five dimensions of personality (except Openness to experience.) In both cases, Naive-Bayes and SVM classification were attempted. In [4] a set of 71 blogs was considered, with accuracy ranging from 45% (random) to 100% depending on the order of the n-gram model and class. In [10], the same experiment was repeated using a set of 1,672 blogs, with a maximum accuracy of 65%.

Finally, a note on NLP resources. Clearly, the computational problem of personality recognition from text is well developed for the English language. By making use of large scale resources such as the *myPersonality* corpus [11], the field has even experienced a number of dedicated scientific events in recent years, including the PAN-CLEF shared tasks series. In the case of the Portuguese language, by contrast, there is no obvious equivalent for the purpose of Big Five personality recognition, and we are not aware of any existing systems that may be regarded as a baseline.

3 Current Work

Given the lack of data and baseline systems for personality recognition in Brazilian Portuguese, we devised an exploratory study in which we first build a suitable corpus, and then we investigate a range of computational methods for the task. This study is described in the next sections.

3.1 Objectives

The objective of the present study is to develop supervised models of human personality recognition from Brazilian Facebook status updates, and to determine which of these models are more suitable for the task. In particular, we would like to investigate recent methods for text representation - namely, those based on word embeddings - as a possible alternative to standard personality recognition based on language-dependent psycholinguistic knowledge.

3.2 Data Acquisition

The computational models under discussion were built from a Facebook corpus labelled with Big Five information obtained from self-report IGFP-5 personality inventories [3]. To this end, a Facebook application was developed. The application requests users to fill in the 44 items of the IGFP-5 inventory as proposed in [12], and from which the five dimensions of personality are computed.

In addition to providing the personality inventories, the application simultaneously collects the user's status updates upon consent. Once the personality inventory is completed, a result page displayed details about the user's personality, and a brief explanation of each trait. The purpose of this page was however merely illustrative, that is, aiming to offer some kind of reward to the participant as a mean to possibly motivate them to further disseminate the Facebook application to their social circle.

As discussed in [7], the accuracy of this form of self-assessment is admittedly lower than third-party evaluation (i.e., performed by Psychology experts.) However, due to the costs of a large-scale professional evaluation of this kind, self-assessment remains the most common method in the field [6,7], and may be considered sufficient for the purposes of the present (exploratory) study as well.

We obtained data from 1,039 participants to create a corpus that is, to the best of our knowledge, the largest data set of this kind for the Brazilian Portuguese language. The corpus - hereby called *b5-post* - contains 2.2 million words in total, and it was subject to a number of pre-processing, spell-checking and normalisation procedures to be described elsewhere.

3.3 Computational Models

We follow a great deal of previous studies such as [4–7] in that we model personality recognition as five independent binary classification tasks, that is, one for each personality dimension of the Big Five model. This decision is motivated both by the type of application intended (i.e., we would like to recognise personality traits exclusively from text, and not from other already known traits), and also by the fact that the personality factors of the Big Five model are, by definition, highly independent [2].

In our models, individuals are classified as either positive or negative for each Big Five dimension based on the mean personality score for that class.

Table 1 presents the number and proportion of positive and negative individuals (or classification instances) in the corpus.

Table 1. Positive and negative learning instances

Class	Positive		Negative		Mean
Extraversion	412	52.9%	366	47.1%	3.15
Agreeableness	406	52.1%	372	47.9%	3.54
Conscientiousness	381	48.9%	397	51.1%	3.25
Neuroticism	389	50.0%	389	50.0%	3.17
Openness to exp.	408	52.4%	370	47.6%	3.78

As a means to provide an overview of possible strategies for personality recognition from text, we envisaged a range of models based on Random Forest classification with different text representations. These models are complemented by an alternative that makes use of long short-term memory neural networks (LSTMs). In the case of Random Forest, we use 10-fold cross validation over the entire dataset. In the case of the LSTMs models, the corpus was split into training (70%) and test (30%) subsets.

We carried out dozens of experiments involving alternative text representations and machine learning algorithms. For brevity, however, the present discussion will be limited to six of the best-performing alternatives and relevant baseline models, hereby called BoW, Psycholinguistics, word2vec-cbow-600, word2vec-skip-600, doc2vec and LSTM-600. Details are provided as follows.

- **BoW:** A bag-of-words model retaining 22,612 terms after lemmatisation.
- **Psycholinguistics:** LIWC-BR word counts [13] and psycholinguistic properties for Brazilian Portuguese [14]. LIWC categories include words that indicate emotions, social relations, cognitive processes, etc. Psycholinguistic properties include word age of acquisition, concreteness, etc.
- **word2vec-cbow-600:** A word2vec cbow model [15] of size 600, trained from a corpus of 50k tweets [16] using the vector component-wise average with the text corpus in lower case.
- **word2vec-skip-600:** A word2vec skip-gram model [15] of size 600, with the same features as the above cbow model.
- **doc2vec:** A doc2vec model [17] in which a document is defined as the set of all Facebook status updates written by each participant.
- **LSTM-600:** A Keras embedding model of size 600 built from the *b5-post* corpus, with combination provided by the LSTM hidden layers.

4 Results

The six models - and a Majority class baseline - were applied to the recognition of the five personality dimensions, resulting in 35 binary classifiers. Table 2 reports mean F1 scores for each model and class (i.e., each personality trait.)

Table 2. Mean F1 scores results

Model	EXT	AGR	CON	NEU	OPE
Majority baseline	0.33	0.34	0.33	0.32	0.34
BoW	0.60	0.53	0.57	0.53	0.55
Psycholinguistics	0.61	0.53	0.57	0.54	0.53
word2vec-cbow-600	0.61	0.56	0.58	0.55	0.59
word2vec-skip-600	0.61	0.55	0.59	0.55	0.57
doc2vec	0.60	0.58	0.57	0.55	0.55
LSTM-600	0.58	0.53	0.45	0.54	0.59

Based on these results, a number of observations are warranted. First, we notice that no single model is capable of providing the best results for all five classes. This may suggest that not all personality traits are equally accessible from text (or at least not in our domain.) Second, we notice that the Majority baseline and BoW never outperform the other models. This may suggest that the use of word embeddings is indeed a suitable approach to the task.

Looking at the classes individually, we notice that Extraversion results are slightly superior to those observed for the other classes, a result that has already been suggested in previous studies devoted to the English language [7]. As for the other classes, we notice that best results are divided between various methods, with a small advantage for the CBOW architecture.

5 Discussion

This article presented an exploratory study on the computational problem of human personality recognition from social network texts in the Brazilian Portuguese language. A corpus of texts labelled with personality information was collected, and subsequently used as training data for a range of supervised machine learning models of personality recognition.

Results suggest that different personality traits may be more or less evident from (Facebook) text, and that there is no single best-performing model for all traits. Despite our relatively small dataset, we notice that models based on word embeddings seem to outperform those based on lexical resources and, perhaps more importantly, we notice that these methods do not require language-specific resources such as psycholinguistic databases.

As future work, we consider improving the models based on word embeddings by making use of deep neural networks such as our current LSTM model. Despite the relatively weak results reported in our initial experiments, we believe that further fine-tuning of the network hyper parameters may provide more significant results in this regard.

The original *b5-post* corpus has been made publicly available for research, and has been reused on a number of related projects. Details regarding the corpus are discussed in [18]. An experiment comparing personality recognition models based on Facebook and other textual sources is presented in [19]. The corpus has also been applied to the task of author profiling (i.e., for predicting author's gender, age group and others) in [20]. Finally, a pilot experiment investigating alternative models of personality appeared in [21].

Acknowledgements. The second author received supported from grant # 2016/14223-0, São Paulo Research Foundation (FAPESP).

References

1. Allport, F.H., Allport, G.W.: Personality traits: their classification and measurement. J. Abnorm. Soc. Psychol. **16**, 6–40 (1921)
2. Goldberg, L.R.: An alternative description of personality: the Big-Five factor structure. J. Pers. Soc. Psychol. **59**, 1216–1229 (1990)
3. de Andrade, J.M.: Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil. Ph.D. thesis, Universidade de Brasília (2008)
4. Oberlander, J., Nowson, S.: Whose thumb is it anyway? Classifying author personality from weblog text. In: COLING/ACL-2006 Poster Sessions, Sydney, Australia, Association for Computational Linguistics, pp. 627–634 (2006)
5. Celli, F.: Adaptive personality recognition from text. Ph.D. thesis, University of Trento (2012)
6. Argamon, S., Dhawle, S., Koppel, M., Pennebaker, J.W.: Lexical predictors of personality type. In: The Joint Annual Meeting of the Interface and the Classification Society of North America (2005)
7. Mairesse, F., Walker, M., Mehl, M., Moore, R.: Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Intell. Res. (JAIR) **30**, 457–500 (2007)
8. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Inquiry and Word Count: LIWC. Lawrence Erlbaum, Mahwah (2001)
9. Coltheart, M.: The MRC psycholinguistic database. Q. J. Exp. Psychol. Sect. A: Hum. Exp. Psychol. **33**(4), 497–505 (1981)
10. Nowson, S., Oberlander, J.: Identifying more bloggers: towards large scale personality classification of personal weblogs. In: Proceedings of the International Conference on Weblogs and Social Media, Boulder, Colorado, USA (2007)
11. Kosinski, M., Matz, S., Gosling, S., Popov, V., Stillwell, D.: Facebook as a social science research tool: opportunities, challenges, ethical considerations and practical guidelines. Am. Psychol. **70**(6), 543–556 (2015)
12. John, O.P., Naumann, L.P., Soto, C.J.: Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues, pp. 114–158. Guilford Press, New York (2008)
13. Filho, P.P.B., Aluísio, S.M., Pardo, T.: An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: 9th Brazilian Symposium in Information and Human Language Technology - STIL, Fortaleza, Brazil, pp. 215–219 (2013)
14. dos Santos, L.B., Duran, M.S., Hartmann, N.S., Candido, A., Paetzold, G.H., Aluisio, S.M.: A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In: Ekštein, K., Matoušek, V. (eds.) TSD 2017. LNCS (LNAI), vol. 10415, pp. 281–289. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64206-2_32

15. Mikolov, T., Wen-tau, S., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL-HLT-2013, Atlanta, USA, pp. 746–751. Association for Computational Linguistics (2013)
16. Ramos Casimiro, C., Paraboni, I.: Temporal aspects of content recommendation on a microblog corpus. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.G. (eds.) PROPOR 2014. LNCS (LNAI), vol. 8775, pp. 189–194. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09761-9_20
17. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of Machine Learning Research, PMLR, Beijing, China, vol. 32, no. 2, pp. 1188–1196 (2014)
18. Ramos, R.M.S., Neto, G.B.S., Silva, B.B.C., Monteiro, D.S., Paraboni, I., Dias, R.F.S.: Building a corpus for personality-dependent natural language understanding and generation. In: 11th International Conference on Language Resources and Evaluation (LREC-2018), ELRA, Miyazaki, Japan, pp. 1138–1145 (2018)
19. dos Santos, V.G., Paraboni, I., Silva, B.B.C.: Big Five personality recognition from multiple text genres. In: Ekštein, K., Matoušek, V. (eds.) TSD 2017. LNCS (LNAI), vol. 10415, pp. 29–37. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64206-2_4
20. Hsieh, F.C., Dias, R.F.S., Paraboni, I.: Author profiling from facebook corpora. In: 11th International Conference on Language Resources and Evaluation (LREC-2018), ELRA, Miyazaki, Japan, pp. 2566–2570 (2018)
21. Silva, B.B.C., Paraboni, I.: Learning personality traits from Facebook text. IEEE Lat. Am. Trans. **16**(4), 1256–1262 (2018)

Portuguese Native Language Identification

Shervin Malmasi[1](✉), Iria del Río[2], and Marcos Zampieri[3]

[1] Harvard Medical School, Boston, USA
shervin.malmasi@mq.edu.au
[2] University of Lisbon, Lisbon, Portugal
[3] University of Wolverhampton, Wolverhampton, UK

Abstract. This study presents the first Native Language Identification (NLI) study for L2 Portuguese. We used a sub-set of the NLI-PT dataset, containing texts written by speakers of five different native languages: Chinese, English, German, Italian, and Spanish. We explore the linguistic annotations available in NLI-PT to extract a range of (morpho-)syntactic features and apply NLI classification methods to predict the native language of the authors. The best results were obtained using an ensemble combination of the features, achieving 54.1% accuracy.

Keywords: Native Language Identification · Learner corpus
Portuguese

1 Introduction

Native Language Identification (NLI) is the task of determining the native language (L1) of an author based on their second language (L2) linguistic productions [1]. NLI works by identifying language use patterns that are common to groups of speakers of the same native language. This process is underpinned by the presupposition that an author's L1, disposes them towards certain language production patterns in their L2, as influenced by their mother tongue. A major motivation for NLI is studying second language acquisition. NLI models can enable analysis of inter-L1 linguistic differences, allowing us to study the language learning process and develop L1-specific pedagogical methods and materials.

NLI research is conducted using learner corpora: collections of learner writing in an acquired language, annotated with metadata such as the author's L1 or proficiency. These datasets are the foundation of NLI experiments and their quality and availability has been a key issue since the earliest work in this area.

A notable research trend in recent years, and the focus of this paper, has been the extension of NLI to languages other than English [2]. Recent NLI studies on languages other than English include Chinese [3], Norwegian [4], and Arabic [5].

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 115–124, 2018.
https://doi.org/10.1007/978-3-319-99722-3_12

Since the learner corpus is a core component of NLI work, extending the task to a new language depends on the availability, or collection, of suitable learner corpora.

Early research focused on L2 English, as it is one of the most widely studied languages and data has been more readily available. However, continuing globalization has resulted in increased acquisition of languages other than English [6]. Additionally, researchers have sought to investigate whether the NLI methods that work for English would work for other languages, and whether similar performance trends hold across corpora. These motivations have led to an extension of NLI research to new non-English languages, of which our research directly contributes.

To the best of our knowledge, this study presents the first detailed NLI experiments on L2 Portuguese. A number of studies have been published on educational NLP applications and learner language resources for Portuguese, but so far none of them have included NLI. Examples of educational NLP studies that included Portuguese range from grammatical error correction [7] and automated essay scoring [8], to language resources such as the Portuguese Academic Wordlist (P-AWL) [9], and the learner corpus COPLE2 [10] which is part of the dataset used in our experiments.

The remainder of the paper is organized as follows: Sect. 2 discusses related work in NLI, Sect. 3 describes the methodology and dataset used in our experiments, and Sect. 4 presents the experimental results. Finally, Sect. 5 presents a brief discussion and concludes this paper with avenues for future research.

2 Related Work

NLI is a fairly recent, but rapidly growing area of research. While some research was conducted in the early 2000s, the most significant work has only appeared in recent years [11–15].

NLI is typically modeled as a supervised multi-class classification task. In this experimental design the individual writings of learners are used as training and testing data while the author's L1 information serves as class labels. NLI has received much attention in the research community over the past decade, with efforts focusing on improving classification [14], studying language transfer effects [16], and applying the linguistic features to other NLP tasks [17]. It has also been empirically demonstrated that NLI is a challenging task even for human experts, with machine learning approaches significantly outperforming humans on the same test data [18].

The very first shared task focusing on NLI was held in 2013, bringing further focus, interest and attention to the field.[1] The competition attracted entries from 29 teams. The winning entry for the shared task was that of [19], with an accuracy of 83.6%. The features used in this system are n-grams of words, parts-of-speech, as well as lemmas. In addition to normalizing each text to unit

[1] https://sites.google.com/site/nlisharedtask2013/home.

length, the authors applied a log-entropy weighting schema to the normalized values, which clearly improved the accuracy of the model. An L2-regularized SVM classifier was used to create a single-model system.

Growing interest led to another edition of the shared task in 2017, where the task was expanded to include speech data.[2] The results of the task showed that various types of multiple classifier systems, such as ensembles and meta-classifiers, achieved the best performance across the different tracks. While a number of participants attempted to utilize newer deep learning-based models and features (e.g. word embeddings), these approaches did not outperform traditional classification systems. Finally, it was also shown that as participants had used more sophisticated systems, results were on average substantially higher than in the previous edition of the task. A detailed report on the findings of the task can be found in [20].

With respect to classification features, NLI research has grown to use a wide range of syntactic, and more recently, lexical features to distinguish the L1. A more detailed review of NLI methods is omitted here for brevity, but a comprehensive exposition of the methods can be found in [21,22]. Some of the most successful syntactic and lexical features used in previous work includes Adaptor Grammars (AG) [23], character n-grams [24], Function word unigrams and bigrams [25], Word and Lemma n-grams, CFG Production Rules [12], Penn Treebank (PTB) part-of-speech n-grams, RASP part-of-speech n-grams [25], Stanford Dependencies with POS transformations [14], and Tree Substitution Grammar (TSG) fragments [13].

NLI is now also moving towards using models based on these features to generate Second Language Acquisition (SLA) hypotheses. In [26] the authors approach this by using both L1 and L2 data to identify features exhibiting non-uniform usage in both datasets, using them to create lists of candidate transfer features. The authors of [16] propose a different methodology, using linear SVM weights to extract lists of overused and underused linguistic features per L1 group.

Most English NLI work has been done using two corpora. The *International Corpus of Learner English* [27] was widely used until recently, despite its shortcomings[3] being widely noted [28]. More recently, TOEFL11, the first corpus designed for NLI was released [29]. While it is the largest NLI dataset available, it only contains argumentative essays, limiting analyses to this genre.

An important trend has been the extension of NLI research to languages other than English [5,30]. Recently, [3] introduced the Jinan Chinese Learner Corpus [31] for NLI and their results indicate that feature performance may be similar across corpora and even L1-L2 pairs. Similarly, [4] also proposed using the ASK corpus [32] to conduct NLI research using L2 Norwegian data.

In this study we also follow this direction, presenting new experiments on L2 Portuguese. Other aspects of our work, such as the classification methodology and features, are largely based on the approaches discussed above.

[2] https://sites.google.com/site/nlisharedtask/home.

[3] The issues exist as the corpus was not designed specifically for NLI.

3 Data and Method

3.1 Data

We used a sub-set of the NLI-PT dataset [33] containing texts for five L1 groups: Chinese, English, German, Italian, and Spanish. We chose these five languages because they are the ones with the greatest number of texts in NLI-PT. The sub-set is balanced in terms of proficiency level by L1. The composition of our data is shown in Table 1.

Table 1. Distribution of the five L1s in the NLI-PT datasets in terms of texts, tokens, types, and type/token ratio (TTR).

L1	Texts	Tokens	Types	TTR
Chinese	215	50,750	6,238	0.12
English	215	49,169	6,480	0.13
German	215	52,131	6,690	0.13
Italian	215	51,171	6,814	0.13
Spanish	215	47,935	6,375	0.13
Total	1,075	251,156	32,597	0.13

Texts in NLI-PT are automatically annotated using available NLP tools at two levels: Part of Speech (POS) and syntax. There are two types of POS: a simple POS with only the type of word, and a fine-grained POS with type of word plus morphological features. Concerning syntactic information, texts are annotated with constituency and dependency representations. These annotations can be used as classification features.

3.2 Classification Models and Evaluation

In our experiments we utilize a standard multi-class classification approach. A linear Support Vector Machine [34] is used for classification and feature vectors are created using relative frequency values, in line with previous NLI research [21]. A single model is trained on each feature type to evaluate feature performance. We then combine all our features using a mean probability ensemble.[4]

Similar to the majority of previous NLI studies, we report our results as classification accuracy under k-fold cross-validation, with $k = 10$. In recent years this has become the accepted standard for reporting NLI results. For generating our folds we use randomized stratified cross-validation which aims to ensure that the proportion of classes within each partition is equal [35]. While accuracy is a suitable metric as the data classes are balanced in our corpus, we also report per-class precision, recall, and F1-scores. We also compare these results against a random baseline.

[4] More details about this approach can be found in [21].

3.3 Features

Previous work on NLI using datasets which are not controlled for L1 and topic [3,5] avoids using lexical features. Using only non-lexicalized features allows researchers to model syntactic differences between classes and avoid any topical cues. For the same reasons, we do not use lexical features (*e.g.* word *n*-grams) as NLI-PT is not topic balanced. While a detailed exposition of this issue is beyond the scope of this paper, a comprehensive discussion can be found in [1, p. 23].

We extract the following topic-independent feature types: function words, context-free grammar production rules, and POS tags, as outlined below.

Function words are topic-independent grammatical words such as prepositions, which indicate the relations between content words. They are known to be useful for NLI. Frequencies of 220 Portuguese function words[5] are extracted as features. We also make this list available as a resource.[6]

Context-free grammar production rules are the rules used to generate constituent parts of sentences, such as noun phrases.[7] These rules can be obtained by first generating constituent parses for sentences. The production rules, excluding lexicalizations, are then extracted and each rule is used as a single classification feature. These context-free phrase structure rules capture the overall structure of grammatical constructions and global syntactic patterns. They can also encode highly idiosyncratic constructions that are particular to an L1 group. They have previously been found to be useful for NLI [12]. Our dataset already includes parsed versions of the texts which we used to extract these features.

Part-of-Speech (POS) tags are linguistic categories (or word classes) assigned to words that signify their syntactic role. Basic categories include verbs, nouns and adjectives, but these can be expanded to include additional morpho-syntactic information. The assignment of such categories to words in a text adds a level of linguistic abstraction. Our dataset already includes POS tags and *n*-grams of size 1–3 are extracted as features. They capture preferences for word classes and their localized ordering patterns. Previous work, and our own experiments, demonstrates that sequences of order 4 or greater achieve lower accuracy, possibly due to data sparsity, so we did not include them.

4 Results

In this section we first report results by individual feature types in terms of accuracy. Subsequently we report the results obtained using all features in an ensemble combination. Finally, we look at the performance obtained by the best system for each L1 class.

[5] Like previous work, this also includes stop words.

[6] http://web.science.mq.edu.au/~smalmasi/data/pt-fw.txt.

[7] They are also known as Phrase Structure Rules or Production Rules.

We first report the results obtained using systems trained on different feature types. Results are presented in terms of accuracy in Table 2. These results are compared against a uniform random baseline of 20%.

Table 2. Classification results under 10 fold cross-validation (accuracy is reported).

Feature type	Accuracy (%)
Random baseline	20.0
Function words	38.5
POS 1-grams	46.3
POS 2-grams	52.8
POS 3-grams	44.9
CFG production rules	43.3
Ensemble combination	54.1

We observed that all features types individually deliver results well above the baseline. POS bigrams are the features that individually obtain the best performance, achieving 52.8% accuracy. This demonstrates the importance of syntactic differences between the L1 groups. The ensemble combination, using all feature types, obtains performance higher than POS bigrams achieving 54.1% accuracy. These trends are very similar to previous research using similar features, but the boost provided by the ensemble is more modest. This is likely because the syntactic features used here are not as diverse as including other syntactic and lexical features, as shown in [36].

We also experimented with tuning the regularization hyperparameter of the SVM mode. This parameter (C) is considered to be the inverse of regularization strength; increasing it decreases regularization and vice versa. The results from the POS bigram model are shown in Fig. 1. We performed a grid search of the parameter space in the range of 10^{-6} to 10^{1}. We observe that model generalization (i.e. cross-validation score) is quite poor with strong regularization and improves as the parameter is relaxed. Generalization plateaus as approximately $C = 1$ and we therefore select this parameter value. Similar patterns hold for all feature types, but results are not included for reasons of space.

In Table 3 we present the results obtained for each L1 in terms of precision, recall, and F1 score as well as the average results on the five classes. Across all classes, we obtain a micro-averaged F1 score of 0.531 and a macro-averaged F1 score of 0.530.

Looking at individual classes, the results obtained for Chinese are higher than those of other L1s. One hypothesis is that as English, German, Italian, and Spanish are Indo-European languages, properties of Chinese, which belongs to the Sino-Tibetan family, are helping the system to discriminate Chinese texts with much higher performance than the other three L1s. To visualize these results

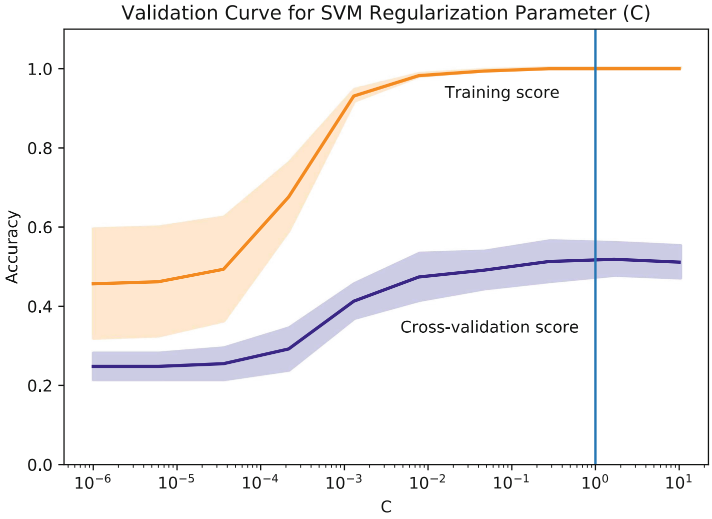


Fig. 1. Results for tuning the regularization hyperparameter (C) of the POS bigram SVM model. The top represents performance on the training set, while the bottom line is the cross-validation accuracy. The vertical line represents the value of C = 1.

and any notable error patterns, in Fig. 2 we present a heatmap confusion matrix of the classification errors.

Table 3. Ensemble system per-class results: precision, recall and the F1-score are reported.

Class	Precision	Recall	F1-score
CHI	0.571	0.796	0.665
ENG	0.507	0.326	0.397
GER	0.542	0.547	0.545
ITA	0.549	0.577	0.562
SPA	0.510	0.460	0.484
Average	0.536	0.541	0.531

Finally, another important finding here is that our results suggest the existence of syntactic differences between the L1 groups. Earlier in Sect. 3.3 we justified the use of non-lexical features to avoid topic bias, and the presence of such bias is also evidenced by the difference between our results and the lexical baseline provided with the dataset description [33]. Such lexical models built using topic-imbalanced datasets may not capture actual L1 differences between the classes. Accordingly, the results are often artificially inflated and may actually represent thematic classification.

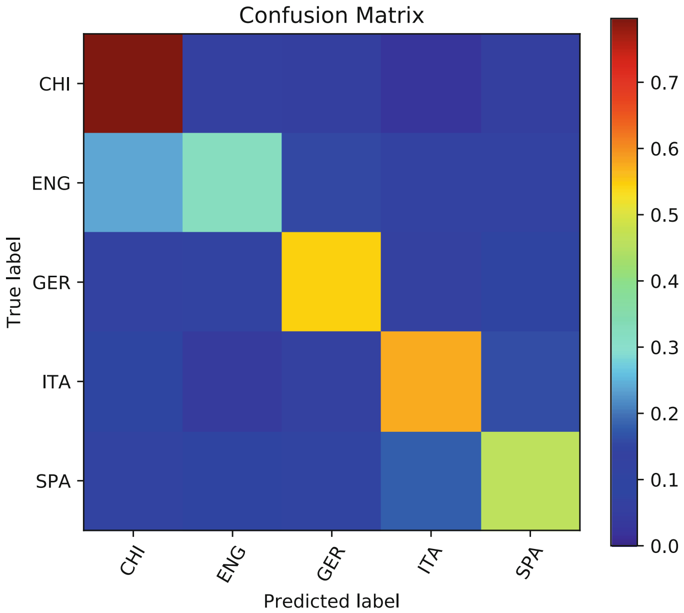


Fig. 2. Confusion matrix for our ensemble system.

5 Conclusion and Future Work

This paper presented the first NLI experiments on Portuguese. These results add to the growing body of evidence that demonstrates the applicability of NLI methods to various languages. The availability of the presented dataset also allows future research and hypotheses to be tested on another NLI corpus, which are valuable resources.

The presented results are comparable to those of other NLI studies [2], but not as high as those on the largest and most balanced corpora [20]. This is likely a limitation of our data, which we will address below.

This study opens several avenues for future research. One of them is investigating the influence of L1 in Portuguese second language acquisition. Such approaches, similar to those applied to English learner data [16], can have direct pedagogical implications. For example, the identification of the most discriminative language transfer features can lead to recommendations for language teaching and assessment methods. Such NLI models can provide the means to perform qualitative studies of the distinctive characteristics of each L1 group, allowing these differences to be described. Following this, further analysis may attempt to trace the linguistic phenomena to causal features of the L1 in order to explain their manifestation.

There are several directions for future work. The evaluation of more features, such as dependency parses, could be helpful. The application of more advanced ensemble methods, such as meta-classification [21], have also proven to be useful for NLI, as well as other tasks [37,38]. However, we believe that the most valuable (and challenging) next step is the refinement and extension of the learner corpus. Having more data is extremely important in improving NIL accuracy. Additionally, well-balanced data is a key component of NLI experiments and having a dataset that is more carefully balanced for topic and proficiency will be of utmost importance for future research in this area.

Acknowledgements. We would like to thank the anonymous reviewers for the suggestions and constructive feedback provided.

References

1. Malmasi, S.: Native language identification: explorations and applications. Ph.D. thesis (2016)
2. Malmasi, S., Dras, M.: Multilingual native language identification. In: Natural Language Engineering (2015)
3. Malmasi, S., Dras, M.: Chinese native language identification. In: Proceedings of EACL. Association for Computational Linguistics, Gothenburg (2014)
4. Malmasi, S., Dras, M., Temnikova, I.: Norwegian native language identification. In: Proceedings of RANLP, Hissar, Bulgaria, pp. 404–412, September 2015
5. Malmasi, S., Dras, M.: Arabic native language identification. In: Proceedings of the Arabic Natural Language Processing Workshop (2014)
6. Block, D., Cameron, D.: Globalization and Language Teaching. Routledge, Abingdon (2002)
7. Martins, R.T., Hasegawa, R., Nunes, M.G.V., Montilha, G., De Oliveira, O.N.: Linguistic issues in the development of ReGra: a grammar checker for Brazilian Portuguese. Nat. Lang. Eng. **4**(4), 287–307 (1998)
8. Elliot, S.: IntelliMetric: From here to validity. In: A Cross-Disciplinary Perspective, Automated Essay Scoring, pp. 71–86 (2003)
9. Baptista, J., Costa, N., Guerra, J., Zampieri, M., Cabral, M., Mamede, N.: P-AWL: academic word list for Portuguese. In: Pardo, T.A.S., Branco, A., Klautau, A., Vieira, R., de Lima, V.L.S. (eds.) PROPOR 2010. LNCS (LNAI), vol. 6001, pp. 120–123. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12320-7_15
10. Mendes, A., Antunes, S., Janssen, M., Gonçalves, A.: The COPLE2 corpus: a learner corpus for Portuguese. In: Proceedings of LREC (2016)
11. Wong, S.M.J., Dras, M.: Contrastive analysis and native language identification. In: Proceedings of ALTA, Sydney, Australia, pp. 53–61, December 2009
12. Wong, S.M.J., Dras, M.: Exploiting parse structures for native language identification. In: Proceedings of EMNLP (2011)
13. Swanson, B., Charniak, E.: Native language detection with tree substitution grammars. In: Proceedings of ACL, Jeju Island, Korea, pp. 193–197, July 2012
14. Tetreault, J., Blanchard, D., Cahill, A., Chodorow, M.: Native tongues, lost and found: resources and empirical evaluations in native language identification. In: Proceedings of COLING, Mumbai, India, pp. 2585–2602 (2012)
15. Gebre, B.G., Zampieri, M., Wittenburg, P., Heskes, T.: Improving native language identification with TF-IDF weighting. In: Proceedings of BEA (2013)
16. Malmasi, S., Dras, M.: Language transfer hypotheses with linear SVM weights. In: Proceedings of EMNLP, pp. 1385–1390 (2014)

17. Malmasi, S., Dras, M., Johnson, M., Du, L., Wolska, M.: Unsupervised text segmentation based on native language characteristics. In: Proceedings of ACL (2017)
18. Malmasi, S., Tetreault, J., Dras, M.: Oracle and human baselines for native language identification. In: Proceedings of BEA (2015)
19. Jarvis, S., Bestgen, Y., Pepper, S.: Maximizing classification accuracy in native language identification. In: Proceedings of BEA (2013)
20. Malmasi, S., et al.: A report on the 2017 native language identification shared task. In: Proceedings of BEA (2017)
21. Malmasi, S., Dras, M.: Native Language Identification using Stacked Generalization. arXiv preprint arXiv:1703.06541 (2017)
22. Malmasi, S., Dras, M.: Native language identification with classifier stacking and ensembles. Computational Linguistics (2018)
23. Wong, S.M.J., Dras, M., Johnson, M.: Exploring adaptor grammars for native language identification. In: Proceedings of EMNLP (2012)
24. Tsur, O., Rappoport, A.: Using classifier features for studying the effect of native language on the choice of written second language words. In: Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition (2007)
25. Malmasi, S., Wong, S.M.J., Dras, M.: NLI shared task 2013: MQ submission. In: Proceedings of BEA (2013)
26. Swanson, B., Charniak, E.: Data driven language transfer hypotheses. EACL **2014**, 169 (2014)
27. Granger, S., Dagneaux, E., Meunier, F., Paquot, M.: International Corpus of Learner English (Version 2). Presses Universitaires de Louvain, Louvian-la-Neuve (2009)
28. Brooke, J., Hirst, G.: Measuring interlanguage: native language identification with L1-influence metrics. In: Proceedings of LREC (2012)
29. Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., Chodorow, M.: TOEFL11: a corpus of non-native English. Educational Testing Service, Technical report (2013)
30. Malmasi, S., Dras, M.: Finnish native language identification. In: Proceedings of ALTA, Melbourne, Australia, pp. 139–144 (2014)
31. Wang, M., Malmasi, S., Huang, M.: The Jinan Chinese learner corpus. In: Proceedings of BEA (2015)
32. Tenfjord, K., Meurer, P., Hofland, K.: The ASK corpus: a language learner corpus of Norwegian as a second language. In: Proceedings of LREC (2006)
33. del Río, I., Zampieri, M., Malmasi, S.: A Portuguese native language identification dataset. In: Proceedings of BEA (2018)
34. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
35. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI **14**, 1137–1145 (1995)
36. Malmasi, S., Cahill, A.: Measuring feature diversity in native language identification. In: Proceedings of BEA (2015)
37. Malmasi, S., Dras, M., Zampieri, M.: LTG at SemEval-2016 Task 11: complex word identification with classifier ensembles. In: Proceedings of SemEval (2016)
38. Malmasi, S., Zampieri, M., Dras, M.: Predicting post severity in mental health forums. In: Proceedings of CLPsych (2016)

Text-Image Alignment in Portuguese News Using LinkPICS

Wellington Cristiano Veltroni and Helena de Medeiros Caseli

Universidade Federal de São Carlos (UFSCar), Rod. Washington Luís Km 235,
CP 676, São Carlos, SP CEP: 13565-905, Brazil
wellingtonveltroni@gmail.com, helenacaseli@ufscar.br

Abstract. Text-image alignment is the task of aligning elements in a text with elements in the image accompanying it. Text-image alignment can be applied, for example, in news articles to improve clarity by explicitly defining the correspondence between regions in the article's image and words or named entities in the article's text. It can also be an useful step in many multimodal applications such as image captioning or image description/comprehension. In this paper we present the LinkPICS: an automatic aligner which combines Natural Language Processing (NLP) and Computer Vision (CV) techniques to explicitly define the correspondence between regions of an image (bounding boxes) and elements (words or named entities) in a text. LinkPICS performs the alignment of people and objects (or animals, vehicles, etc.) as two distinct processes. In the experiments present in this paper, LinkPICS obtained a precision of 97% in the alignment of people and 73% in the alignment of objects in articles in Portuguese from a Brazilian news site.

Keywords: Text-image alignment · Aligner · LinkPICS
Brazilian Portuguese · Alignment of people · Alignment of objects

1 Introduction

Text-image alignment is the task of finding explicit correspondences between elements in a text (words, expressions, etc.) and visual elements found in the image accompanying it (delimited by bounding boxes).

Traditionally, the alignment has been performed with parallel texts[1] to find the correspondences between source and target sentences or words (e.g. GIZA++ [17]). In the visual domain, the alignment arose as a demand for the preservation and comprehension of historical manuscripts, by means of the digitization and transcription of the manuscripts followed by the alignment between the scanned image and the transcribed text [9,12,25,30,31].

Another application quite related to the text-image alignment is the image annotation. In automatic image annotation, the images are annotated with

[1] Parallel texts are texts written in one language accompanied by their translations to another language.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 125–135, 2018.
https://doi.org/10.1007/978-3-319-99722-3_13

keywords. However, different from what occurs in text-image alignment, in image annotation there is no clear link between a region in the image and the word used to annotate the whole image. Works on image annotation [3,6,7,16,19,20,24,26–28] use Natural Language Processing (NLP), Computer Vision (CV) and Machine Learning (ML) techniques to find the best set of words to define (annotate) an image.

The main goal of this paper is to present the LinkPICS: an automatic text-image aligner which combines NLP and CV techniques to explicitly define the correspondence between regions in an image (bounding boxes) and elements (words or named entities) in a text. LinkPICS performs the text-image alignment as two distinct processes: (1) alignment of people and (2) alignment of objects (animals, vehicles, etc.). LinkPICS obtained 97% of precision in the alignment of people and 73% in the alignment of objects as detailed in Sect. 4.3.

This paper is organized as follows: Sect. 2 brings some related work on text-image alignment and image captioning; Sect. 3 describes the LinkPICS's architecture and Sect. 4 presents the experiments carried out to evaluate the proposed aligner. Finally, Sect. 5 concludes this paper with some final remarks and proposals for future work.

2 Related Work

Several related works perform the text-image alignment in corpora composed of news related articles [16,19,24,27]. Thus, we also chose to test our proposed aligner in this domain. However, different from the related works, in our case the alignment is explicitly defined.

In [24], objects occurring in the image are annotated based on texts extracted from the New York Times, in English, and pictures of sports from [13]. The text processing phase includes the identification of nouns and adjectives followed by the filtering of those that could not be mapped to objects in the image. This filtering step is carried out based on the WordNet [8]. The bounding boxes are detected applying a segmentation technique followed by the feature extraction step. Filtered words and image features are used to generate a mapping in a common latent space and the alignment is performed based on the conditional probability. The main difference between this related work and ours is that our alignment is explicitly (not implicitly) defined between the text and the regions of the image. But, similar to it, our approach also filters out the not physical words (the words that can not describe an object in the image) based on the OpenWordNet-PT [18]. The object alignment approach proposed in [24] achieved 35% of precision and 47% of F-Measure.

In [19], the names in the image caption are aligned with the bounding boxes containing faces in the associated image. To do so, a named entity recognizer is applied together with a face detector followed by an extractor of facial features. The alignment is performed based on the co-occurrence of names and faces by means of Expectation-Maximization [5]. The evaluation was carried out with news from the Yahoo! News site [2] and the LFW (Labeled Faces in the Wild) dataset [11]. The authors reported a precision of 71%. Similar to [19], our approach also uses the LFW dataset and also performs the facial features extraction step during the image processing phase.

Finally, in [15] an implicit alignment is performed during the process of image description generation and image comprehension. The alignment is performed by convolutional neural networks (CNN) and recurrent neural networks (RNN). The experiments were performed using a new dataset with images and referencing expressions based on the MS-COCO [14] images. Since their intended application was not the text-image alignment, we do not report their results here because it would not be possible to compare them to ours neither to the results of the other related works presented in this section.

3 The LinkPICS

This section describes the architecture of the LinkPICS text-image aligner, which can be seen in Fig. 1.

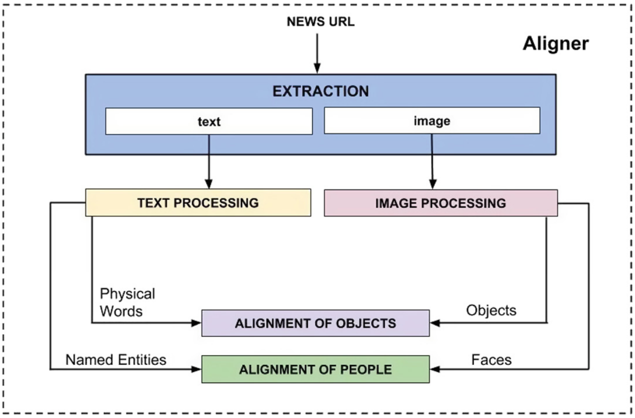


Fig. 1. LinkPICS's architecture. After collecting the news, text and image are processed separately and the alignment is performed in two separate processes: one for aligning people and other for aligning objects.

The first step in the LinkPICS's workflow is to extract the most relevant information from a news article. To do so, we built a web crawler that extracts textual elements – (1) the title of the article (its headline), (2) the text of the article and (3) the image caption – and the image associated with the text.

In the **text processing phase**, first, the textual elements extracted from the news article are part-of-speech tagged and lemmatized. Then, similar to [24], the filter of physical words selects only the nouns and filters out all the nouns that are not marked as physical words in the WordNet hierarchy. The output is a list of physical words which is the textual input of the alignment of objects.

Also in the text processing phase, a named entity (NE) recognizer is applied then, to the textual elements extracted from the news article. The recognized NEs are grouped together whenever they refer to the same person (e.g., Donald Trump and Trump). The grouping is performed based on a simple heuristic that verifies if a NE is contained in another one and, if so, they are grouped together (e.g., as <Donald Trump, Trump>). The output is a list of named entities which is the textual input of the alignment of people.

In the **image processing phase**, an object detector identifies the bounding boxes in the image extracted from the news article. The detected bounding boxes are separated into two classes: bounding boxes containing people and bounding boxes containing objects or animals, for example. The bounding boxes containing people are processed by a face detector and, for each detected face, its features are extracted and represented in a vector. For the bounding boxes containing objects/animals, a set of possible labels is generated. So, the image processing phase outputs two sets of bounding boxes: one containing the bounding boxes together with their facial vectors which is the input for the alignment of people; and other containing the bounding boxes and the proposed labels, which is the input for the alignment of objects.

The **alignment of people** in LinkPICS is performed as follows. Each named entity output by the text processing phase is used as a search key in a image database containing labeled faces. The images retrieved by this search are converted into facial vectors following the same process described in the image processing phase. The facial vectors generated for a named entity are compared with the facial vectors generated for each bounding box containing a person using a measure of similarity. The best matches give the alignments.

The **alignment of objects** in LinkPICS is performed as follows. The labels proposed for objects as the output of the image processing phase are compared with the physical words output by the text processing phase by means of a similarity measure. The five candidate words with the highest similarity scores form a TOP-5 list of candidate words for alignment. Finally, the best ranked word (TOP-1) is aligned with the region of the object in the image.

4 Experiments and Results

This section describes the experiments carried out to evaluate the alignment of people and objects performed by LinkPICS. Firstly, in Subsect. 4.1, we describe all the resources and tools applied in these experiments. Then, we present the baseline aligner developed to allow the comparison with LinkPICS (in Subsect. 4.2). Finally, in Sect. 4.3, we show the experiments performed and their results.

4.1 Resources and Tools

The resources, and similarity measures tools applied in our experiments are described bellow:

Resources

- **G1 Corpus:** The experiments described in this paper were carried out with a corpus composed of 390 news articles written in Portuguese extracted from the G1 news site[2].
- **LFW dataset:** The LFW (Labeled Face in the Wild) [11] dataset contains 13,000 labeled images of famous people. This resource is used during the alignment of people: each named entity in the text is looked up in this dataset and, if found, the corresponding image retrieved is converted into a facial vector.
- **Google images:** As we noticed that some of the named entities in our corpus do not occur in the LFW dataset, we implemented a complementary approach based on a survey done on the Google Images website[3]. Thus, if a named entity is not present in the LFW dataset, Google Images is used as a plan-B and facial vectors are generated for the first 20 images retrieved by Google.

Tools for Image Processing

- **Object detector and label proposal:** We used the YOLO [21] CNN to detect bounding boxes containing people and objects in our experiments. Yolo can detect 80 classes of objects (e.g., vehicles, animals, people). However, since we noticed that the YOLO's labels for objects did not have a good intersection with the words in our corpus, we combined YOLO with other three CNNs: DenseNet [10], DarkNet and Extraction[4]. These CNN's were trained in ImageNet dataset[5] and can classify 1000 different classes of objects. Due to this combination, LinkPICS can use more than 1000 different labels to predict the alignment of objects. Each extra CNN is applied to the bounding box detected by YOLO containing an object and 5 new labels are provided for that object. Thus, the final set of proposed labels is composed of at most 16 labels: 5 proposed by each of the three extra CNN and the one output by YOLO.
- **Face detector and generation of facial vector:** After the detection of a bounding box containing a person performed by YOLO, the face detection and generation of the facial vector are performed using FaceNet [23]. For each face, facial features are extracted and these features are mapped into a 128-position vector.

[2] http://g1.globo.com/.

[3] https://www.google.com/.

[4] http://pjreddie.com/darknet/imagenet/.

[5] http://www.image-net.org/.

Tools for Text Processing

- **Part-of-speech tagger and lemmatizer:** We used the TreeTagger[6] [22] in Portuguese for part-of-speech tagging and lemmatization.
- **Filter of physical words:** We used the lexical database OpenWordNet-PT [18] to filter physical words.
- **Named entity recognizer:** For the recognition of named entities we used Polyglot [1] in Portuguese.
- **Word embeddings:** We applied the MUSE [4] multilingual word embeddings to translate the labels proposed for objects from English to Portuguese since YOLO was trained with an English labeled dataset and retraining it for a Portuguese one was inviable.

Similarity Measures

- **Alignment of people:** The similarity measure applied to compare the facial vectors generated for a named entity and the facial vectors generated for each bounding box containing a person was the Euclidean distance. The similarity threshold was defined empirically as 0.9. Thus, if the distance between the two vectors was less than 0.9, the named entity was aligned with the bounding box containing the person.
- **Alignment of objects:** The similarity measure applied to compare the labels of a bounding box for an object and the physical words in the text was the WUP similarity [29]. WUP is based on WordNet structure and since each word in WordNet can be associated with several *synsets*, it is not possible to determine unequivocally which *synset* corresponds to the word in the text. Therefore, we calculated the similarity between all the possible *synsets* for the physical word and the set of proposed labels for the bounding box. The similarity threshold was defined empirically as 0.8. In the case of WUP similarity, the higher the value, the more similar the words are. Thus, only the pairs of words with similarity score greater than 0.8 were aligned. The highest scored word was the one chosen to be aligned with the bounding box containing an object.

4.2 Baseline

To serve as a basis for comparison, we created a baseline aligner that follows the same architecture of LinkPICS and uses the same tools and resources of it. However, the alignment approach established for the baseline is quite simple and relies on the idea of aligning the most important regions of the image (bounding boxes) with the most relevant physical words or named entities in the text. For the baseline, the importance of a bounding box or a physical word or named entity is defined based on the following criteria:

[6] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

- **Bounding box containing a person:** The importance of a bounding box containing a person is directly related to its position in the image: the closer to the lower center, the more important is the bounding box (the person) in the image.
- **Bounding box containing an object:** The importance of a bounding box containing an object, in turn, is related to the size of the object detected in the image: the larger the object, the greater is its importance.
- **Physical words and named entities:** Physical words and named entities are classified by relevance according to the following criteria[7]:
    1. Presence of the physical word or named entity in the image caption and in the headline;
    2. Presence of the physical word or named entity in the image caption;
    3. Presence of the physical word or named entity in the headline;
    4. Frequency of the physical word or named entity.

After generating the lists of bounding boxes and physical words or named entities sorted by their importance, following an approach similar to [26], the baseline aligner performs the one-to-one mapping following the order of the items in these lists. More specifically, the baseline aligner aligns the best ranked bounding box containing a person with the best ranked named entity. Similarly, the baseline aligner aligns the best ranked bounding box containing an object with the best ranked physical word. This process is repeated until it is not possible to perform any new alignment.

So, the only difference between the baseline and the LinkPICS relies on the alignment criteria which, in LinkPICS is based on the similarity measures and in the baseline aligner is based on the criteria previously described.

4.3 Evaluation

Table 1 brings the values for precision[8] obtained by the baseline aligner and LinkPICS in the alignment of people and objects. It is worth mentioning that, for this work, greater precision is more important than greater recall, since the ultimate goal of LinkPICS is to enrich the news by establishing a correct correspondence between the named entities (or physical words) in the text and the people (or objects) in the image. High precision means correct alignment, preventing the news reader from learning something wrong about a person (or an object) mentioned in the text.

As we can see from the values on this table, LinkPICS outperformed the baseline aligner in both the alignment of people and the alignment of objects, proving that the proposed approach is a promising one.

We also evaluated the precision regarding the number of bounding boxes (BB) detected in the image. This evaluation was performed because the baseline

[7] If there are more than one physical word or named entity that meet some of these criteria, the tiebreaker is performed choosing the one with the highest frequency.

[8] Precision is calculated as the number of corrected aligned instances divided by the total amount of instances.

Table 1. Precision values for the alignment of people and objects performed by the baseline and the LinkPICS in the G1 corpus also considering the number of bounding boxes (BB) detected in the image.

	Aligner	All	1 BB	2 or more BBs
People	Baseline	37.46%	76.79%	17.35%
	LinkPICS	95.52%	97.04%	90.74%
Objects	Baseline	35.65%	42.66%	30.43%
	LinkPICS	73.39%	73.02%	75.86%

aligner has a higher chance to correctly align news with only 1 BB, due to its characteristic of aligning the region of interest to the most important physical word or named entity in the text. The complexity of the alignment increases with the number of BBs. From these results, it is possible to notice that unlike the baseline aligner, LinkPICS can align news articles with several people or objects with a high performance. For people, it achieved 90.74% precision while the baseline aligner was not able to disambiguate and correctly align the news with several people (only 17.35% of precision). To illustrate this fact, Fig. 2 show some examples of articles in which the bounding boxes were correctly aligned by LinkPICS.

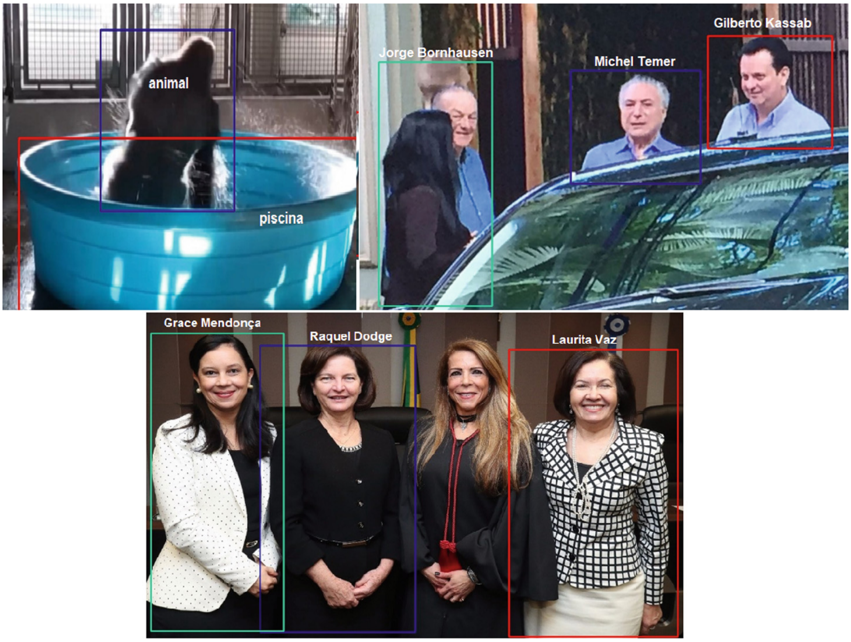


Fig. 2. LinkPICS's alignment. These examples show that our proposed approach has the ability to correctly align more than one object or people in a picture.

Another conclusion regarding the alignment is related to our proposed approach of using CNNs together with YOLO to align objects. This approach has proved to be very effective. To illustrate this, Fig. 3 presents four examples of bounding boxes incorrectly labeled by YOLO but which were successfully aligned thanks to our approach of usings extra CNNs.

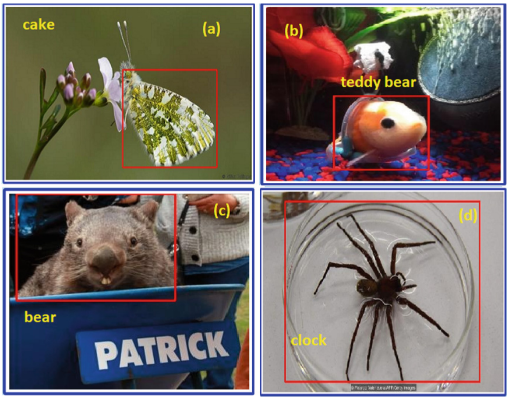


Fig. 3. Examples of bounding boxes containing animals that were incorrectly labeled by YOLO but correctly labeled by the extra CNNs, an approach proposed in this paper

5 Conclusions and Future Work

This paper addressed the text-image alignment, an essential step in many multimodal applications such as image captioning and image description/comprehension.

The main contributions of this work are: (1) the LinkPICS text-image aligner, (2) the G1 aligned corpus and (3) the LinkPICS's database containing pairs of <named entity, face> and <physical word, object> which can be incrementally increased and also exported to a visual dictionary which could be useful for many human and automatic applications.

As possible extensions of this work we give emphasis to four: (1) the extension of the text processing phase to be able to output multiword and other expressions, such as the referring expressions of [15]; (2) the extension of LinkPICS's alignment process to allow n:1 alignments in which several words could be aligned with the same bounding box, (3) alignment of landscapes (e.g., buildings, mountains, trees) and (4) alignment of objects which are hard to identify (e.g., drugs).

Acknowledgments. This research is part of the MMeaning project, supported by São Paulo Research Foundation (FAPESP), grant #2016/13002-0, and was also partly funded by the Brazilian Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

1. Al-Rfou, R., Kulkarni, V., Perozzi, B., Skiena, S.: POLYGLOT-NER: massive multilingual named entity recognition. In: Proceedings of the 2015 SIAM International Conference on Data Mining, pp. 586–594. SIAM (2015)
2. Berg, T.L., et al.: Names and faces in the news. In: 2004 Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, p. II-848. IEEE (2004)
3. Choi, D., Kim, P.: Automatic image annotation using semantic text analysis. In: Quirchmayr, G., Basl, J., You, I., Xu, L., Weippl, E. (eds.) CD-ARES 2012. LNCS, vol. 7465, pp. 479–487. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32498-7_36
4. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B (Methodol.) 1–38 (1977)
6. Deschacht, K., Moens, M.F.: Text analysis for automatic image annotation. ACL **7**, 1000–1007 (2007)
7. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-47979-1_7
8. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
9. Fischer, A., Frinken, V., Fornés, A., Bunke, H.: Transcription alignment of Latin manuscripts using hidden Markov models. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pp. 29–36. ACM (2011)
10. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv preprint arXiv:1608.06993 (2016)
11. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07–49, University of Massachusetts, Amherst (2007)
12. Leydier, Y., Eglin, V., Bres, S., Stutzmann, D.: Learning-free text-image alignment for medieval manuscripts. In: 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 363–368. IEEE (2014)
13. Li, L.J., Socher, R., Fei-Fei, L.: Towards total scene understanding: classification, annotation and segmentation in an automatic framework. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2036–2043. IEEE (2009)
14. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
15. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, pp. 11–20 (2016)
16. Noel, G.E., Peterson, G.L.: Context-driven image annotation using ImageNet. In: The Twenty-Sixth International FLAIRS Conference (2013)
17. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. **29**(1), 19–51 (2003)

18. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: an open Brazilian Wordnet for reasoning. In: Proceedings of COLING 2012: Demonstration Papers, The COLING 2012 Organizing Committee, Mumbai, India, pp. 353–360, December 2012. http://www.aclweb.org/anthology/C12-3044. Published also as Technical report http://hdl.handle.net/10438/10274
19. Pham, P., Moens, M.F., Tuytelaars, T.: Linking names and faces: seeing the problem in different ways. In: Proceedings of the 10th European Conference on Computer Vision: Workshop Faces In'real-life'images: Detection, Alignment, and Recognition, pp. 68–81 (2008)
20. Ramisa, A., Yan, F., Moreno-Noguer, F., Mikolajczyk, K.: Breakingnews: article annotation by image and text processing. arXiv preprint arXiv:1603.07141 (2016)
21. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
22. Schmid, H.: Probabilistic part-ofispeech tagging using decision trees. In: New methods in Language Processing, p. 154. Routledge (2013)
23. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, pp. 815–823 (2015)
24. Socher, R., Fei-Fei, L.: Connecting modalities: semi-supervised segmentation and annotation of images using unaligned text corpora. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 966–973. IEEE (2010)
25. Socher, R., Fei-Fei, L.: Connecting modalities: semi-supervised segmentation and annotation of images using unaligned text corpora. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 966–973. IEEE (2010)
26. Tegen, A., et al.: Image segmentation and labeling using free-form semantic annotation. In: ICPR, pp. 2281–2286 (2014)
27. Tirilly, P., Claveau, V., Gros, P., et al.: News image annotation on a large parallel text-image corpus. In: LREC (2010)
28. Tiwari, P., Kamde, P.: Automatic image annotation and retrieval using contextual information (2015)
29. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics (1994)
30. Yin, F., Wang, Q.F., Liu, C.L.: Transcript mapping for handwritten chinese documents by integrating character recognition model and geometric context. Pattern Recogn. **46**(10), 2807–2818 (2013)
31. Zinger, S., Nerbonne, J., Schomaker, L.: Text-image alignment for historical handwritten documents. In: IS&T/SPIE Electronic Imaging, pp. 724703–724703. International Society for Optics and Photonics (2009)

When, Where, Who, What or Why? A Hybrid Model to Question Answering Systems

Eduardo G. Cortes(✉), Vinicius Woloszyn, and Dante A. C. Barone

PPGC, Institute of Informatics, Federal University of Rio Grande Do Sul (UFRGS),
Caixa Postal 15.064, Porto Alegre, RS 91.501-970, Brazil
egcortes@inf.ufrgs.br

Abstract. Question Answering Systems is a field of Information Retrieval and Natural Language Processing that automatically answers questions posed by humans in a natural language. One of the main steps of these systems is the Question Classification, where the system tries to identify the type of question (i.e. if it is related to a person, time or a location) facilitate the generation of a precise answer. Machine learning techniques are commonly employed in tasks where the text is represented as a vector of features, such as bag–of–words, Term Frequency-Inverse Document Frequency (TF-IDF) or word embeddings. However, the quality of results produced by supervised algorithms is dependent on the existence of a large, domain-dependent training dataset which sometimes is unavailable due to labor-intense of manual annotation of datasets. Normally, word embedding presents a related better performance on small training sets, while bag-of-words and TF-IDF presents better results on large training sets. In this work, we propose a hybrid model that combines TF-IDF and word embedding in order to provide the answer type to text questions using small and large training sets. Our experiments using the Portuguese language, using several different sizes of training sets, showed that the proposed hybrid model statistically outperforms bag-of-words, TF-IDF, and word embedding approaches.

Keywords: Question answering · Question classification
Word embedding

1 Introduction

Question answering (QA) is a specific Computer Science task within the fields of Information Retrieval and Natural Language Processing that aims to provide a precise answer to an input question posed by humans in a natural language. A QA system implementation usually involves different areas of Computer Science that vary from advanced natural language processing, information retrieval, knowledge representation, automated reasoning, to machine learning. Typically it includes three main components: (i) ***Question Processing*** which classify

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 136–146, 2018.
https://doi.org/10.1007/978-3-319-99722-3_14

the question in different classes (e.g. person, time and location) and create a query for Information Retrieval (IR) system with information from the question text; (ii) *Passage Retrieval* which recovers the text passages from a collection of documents that likely contains the answer to input question; and (iii) *Answer Processing* which generates a final answer, usually it extracting from the passage of texts to the words for the answer [2,4,8].

Question classification or Answer type recognition is an important stage in a ***Question Processing*** to provide meaningful guidance on the nature of the response required [20]. The task consists in determining the question class, normally related to the 5WH, the acronyms used in the English language for the main types of questions (Who?, What?, Where?, When?, Why?, and How?). For example, the question "Who won the last Nobel Peace Prize?" expects an answer of type PERSON while the question "When did the man step on the moon?" expects an answer of type TIME. Once the question class is determined, this information will be used in the next stages of the QA pipeline [8]. According to [12], this stage is useful in Passage Retrieval stage to determine the research strategy to retrieval candidate passages and useful in Answer Processing in order to select the answers candidates.

Approaches addressed to Answer type recognition commonly rely on rule-based or supervised learning techniques. In rule-based models, hand-written rules are manually created by an empirical observation of the questions to determine patterns in the associated text with the class of question type [6,8]. On the other hand, approaches based on machine learning, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Support Vector Machine (SVM), have been showing excellent results in question classification task [11,13,19,22]. However, the quality of results produced by these supervised algorithms is highly dependent on the existence of a large, domain-dependent training dataset. Normally, word embedding presents a better performance on small training sets, while bag-of-words and TF-IDF present better results on large training set [15].

In this work, we propose to use a combination between TF-IDF and Word2Vec on a hybrid model to question classification, once we have observed that the two approaches complement each other. To evaluate our proposed model, we used a Portuguese dataset [17] and a linear SVM. The results showed that the hybrid model statistically outperforms the individual models in different sizes of tested training sets. The main contributions of this work are listed below:

- a comparison between Word2Vec, TF–IDF and bag–of–words for question classification task using a Portuguese data set.
- a hybrid model for question classification task that statistically outperforms bag–of–words, TF–IDF and Word2Vec using different sizes of training set.
- a end-to-end testing of a QA system on a Portuguese dataset using different strategies for question classification.

The rest of this paper is organized as follows. Section 2 discusses related approaches for question classification task. Section 3 presents details of our proposed hybrid approach. Section 4 describes the design of our experiments, and

Sect. 5 show and discusses the results. Section 6 summarizes our conclusions and presents future research directions.

2 Related Works

We found in the literature several works in question classification task for QA systems. Currently, the majority of works are using approaches based on machine learning. Most of these support the English language but many others are focusing on a non-English language, like Chinese and Spanish.

We have observed that the first approaches in question classification used hand-written rules. In [6] was present the QA Topology that employs about 270 hand-written rules to classify a question in approximately 180 categories. [7] made a hybrid model that uses a rule-based approach to provides semantic features to a classifier. [1] uses rules based on a sequence of terms with the possible types in the text to classify a question.

Nowadays, the rule-based approach is no longer widely used due to the significant effort to create a large number of rules and, currently, the use of machine learning is achieving excellent results to this task [12]. In [27], is proposed a question classification to enhance a QA system that uses an SVM and a question semantic similarity model to classification. The results showed that the approach has the accuracy of 91.49% better than baseline approaches. A hybrid approach is proposed in [16] with a model that combines Wh-words, Wh-words position and question length to increase the accuracy of existing question classification system in Bangla language.

Deep learning has been widely used in question classification task with the models CNN and RNN. [9] reports that a CNN with little hyperparameter tuning and static vectors produces excellent results on different benchmarks. In order to deal with long-range dependencies in LSTM (Long short-term memory) models, [26] proposes a novel architecture that connects continuous hidden states from previous steps to the current step and brings a consideration mechanism to the hidden states. [21] proposes a novel method to model short texts based on semantic clustering and CNN. The results show that the strategy achieves the best performance in datasets TREC and Google snippets. [7] proposed a group of sparse CNN by embedding a neural version of a dictionary learning to represent the input question taking into account the answer set. The results showed that the model outperforms baselines on four datasets.

One important aspect that high influence on the performance of a machine learning model is how the input information is represented. [13] propose to consider answer information in question classification using a novel group sparse CNN in which significantly outperform strong baselines model in four datasets. According to [12], the syntactic and semantic features can usually improve the question classification but augment the number of features can introduce noisy information that can make misclassification. Nonetheless, a common drawback of supervised learning approaches is that the quality of the results is heavily influenced by the availability of a large, domain-dependent annotated corpus to train the model. Unsupervised and semi-supervised learning techniques, on the other hand, are attractive because they do not imply the cost of corpus annotation [18,23–25].

3 Proposed Approach

Regarding machine learning approaches, usually, a document is represented as a vector of features taken from the text. A typical approach is bag-of-words and Term Frequency-Inverse Document Frequency (TF-IDF), where each word from the vocabulary receives a weight value according to the presence in the document. Another typical approach is related to word embedding models where each word is represented in a semantic space, for instance, Word2Vec models [14]. Despite good representation of the semantic space, in our experiments, we observed that Word2Vec models do not represent well keywords in a question sentence for question classification problem since words like "when", "who" or "how" is not well represented as bag–of–words and TF-IDF models. Based on that, our hypothesis is that TF-IDF has a good representation of words that determine the type of answer of a question, but doesn't represent well the semantics while Word2Vec has opposing characteristics. By observing the advantages and disadvantages of there models, it is possible to presume that they complement each other.

A bag-of-words is a conventional model for representing texts that employ a vector of features where each feature represents a word in the vocabulary. Normally, a feature is represented by its word frequency in the text document. Another way to determine the value of a feature is the TF-IDF, where we set a weight to the word according to its frequency in the document (TF) and its inverse frequency in the collection of documents (IDF). The TF-IDF $w_{f,d}$ is represented by:

$$w_{f,d} = freq_{f,d} \times log_{10} \frac{N}{n_f} \tag{1}$$

where $freq_{f,d}$ is the word frequency (feature) f in the document d, N is the total of documents in the dataset and n_f is the number of documents that contains the word f.

An alternative model to overcome the disadvantages of bag–of–words and TF-IDF is Word2Vec [14], a word embedding model that allows making semantic analogies similar to the real world. In order to get advantages from TF-IDF model and Word2Vec model, our approach consists of a combination of two models in a single one that contains the semantic representation of words and TF-IDF representation. A question is a query q represented by a sentence of words $w_1, \ldots, w_n$ where each word w_i has a vector v_i that represents its position in the semantic space. Therefore, the concatenation between the two models can be expressed by $\langle tfidf_q, w2v_{q,v} \rangle$ where:

$$w2v_{q,v} = \frac{1}{n} \sum_{i=1}^{n} v_i \tag{2}$$

$tfidf_q$ is a vector with s dimensions, where s is the size of words vocabulary, and each dimension receives the TF-IDF weight of the word w_i, in the question q. $w2v_{q,v}$ is a vector with 300 dimensions, where each dimension receives the arithmetic average of words in question q for its dimension.

4 Experiment Design

This section describes the dataset used to evaluate the approaches and presents the models used as baselines. For the tests, we used an open domain dataset in the Portuguese language. The baseline models selected are approaches commonly used in question classification task presenting good performances. Due to the low quantity of works in the literature using a dataset in Portuguese with these features, it was not possible to compare the approaches with other.

4.1 Dataset

Most QA systems in the literature are working with English datasets, on the other hand, there are few systems working with the Portuguese language. Most of the system that works with the Portuguese use the collection Chave [17] for testing and validation. This collection provides question with their answers class, a collection of documents to consult, and their respective answers. Also, the collection was built by Linguateca for the tasks in CLEF, two recognized institutions, what makes this dataset a standard for QA in the Portuguese language. The collection was used in CLEF from 2004 to 2008, where the QA systems needed to provide an answer to an input question looking for this answer in a raw text in a set of documents made available by the collection.

The Chave collection contains about 4000 questions in Portuguese where about 1000 provide at least one response. Each question has included a category and type as well as other information such as identification code and year of creation. This collection is provided by Linguateca, a center of resource for the computational processing of the Portuguese language where, in order to add Portuguese on ad-hoc IR task and QA task, the Linguateca has created the Chave collection. For this research, we selected 2.973 questions and their respective answers class from the Chave collection. We did not use all the questions because some did not provide the correct answer class and also some questions were discarded due to the low number of samples. The distribution of the dataset used is depicted in Table 1.

The Word2Vec model used in this works was obtained in NILC-Embeddings, a repository that storage and sharing word embeddings models generated for the Portuguese language. We used the Word2Vec model with 300 dimensions and trained with continuous bag-of-words from a large corpus of Brazilian Portuguese and European Portuguese, from varied sources [5].

Table 1. Dataset class distribution

Answer type (class)	Amount
DEFINITION	624
LOCATION	545
MEASURE	502
ORGANIZATION	356
PERSON	582
TIME	364
Total	**2973**

4.2 Baselines

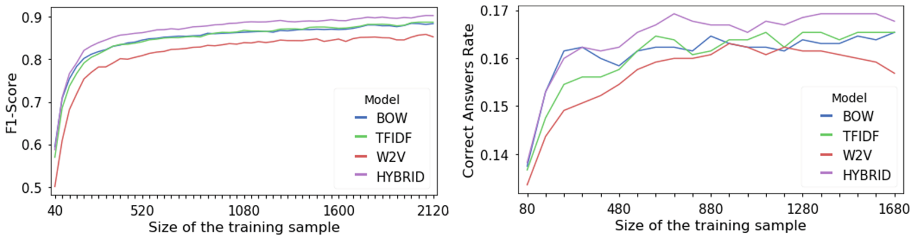
A conventional approach for question classification is bag-of-words, TF-IDF or word embedding with an SVM as classifier. Normally, on this task, SVM outperforms other models like Naive Bayes and Decision Tree [27]. Using the collection Chave for question classification, we observed that linear kernel outperforms the other kernels, following the conclusions of [28], which showed that SVM based on linear kernel achieve better results than SVM based on polynomial kernel, RBF kernel or sigmoid kernel. In this way, the model used in this work uses a linear SVM from scikit-learn library for Python.

To evaluate our proposed hybrid model, we used three commonly and strong models employed on QA as the baseline with an SVM classifier based on a linear kernel, as follows: (i) Bag-of-words with word frequency, namely *BOW*, where each word in the bag-of-words received its frequency in the question; (ii) TF-IDF, namely *TFIDF*, where each vector dimension received the TF-IDF weight of the word in the question; and (iii) Word2vec, namely *W2V*, where each sentence is mapped as the arithmetic average of the vector of question words in semantic space.

5 Results and Discussions

This section presents the results of tests and discussions. We divided the test into two steps: the first, showed in Fig. 1(a), test the questions classification models; the second, showed in Fig. 1(b), test these models in a complete QA system for the Portuguese language. In order to get reliable results and robust conclusions the results presented were generated varying the size of training samples for the question classification model.

In addition to the baselines models, referred as *BOW*, *TFIDF* and *W2V*, we present here the results from the proposed approach, referred as *HYBRID*, that is a combination between *TFIDF* and *W2V*.



(a) (b)

Fig. 1. (a) F1-Score over different training set size in question classification task. (b) Correct Answers rate in QA system over different question classification model and training set.

5.1 Question Classification Results

In the first test, we measure the models with 53 different sampling sizes, ranging from 40 to 2120 the size of the training set. For each sample interaction, was extracted the corresponding number of samples from the dataset in a random position. This process is repeated 5 times and is performed through the calculation of the arithmetic average of the results with regard to reducing the noise by bias in the results and try as many as possible to train and to test all the dataset. Thus for each interaction, the data that is not used for training will be used for testing.

The measure used in the graphic in Fig. 1(a) for model evaluation is F1-Score, that is the harmonic mean of precision and recall measures. In Table 2 is presented the final precision, recall and F1-Score from results in Fig. 1(a) using 2120 training samples.

The graphic in Fig. 1(a) shows that the *HYBRID* approach got better results than the all other models and in all training sample sizes in question classification task. The difference is 3% points (pp) for F1-Score when compared to the runner-up method, namely *TFIDF*. Regarding the *BOW*, the difference if 3 pp, and *W2V* is **8** pp. Using the Wilcoxon statistical test with a significance level of 0.05, we verified that the results got by *HYBRID* model are statistically superior to all baselines models. The Table 2 shows that the *HYBRID* approach has better results in recall and precision measures when uses 2120 training samples as well.

When comparing the results of baseline models, *BOW* and *TFIDF* had similar results and outperform the model *W2V* (Word2Vec). With an F1-Score about 5 pp lower than another baseline models in most of the training samples, it is possible to assume that *W2V* model alone is not the best option for question classification task. Although the *BOW* and *TFIDF* models got the best baseline results, in question classification task the amount of information get from an input question is lower than a complete text, so the *W2V* model can extract and complement a model with important semantic information.

While bag–of–words and TF-IDF can handle with important words like “Who” or “When”, Word2Vec can handle with problems related with semantic as synonyms. Thus, based on results, it is possible to assume that the combination of TF-IDF with a Word2Vec model can handle with the disadvantages that the individual models cannot deal.

Table 2. Recall, Precision and F1-Score with 2120 training samples

Model	Recall	Precision	F1-score
BOW	0.8727	0.8717	0.8685
TFIDF	0.8757	0.8752	0.8724
W2V	0.8487	0.8472	0.8462
HYBRID	**0.8964**	**0.8921**	**0.8923**

In Fig. 2 is shown the F1-Score of each class for each question classification model. The classes PERSON and ORGANIZATION had the worst results once we observed that this type of question has a similar syntactic and semantic structure. For example, the question "Who won the last championship?" can expect the name of a people (PERSON class) or a soccer team (ORGANIZATION class). Also, is possible to observe that the *HYBRID* can join the bests performances of *TFIDF* and *W2V* models once it is the merge between the two models.

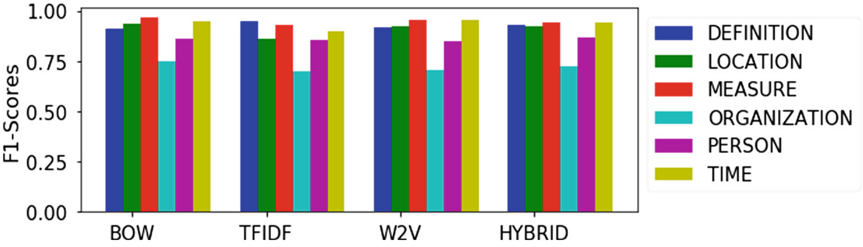


Fig. 2. F1-Score of each class and for each question classification model using 2120 samples for training.

5.2 QA System Results

In order to measure the proposed model in a complete QA system, we have built a simple QA system for Portuguese that used the models tested in the step of question classification. The QA system was built for open domain and data unstructured, following a default QA architecture [8]. In this task, we expect to get a graphic with similar results to the first test, once the question classification is an important step in QA system pipeline. The results from QA system should reflect the question classification model performance. The second test has the same configuration as the first one, except that the measure used for evaluation was the accuracy of the answers of the system. This test uses 21 different sampling sizes, ranging from 80 to 1680 the size of the training set.

The QA system developed for this test uses approaches in each pipeline steps once in this work the main objective is the question classification evaluation. The approaches used in each step of the QA system is described as follow:

- Question Processing: Each question classification models exposed in this work was used for the question classification task. The query for IR system was generated considering all words in the question that was not a stopword (irrelevant word).
- Information Retrieval: The Solr search platform was used to index and query the documents from the collection Chave. For passage retrieval, each phrase from the retrievals documents that contain at last one entity of the same type of question class will be selected. For named entity recognition is used the collection Harem [3] to train a Conditional Random Field model [10] to identify and classify the entities of each passage.
- Answer Processing: This step retrieves the entities from retrieved passages with the same class type as the question. Thus, is created a rank of these entities ranked by the votes and document rank where the passage and entity were retrieved.

The question classification approaches used in the QA system has a good performance, however, the other steps have low performance compared to it. Thus, the results of this test must have much more noise than the first test. Even so, the graphic in Fig. 1(b) shows that the performances of this test followed a similar behavior than the first one, adding more reliability in the results obtained in the first test. Regarding the accuracy, the *HYBRID* model shows values statically relevant, where the differences are 4 pp better than the runner-up baseline. We verified that the results got by the model are statistically superior using the Wilcoxon statistical test with a significance level of 0.05.

6 Conclusion

In this work, we proposed a hybrid model combining the features from TF-IDF and Word2Vec to represent texts for question classification task, an important step in a Question Answering system. To evaluate our approach, we used a linear Support Vector Machine classifier for all tested models, including the proposed approach. We used the dataset Chave, a Portuguese collection of questions and their respective answers where each question in the collection contained the type of response information.

For evaluating, we have varied the size of training samples generating a graphic with the F1-Score of each model for each sample. The results obtained in our experiments, testing the models individually and in a full QA system, showed that the proposed Hybrid approach overcomes the performance of the other baseline models evaluated, which indicates that the concatenation approach between TF-IDF and Word2Vec is promising in the question classification task for Question Answering systems.

For future works, we consider evaluating our approach using others dataset in another language, for instance in English, in order to compare our approach with other known models with a more used dataset. Also, we intend to use this approach in Deep Learning models, like LSTM and CNN.

References

1. Amaral, C., et al.: Priberam's question answering system in QA@CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 364–371. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85760-0_46
2. Cavalin, P., et al.: Building a question-answering corpus using social media and news articles. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 353–358. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_36
3. Freitas, C., Mota, C., Santos, D., Oliveira, H.G., Carvalho, P.: Second harem: advancing the state of the art of named entity recognition in portuguese. In: LREC. Citeseer (2010)
4. Gonçalves, P.N., Branco, A.H.: A comparative evaluation of QA systems over list questions. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 115–121. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_11
5. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: evaluating on word analogies and natural language tasks. arXiv preprint arXiv:1708.06025 (2017)
6. Hovy, E., Hermjakob, U., Ravichandran, D.: A question/answer typology with surface text patterns. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 247–251. Morgan Kaufmann Publishers Inc. (2002)
7. Huang, Z., Thint, M., Qin, Z.: Question classification using head words and their hypernyms. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 927–936. Association for Computational Linguistics (2008)
8. Jurafsky, D., Martin, J.H.: Speech and Language Processing, vol. 3. Pearson, London (2014)
9. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
10. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data (2001)
11. Lee, J.Y., Dernoncourt, F.: Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint arXiv:1603.03827 (2016)
12. Loni, B.: A survey of state-of-the-art methods on question classification (2011)
13. Ma, M., Huang, L., Xiang, B., Zhou, B.: Group sparse CNNs for question classification with answer sets. arXiv preprint arXiv:1710.02717 (2017)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)

15. Mouriño-García, M., Pérez-Rodríguez, R., Anido-Rifón, L., Gómez-Carballa, M.: Bag-of-concepts document representation for Bayesian text classification. In: 2016 IEEE International Conference on Computer and Information Technology (CIT), pp. 281–288. IEEE (2016)
16. Nirob, S.M.H., Nayeem, M.K., Islam, M.S.: Question classification using support vector machine with hybrid feature extraction method. In: 2017 20th International Conference of Computer and Information Technology (ICCIT), pp. 1–6. IEEE (2017)
17. Santos, D., Rocha, P.: The key to the first CLEF with Portuguese: topics, questions and answers in CHAVE. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 821–832. Springer, Heidelberg (2005). https://doi.org/10.1007/11519645_80
18. dos Santos, H.D., Ulbrich, A.H.D., Woloszyn, V., Vieira, R.: DDC-Outlier: Preventing medication errors using unsupervised learning. IEEE J. Biomed. Health Inform. (2018)
19. Sarrouti, M., El Alaoui, S.O.: A machine learning-based method for question type classification in biomedical question answering. Methods Inf. Med. **56**(03), 209–216 (2017)
20. Solorio, T., Pérez-Coutiño, M., Montes-y-Gómez, M., Villaseñor-Pineda, L., López-López, A.: Question classification in Spanish and Portuguese. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 612–619. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-30586-6_66
21. Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F., Hao, H.: Semantic clustering and convolutional neural network for short text categorization. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), vol. 2, pp. 352–357 (2015)
22. Wang, S., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pp. 90–94. Association for Computational Linguistics (2012)
23. Woloszyn, V., Machado, G.M., de Oliveira, J.P.M., Wives, L., Saggion, H.: Beatnik: an algorithm to automatic generation of educational description of movies. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), vol. 28, p. 1377 (2017)
24. Woloszyn, V., Nejdl, W.: Distrustrank: spotting false news domains. In: Proceedings of the 10th ACM Conference on Web Science, pp. 221–228. ACM (2018)
25. Woloszyn, V., dos Santos, H.D., Wives, L.K., Becker, K.: MRR: an unsupervised algorithm to rank reviews by relevance. In: Proceedings of the International Conference on Web Intelligence, pp. 877–883. ACM (2017)
26. Xia, W., Zhu, W., Liao, B., Chen, M., Cai, L., Huang, L.: Novel architecture for long short-term memory used in question classification. Neurocomputing **299**, 20–31 (2018)
27. Xu, J., Zhou, Y., Wang, Y.: A classification of questions using SVM and semantic similarity analysis. In: 2012 Sixth International Conference on Internet Computing for Science and Engineering (ICICSE), pp. 31–34. IEEE (2012)
28. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 26–32. ACM (2003)

Indexing Names of Persons in a Large Dataset of a Newspaper

Juliana P. C. Pirovani, Matheus Nogueira, and Elias de Oliveira(✉)

Programa de Pós-Graduação em Informática,
Universidade Federal Do Espírito Santo, Av. Fernando Ferrari 514 – Goiabeiras,
Vitória, ES 29075-910, Brazil
juliana.campos@ufes.br, elias@lcad.inf.ufes.br

Abstract. An index is a very good tool for finding the necessary information from a set of documents. So far, the extant index tools in both the printed and digital newspaper versions are not sufficient to help users find information. Users must browse the entire newspaper to fulfill their needs or discover later on, after spending a considerable amount of energy, that the information they had been seeking is not available. We propose here to use state-of-the-art strategies for extracting named entities specifically for person names and, with an index of names, provide the user with an important tool to find names within newspaper pages. The state-of-the-art system considered used the Golden Collection of the First and Second HAREM, a reference for Named Entity Recognition systems in Portuguese, as training and test sets respectively. Furthermore, we created a new training dataset from the actual newspaper's articles. In this case, we processed 100 articles of the newspaper and managed to correctly find 87.0% of the extant names and their respective partial citations.

1 Introduction

The identification of a person's name in free written text is a tough task within the Named Entity Recognition (NER) research area. It has been more than twenty years since the *Message Understanding Conference – 6* (MUC-6) added the task of identification and classification of the Named Entities (Named Entity Task – NE) [8], due to its importance in promoting the development in this area.

This task can be unfolded into three other subtasks: entity names, temporal expressions, and number expressions. We are interested in the first group, which we tagged with `<EM ID="xxx" CATEG="PESSOA">` when found in the text, where ID is a unique identifier for this occurrence.

This problem is relevant in many real-life areas. For instance, according to [2], NER is a fundamental step of preprocessing tasks in Information Retrieval (IR) for several other applications (e.g., relation and event extraction). Another example is the work developed in [12], where the objective is to extract gene, protein and other biomedical entities mentions using a machine learning algorithm.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 147–155, 2018.
https://doi.org/10.1007/978-3-319-99722-3_15

The problem can get more complex depending on the domain of the texts one aims to work on. This is the case in texts in social media, where there is no strict pattern to refer to names, in particular people's names. A similar situation is the extraction of names from newswired texts. Names can appear in many forms: (a) *complete* – where all parts are always presented, the first, middle name and surname, and all of them have initial capitalization; (b) *partial* – where sometimes only the surname is used, or alternatively a combination of the use of the first name together with the surname; and (c) *a reference* – this is the case when one decides to use a nickname instead of using the person's real name.

The main approaches used by systems that automatically identify names in texts are: (a) linguistic methods based on manually built rules [6], (b) probabilistic methods based on machine learning strategies [4], and (c) hybrid methods that combine both methods [17]. In 2004, the linguistic approach was considered by [7] as the most frequent approach used for NER. The rules of this approach allow one to build a very neat local grammar (LG) for the specific problem at hand. In 2012, the most recent works in NER already used statistical machine learning methods such as Hidden Markov Models and Conditional Random Fields [2].

In this paper, we adopted a hybrid approach proposed in [17] to find person names from pages of a local online newspaper called "A Tribuna"[1], in Vitória – Espírito Santo, Brazil. In addition to presenting state-of-the-art results for Portuguese, the strategy combines advantages of the linguistic and machine learning approaches. The names found were used to create a webpage of person names index. This webpage allows the user to easily access all the newspaper articles where their names appear. This webpage will be also used to improve the proposed strategy by offering the user the opportunity of pointing us out a name we have missed out from any newspaper's page, when this happens.

This paper is organized as follows. In Sect. 2, we present a brief revision of the main related works to what we are pursuing here. Next, we describe our work methodology, the metrics we used and give an idea of the application we built as a result of our work. Next, in Sect. 4, we describe the process of experimentation we used to achieve the results we discuss in this work as well as the new training dataset we introduced as a possible benchmark for future works. To sum up, in Sect. 5, our conclusions and future work are presented.

2 The Literature Review

In [5], some linguistic properties of proper names considering European Portuguese are presented. The goal is to assist the automatic processing of texts in this language. Some properties of the formal variations (concordance of number and gender) are also presented, as in this language some names accept the plural form (*e.g.*, *Antônio* is smart, the singular, and – The *Antônio*s are smart, the plural form). In addition, some combinatorial restrictions are also presented as

[1] https://tribunaonline.com.br/.

the existence of prepositions between names (*e.g.*, Maria *de* Lurdes) and the existence of a few compound nouns connected by a hyphen (*e.g., Jose-Maria*). This information is represented by the finite state automata and, furthermore, they presented a proposal to formalize a way to describe people's names in dictionaries.

One related work regarding identifying names in newspaper articles written in French is described in [7]. In this work, the author carried out a linguistic text analysis to build a series of cascaded finite state transducers in which each one is capable of transforming input texts into other suitable texts for some information extraction. This analysis sought to identify the context *prior to* and *after* people's names. The system, called CasSys, was implemented upon the Intex[2] tool. Two are the transducers used for extracting names in [7]. The first one uses a list of rules describing a local grammar and the second uses the names identified by the first transducer to search for the remaining names.

Local Grammars (LG) were created by [11] for use in an NER system in the Serbian language. The system uses electronic dictionaries with a rich set of semantic tags and LGs that describe the context of NEs to recognize: person names, geopolitical names, temporal and numerical expressions. They produced special LGs for recognizing a person's position in society. The system was evaluated in short agency news, and Recall and Precision metrics were manually computed. The results suggest that the system prioritizes Precision.

Another approach presented by [6] is designed to identify person names in Portuguese texts using an LG created based on a linguistic study of these texts. Initially, the authors built a LG for the book titled Senhora by José de Alencar, and then later applied this same grammar to articles to the A Tribuna, a local newspaper in Espírito Santo. The goal was to observe the appropriation of an LG built from one context to another. While this is possible, some adaptations were necessary. The performance achieved for the newspaper articles was lower because the LG was not built specifically for that corpus owing to its particularities in the ways names were presented, confirming that the automatic identification of names is corpus dependent.

On the other hand, some machine learning approaches (e.g., [4,14,16]) have used machine learning techniques such as Hidden Markov Models (HMM), Transformation-Based Learning (TBL), Support Vector Machines (SVM), Conditional Random Fields (CRF), Naive Bayes and Decision Table for NER over Portuguese texts.

Despite these two basic approaches, some hybrid approaches are found in the literature. For instance, the CRF for Portuguese NER was used by [18] to identify and classify the 10 categories of HAREM[3] NEs. The IO [10] notation with the HAREM-defined categories, the corpus annotated with part-of-speech tags (POS-tagging) and a feature vector are used as input for the training phase. In the testing phase, the HAREM-defined categories are removed. The HAREM corpora was used for training and testing.

[2] http://www.nyu.edu/pages/linguistics/intex/.

[3] http://www.linguateca.pt/HAREM/.

In this paper, we show a very useful application of NER. This application is in a form of an index of person names, identified by a hybrid approach considering the specificity of Portuguese language. This is our first step toward a more general research goal, which consists of the automatic knowledge base extraction from free texts.

3 The Methodology

In our case study, a name can appear many times within the newspaper article as the journalist decides to cite a person more than once. Our tool will tag all the occurrences and build a webpage with all of them highlighted so that the user can easily locate the name. Figure 1 shows the step-by-step process to mark the names on the newspaper page.

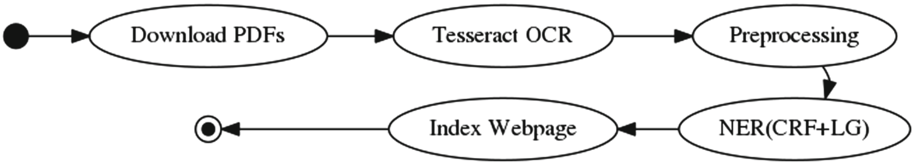


Fig. 1. The processing flow for extracting names from the online newspaper

The first step is done by daily downloading PDFs of the newspaper articles from the A Tribuna newspaper public site. These files are scanned by the Tesseract API [1]. Tesseract is an open source tool that performs Optical Character Recognition (OCR), which allows us to obtain plain ASCII searchable-text files from the PDFs. After this, these texts are preprocessed. This is the process in charge of tasks such as removing empty lines and accounting for the hyphens at the ends of lines, which are common in newspaper articles due to the column layout style [15]. All these steps were carried out by a set of shell-script codes, which we are planning to make fully automatic soon.

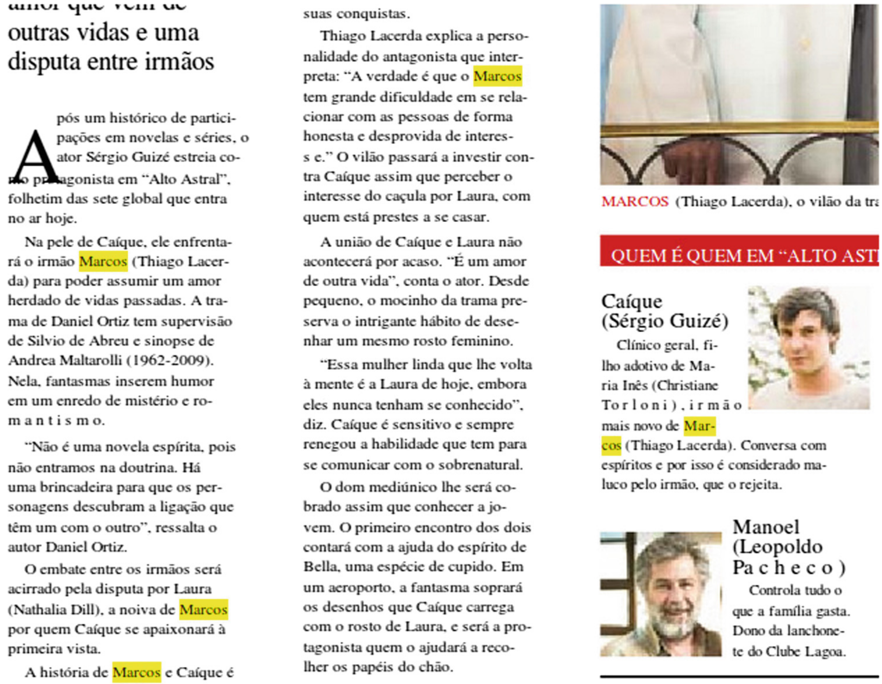
The next step is the Named Entity Recognition (NER). To the best of our knowledge, the state-of-the-art solution for the extraction of names in Portuguese language is the work presented in [17]. This is a hybrid approach where Conditional Random Fields (CRF) is used in combination with handmade Local Grammars (LGs) to capture the logic behind the process of naming recognition. Their results significantly surpass previous results [3,20]. In addition, [19] also successfully used this strategy at automatic question generation from entities named.

CRF is a probabilistic method for structured prediction proposed by [13], which has been successfully used in several Natural Language Processing (NLP) tasks and LGs [9] are one means of representing rules of the linguistic approach in which NEs can appear. The classification obtained from LG is sent as an additional feature for the learning process of the CRF prediction model. That is, the classification obtained from LGs can be seen as a suggestion for the CRF.

The CRF model was trained using the HAREM corpus. HAREM is a joint assessment for Portuguese and the annotated corpora used in the First and Second HAREM, known as the Golden Collections (GC), have served as a golden standard reference for NER systems in Portuguese. The HAREM standard was used for the NE marking. Thus, after the NER, each text sentence has the NE annotated between the <EM> and </EM> tags containing the NE category PESSOA (PERSON) as in the following example:

```
   According to the author, <EM ID=‘‘H2-bbb-3’’ CATEG=‘‘PERSON’’>
José Mourinho </EM> is different because of a new paradigm of
thought.
   (Segundo o autor, <EM ID=‘‘H2-bbb-3’’ CATEG=‘‘PESSOA’’> José
Mourinho </EM> é diferente por partir de um novo paradigma de
pensamento.)
```

Our system can be seen on the website[4]. Figure 2 is an example of how the results are shown to the user. First, the user looks for the target name on the project webpage. On the right side of each name there are some links (pages of



amor que vem de outras vidas e uma disputa entre irmãos

Após um histórico de participações em novelas e séries, o ator Sérgio Guizé estreia como protagonista em "Alto Astral", folhetim das sete global que entra no ar hoje.

Na pele de Caíque, ele enfrentará o irmão Marcos (Thiago Lacerda) para poder assumir um amor herdado de vidas passadas. A trama de Daniel Ortiz tem supervisão de Silvio de Abreu e sinopse de Andrea Maltarolli (1962-2009). Nela, fantasmas inserem humor em um enredo de mistério e romantismo.

"Não é uma novela espírita, pois não entramos na doutrina. Há uma brincadeira para que os personagens descubram a ligação que têm um com o outro", ressalta o autor Daniel Ortiz.

O embate entre os irmãos será acirrado pela disputa por Laura (Nathalia Dill), a noiva de Marcos por quem Caíque se apaixonará à primeira vista.

A história de Marcos e Caíque é

suas conquistas.

Thiago Lacerda explica a personalidade do antagonista que interpreta: "A verdade é que o Marcos tem grande dificuldade em se relacionar com as pessoas de forma honesta e desprovida de interesse." O vilão passará a investir contra Caíque assim que perceber o interesse do caçula por Laura, com quem está prestes a se casar.

A união de Caíque e Laura não acontecerá por acaso. "É um amor de outra vida", conta o ator. Desde pequeno, o mocinho da trama preserva o intrigante hábito de desenhar um mesmo rosto feminino.

"Essa mulher linda que lhe volta à mente é a Laura de hoje, embora eles nunca tenham se conhecido", diz. Caíque é sensitivo e sempre renegou a habilidade que tem para se comunicar com o sobrenatural.

O dom mediúnico lhe será cobrado assim que conhecer a jovem. O primeiro encontro dos dois contará com a ajuda do espírito de Bella, uma espécie de cupido. Em um aeroporto, a fantasma soprará os desenhos que Caíque carrega com o rosto de Laura, e será a protagonista quem o ajudará a recolher os papéis do chão.

MARCOS (Thiago Lacerda), o vilão da tra

QUEM É QUEM EM "ALTO AST

Caíque (Sérgio Guizé)

Clínico geral, filho adotivo de Maria Inês (Christiane Torloni), irmão mais novo de Marcos (Thiago Lacerda). Conversa com espíritos e por isso é considerado maluco pelo irmão, que o rejeita.

Manoel (Leopoldo Pacheco)

Controla tudo o que a família gasta. Dono da lanchonete do Clube Lagoa.

Fig. 2. A screenshot of a section of the newspaper page

[4] http://vitoriavirtual.com.br/indexingNewsPapers/.

the newspaper where the searched name appears) to take the user to another webpage where they can see the name highlighted, as presented in Fig. 2. In this figure, we searched for the name *Marcos*. Our search was nearly 100% accurate. We missed the name *MARCOS* on the bottom left side of the first picture on the page, as it appears in all capital letters.

4 The Experiments

We conducted some initial experiments to evaluate the performance of the NER approach used in the articles from the A Tribuna newspaper since the experiments performed in [17] used the HAREM corpus as a test dataset and considered all the categories of the HAREM.

We prepared a set of 100 news articles to annotate from the actual newspaper we are interested in – the *A Tribuna* corpus. The articles were randomly selected among the articles of politics and economics and the resulting text contains 101733 words. We asked a number of undergraduate students to annotate all the person's name in all the articles by using `Etiquet(H)arem`[5]. The students found 2714 person's name in the 100 documents. The metrics of Precision (P), Recall (R) and F-Measure (F) were computed using the evaluation scripts from the Second HAREM.

In the first experiment, the 100 new annotated articles were used as a test dataset. We applied the LG built by the authors in [17] solely, and we also applied the CRF+LG proposed by [17] for the NER in the *A Tribuna* corpus. CRF+LG was applied considering the GC of the First HAREM as training set and considering all GCs of the HAREM (First HAREM, Mini HAREM and Second HAREM) as training set. We show these results (*Experiment 1*) in Table 1.

Table 1. Evaluation of LG and CRF+LG

Experiment 1	P (%)	R (%)	F(%)
LG	83.51	25.52	39.10
CRF+LG (Training: GC of the First HAREM)	76.15	29.19	42.20
CRF+LG (Training: GCs HAREM)	79.85	40.47	53.71
Experiment 2			
CRF+LG (Training: original ATribuna)	82.70	47.16	60.07
CRF+LG (Training: improved ATribuna)	87.34	48.21	62.13

The Recall value obtained by LG individually was lower because LG captures only some general heuristics for NER and because this LG was not built specifically for the *A Tribuna* corpus that has its own particularities concerning the way the names are written. As expected, the gain obtained (11.5% in F-measure)

[5] http://www.linguateca.pt/poloCoimbra/recursos/etiquetharem.zip.

by CRF+LG using all GCs of the HAREM as training was higher in comparison to the training using only the GC of the First HAREM because the training set is much larger and allows the number of entities identified (Recall) to increase considerably.

After the first experiment, we also considered using articles of the *A Tribuna* corpus for training and testing because we know that NER depends on the corpus. Then, in the second experiment, we defined training and test subsets for the *A Tribuna* corpus using the holdout sampling method. We used the most common split, 2/3 of the corpus was used for training and 1/3 for testing. The results can be seen in Table 1 (*Training: original ATribuna*).

Note that the results obtained outperform the best results of the first experiment in almost 3% for the Precision metric and more than 6% for the Recall and F-measure metrics, representing considerable gain.

By analyzing the false positives and false negatives, we observed that some persons' names were correctly annotated by CRF+LG but were considered wrong when computing the Precision metric because they were not annotated by the students. That is, the students missed some names during the annotation. We thus annotated these names after obtaining a new *A Tribuna* annotated corpus and performed this experiment again. The results presented in Table 1 (*Training: improved ATribuna*) shows that we achieved a gain of approximately 5% in the Precision metric.

Despite preliminary results, we consider them to be promising. We believe that with a larger annotated corpus, the Recall value would be further increased. In addition, a further improvement of the LG to recognize additional names in the *A Tribuna* corpus can enable us to achieve even better results. We also observed that the automatic tool, CRF+LG, can be used for debugging in the process of building a good training dataset.

5 Conclusions

We showed in this work a combined approach for indexing peoples' names within newspaper pages. For this aim, we trained our algorithms with the HAREM very well-known dataset collection for benchmarks and, in addition, we created a new collection for training and test from actual articles extracted from the newspaper.

We tested our approach over 100 newspaper pages. The names we searched for are now posted on our web page, where any user can browse for a name and then go straight to the point where the name is cited within the newspaper page.

The quality of the results we produced is promising, as on average, we yielded a 87.34% of Precision and 48.21% of Recall which will be improved as we train CRF+LG from a larger annotated corpus. The index of names gives us a powerful tool to help the experts find newspapers articles which mention given target names.

Our plans for the future of this work are as follows: first, we want to better improve our capacity of quickly building tailored LGs for the identification of person names. For the current work, we already used the concordance comparison tool built in the Unitex as a computational aid in the manual composition of LGs. We claim that this is a promising tool to be mastered and combined within our automatic framework. Second, we want to improve the process of building and correcting the training data from a given newspaper corpus by being able to incrementally learn from the comparison of the human versus the automatic annotation approach.

Acknowledgments. We would like to thank the reviewers for their insightful comments on the paper, as these comments led us to an improvement of the work.

References

1. Tesseract (2015). https://code.google.com/p/tesseract-ocr/. Acesso em 06 May 2015
2. Jiang, J.: Information extraction from text. In: Aggarwal, C., Zhai, C. (eds.) Mining Text Data, pp. 11–41. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-3223-4_2
3. Amaral, D., Fonseca, E., Lopes, L., Vieira, R.: Comparative analysis of Portuguese named entities recognition tools. In: Calzolari, N., et al. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 2554–2558. European Language Resources Association (ELRA), Reykjavik, May 2014
4. Amaral, D., Vieira, R.: O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa. In: Proceedings of the IX Brazilian Symposium in Information and Human Language Technology - STIL, Fortaleza, CE, October 2013
5. Baptista, J.: A local grammar of proper nouns. Seminários de Linguística **2**, 21–37 (1998)
6. Campos, J., Oliveira, E.: Extração de Nomes de Pessoas em Textos em Português: uma Abordagem Usando Gramáticas Locais. In: Computer on the Beach 2015. SBC, Florianópolis, March 2015
7. Friburger, N., Maurel, D.: Finite-state transducer cascades to extract named entities in texts. Theor. Comput. Sci. **313**(1), 93–104 (2004)
8. Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: Proceedings of the 16th Conference on Computational Linguistics, COLING 1996, vol. 1, pp. 466–471. Association for Computational Linguistics, Stroudsburg (1996)
9. Gross, M.: The construction of local grammars. In: Roche, E., Schabes, Y. (eds.) Finite-State Language Processing, Language, Speech, and Communication, Cambridge, Mass, pp. 329–354 (1997)
10. Konkol, M., Konopík, M.: Segment representations in named entity recognition. In: Král, P., Matoušek, V. (eds.) TSD 2015. LNCS (LNAI), vol. 9302, pp. 61–70. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24033-6_7
11. Krstev, C., Vitas, D., Obradović, I., Utvić, M.: E-dictionaries and finite-state automata for the recognition of named entities. In: Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, FSMNLP 2011, pp. 48–56. Association for Computational Linguistics (2011). ISBN 978-3-642-14769-2

12. Kulkarni, A.: CRF based bio-medical named entity recognition. Int. J. Emerg. Technol. Comput. Sci. **3**(2), 135–139 (2018)
13. Lafferty, J., McCallum, A., Pereira, F.: conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001, vol. 1, pp. 282–289 (2001)
14. Milidiú, R.L., Duarte, J.C., Cavalcante, R.: Machine learning algorithms for Portuguese named entity recognition. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial **11**(36), 65–75 (2007)
15. Nogueira, M., Oliveira, E.: Estratégias de Correção de Erros de Extratores de Palavras em Português. In: 5^{th} Symposium on Knowledge Discovery, Mining and Learning - KDMILE. SBC, October 2017
16. Pellucci, P.R.S., Paula, R.R.d., Silva, W.B.d.O., Ladeira, A.P.: Utilização de Técnicas de Aprendizado de Máquina no Reconhecimento de Entidades Nomeadas no Português. e-Xacta **4**(1), 73–81 (2011)
17. Pirovani, J.P.C., de Oliveira, E.: CRF+LG: a hybrid approach for the portuguese named entity recognition. In: Abraham, A., Muhuri, P.K., Muda, A.K., Gandhi, N. (eds.) ISDA 2017. AISC, vol. 736, pp. 102–113. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76348-4_11
18. Pirovani, J., Oliveira, E.: Portuguese Named Entity Recognition using Conditional Random Fields and Local Grammars. Miyazaki, Japan (2018)
19. Pirovani, J., Spalenza, M., Oliveira, E.: Geração Automática de Questões a Partir do Reconhecimento de Entidades Nomeadas em Textos Didáticos. In: XXVIII Simpósio Brasileiro de Informática na Educação (SBIE). SBC, Ceará (2017)
20. Santos, C.D., Zadrozny, B.: Learning character-level representations for part-of-speech tagging. In: Proceedings of the 31^{st} International Conference on Machine Learning - ICML, pp. 1818–1826, Beijing, China (2014)

ACA - Learning with Alternative Communication

Maria Renata de Mira Gobbo[1](✉),
Cinthyan Renata Sachs C. de Barbosa[1](✉),
José Luiz Villela Marcondes Mioni[1](✉), and Fernanda Mafort[2](✉)

[1] Department of Computer Science,
State University of Londrina, Londrina, Brazil
mr.gobbo1@gmail.com, sachs.cinthyan@gmail.com,
luiz.vmm@gmail.com
[2] School of Life Sciences, Pontifical Catholic University of Paraná,
Londrina, Brazil
fernanda-mafort@hotmail.com

Abstract. Children with Autism Spectrum Disorder (ASD) have a language acquisition difficulty that can already be observed in the first years of life. Impairment in spoken language can negatively affect the social development of these children. Some studies acknowledge that even with a delay in language, children with ASD can acquire speech as soon as an intervention is made early in their lives. This paper presents an application called ACA (Learning with Alternative Communication) that would help these children to acquire an essential linguistic repertoire for the accomplishment of their Daily Life Activities (ADLs). ACA is an Android application that uses alternative communication with voice synthesis, which was developed to help children within the autistic spectrum to identify objects, memorize their names, learn phonemes, visualize some morphological information related to some objects and also help them in the syllabic and pre-syllabic phases of the literacy process.

Keywords: Morphology · ASD · Alternative communication
Acquisition of language

1 Introduction

The development of verbal communication in children normally begins at the age of 12 months, when the child already emits sounds repeatedly, giving meaning to them and this should happen to all children at that age, regardless of their culture, social factors, etc. This development can be recognized in two distinct phases [1]: the pre-linguistic, in which only phonemes are vocalized (without words), that persists until the eleventh or twelfth month; followed by the linguistic phase, when the child begins to speak isolated words with understanding. Subsequently, the child progresses in the complexity escalation of expression. This process is continuous and occurs in an orderly and sequential manner, with considerable overlap between the different stages of this

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 156–163, 2018.
https://doi.org/10.1007/978-3-319-99722-3_16

development. When communication development is delayed, initially called language delay, it is clear that there is some developmental deficiency [2].

It should also be noted that in addition to the communication deficit, the child presents impairments in social interaction and behavior, since these patterns are usually part of the diagnosis of Autism Spectrum Disorder (ASD). In these areas we form classic criteria for childhood diagnosis [3] according to the International Classification of Diseases (ICD) and the Statistical Manual of Mental Disorders (DSM) requiring very early, intensive and multidisciplinary intervention [4]. When there is some language development delay, the areas most affected by the delay are, according to [5], pragmatics, semantics [6] and suprasegmental phonology [7], however, other aspects of phonology, as well as morphology and syntax are also changed. The pattern of language compromise of autistic children involves a dissociation between form (language structure) and function (language use), and they are usually more adept at the formal aspects of language, namely syntactic and morphological aspects [8].

In order for the subject with ASD to be able to communicate, [9] suggests the adoption, in the case of pre-verbal or nonverbal subjects, of a program that privileges communication through gestures, words or an alternative or complementary communication system, which allows the subject to communicate their needs and desires, ensuring the reciprocal interaction between the subjects participating in the communication. It adds the importance of taking advantage of the interests of the subject and of providing a calm environment, working with natural contexts so that the acquisitions can be meaningful and spontaneous [10].

Difficulties in communication occur to varying degrees, both in verbal and nonverbal ability to share information with others. Some children with ASD do not develop communication skills (up to 25% of these children will never develop functional speech), while others often repeat words or phrases (echolalia), make pronominal reversal errors ("you" for "I", "I" for "he," etc.), use the words in an unusual way (idiosyncratic), invent words (neologisms) and use ready-made phrases and questions repetitively. Usually the autistic person does not have a conversation and simply talks to someone else. Some use the verbal expression just to ask for things; others do not realize that the listener has no further interest in the subject. The gestures are reduced and little integrated with what is being said. Half of all autistic children, according to [11], develop an understandable speech up to five years, but those who have not, will hardly have an appropriate verbal expression. There are cases of children diagnosed with ASD and associated hyperlexia (very common among the cadres of the former denomination of Asperger's syndrome) [12] perceived the presence of this phenomenon when it first described and defined the ASD and [13] defined the term "hyperlexia" to describe children who read at levels beyond those expected for their age in the presence of disordered oral communication.

In addition, individuals with ASD have difficulties in understanding non-verbal social interactions such as body language, gestures, facial expressions. The impairment of this verbal communication skill may negatively affect the ability of these children to develop meaningful social interactions, causing adaption problems, poor academic performance and lack of affective interaction.

Even with a delay in language, children with ASD, according to [14], can develop at least some oral language, as long as these children undergo some early intervention and in this way, alternative communication devices can bring positive results for verbal communication in individuals with ASD who verbally reproduce the sounds echoed by the devices.

Alternative communication is the term used to describe adapted methods, strategies, and media made to aid or replace communication of people who have inefficient speech and writing [15]. It can be used both to help one person understand what the other wants to say, and it can also be a means of expression.

An application for Android devices has been developed, that will assist children with ASD in the process of communication and literacy in the syllabic and pre-syllabic phases, allowing them to have access to morphological information as well as exercise activities related to Daily Life Activities.

The paper is organized as follows: Sect. 2 presents the theoretical foundation on language and ASD; Sect. 3 presents the methodology addressed in the research; Sect. 4 presents the application development and Sect. 5 presents the conclusions of the work.

2 Language and ASD

Several studies on the acquisition of grammatical morphemes were performed comparing children with ASD and children with typical development, but no differences were found [5]. Some studies pointed out that children with ASD spoke fewer morphemes, especially articles and morphemes related to verbal tense [16]. That is, children with ASD may present difficulties in the correct use of morphemes and auxiliaries [17].

Autistic children, at the beginning of language development, tend to use specific nouns more often than closed classes of words, such as auxiliary verbs, conjunctions, determinants, prepositions, and pronouns. [18] called our attention to the fact that these children have difficulty in using prepositions, avoiding them and using few pronouns, referring to people by their proper name or another name that describes them, and when they use them, they make frequent inversions, as pointed out in [19].

It is also typical to use verbs without inflection, as well as difficulties in other inflections, especially in the past tense [18], especially in children with more severe degrees of ASD, who also tend to make more errors of verbal inflection and have a shorter vocabulary, using few markers of intention, possibly because of the difficulty in understanding the intentions of the other [20].

In children with ASD it is also common to have difficulties with the small binding words (or connectors), as in “before” and “because”, which are often omitted or misused [21].

Studies indicate [22] that there may be significant lexical increase in autistic children, effective use of oral language and the insertion of new elements in their speech and emission of sentences (subject-verb), presenting significant improvement in their communication after a period of speech therapy.

Children with ASD tend to confuse the meaning of words (this is more frequent with pronouns and prepositions) it is easier for them to learn words about objects than people and emotions, they always prefer concrete concepts rather than abstract [19].

This work tries to provide support for children with ASD to improve their communication skills, especially in the vocabulary that the child already has or needs to acquire in order to communicate in activities that may occur in their daily lives.

3 Methodology

Initially a bibliographical survey was made on the characteristics, peculiarities, strengths and weaknesses that generally the children within the autistic spectrum have. Later, existing literacy methods were studied to see which of them would fit best for children with ASD.

Also, some software were analyzed through the methodology addressed by [23], designed for people with ASD, to verify if there was a need to create a new application or if the existing ones were enough to perform interventions that work the ADLs. As none of them worked with ADLs and much less assist the literacy process through these, an application called ACA (***A***_prendendo com_ ***C***_omunicação_ ***A***_lternativa_) has been developed to assist in the alternative communication of children within the autistic spectrum.

The literacy methodology implemented in the ACA application was the one used by [24, 25] in his studies with children with ASD. In [25] the TEACCH intervention is presented. Thus, the naming of objects will first be taught, since these children have difficulties in naming objects and not being able to name them can generate problems in reading with understanding. Identical image stacking was used to teach such naming so by clicking and dragging a figure from one side of the screen to the other the child will receive the sound-figure stimulus. Using the same technique, the letters of the alphabet and syllables will be taught. Only after that will the full word reading be taught.

Finally, the literacy functionality of the ACA application was developed, which will be validated in a special school for children with ASD.

4 ACA- Learning with Alternative Communication

The game will use pictographs (of the PECS type that are System of Communication by Exchange of Figures), considering the children's difficulty in recognizing a figure by its name. With this association the child can then learn to symbolize.

The use of pictograms in the first stage of literacy works as ideograms, which will bring sense to the text in the future. The color method for the pictograms suggested in [26] will be used. In this method each grammar class will be represented as follows:

- white pictograms will be used for nouns without movement;
- yellow pictograms will be used for articles;
- orange pictograms will be used for verbs;
- gray pictograms for adjectives that do not contain figurative representations;
- light blue pictograms will be used for adjectives;
- white pictograms with pink frames for personal pronouns;

- gray pictograms with pink frames for possessive, interrogative, demonstrative, relative, and indefinite pronouns.

Examples of orange and white pictograms can be seen in Fig. 1.

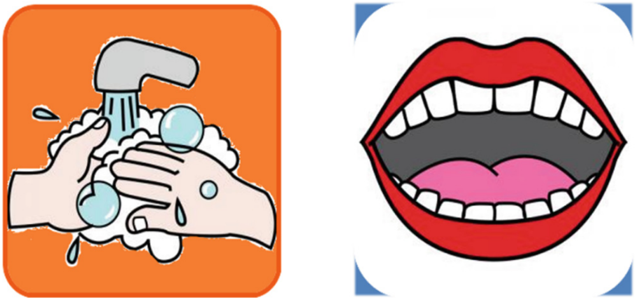


Fig. 1. Examples of the pictographs used. (Color figure online)

ACA software has three features, which will be described next.

One of the features is a digital alternative communication board that has a voice output system. This board is separated into several categories that are part of the ADLs of a child with autism.

ADLs [27–29] refer mainly to the tasks performed by the individuals in their daily life as: food, mobility, environmental control, personal hygiene, clothing, games, and we will also address feelings.

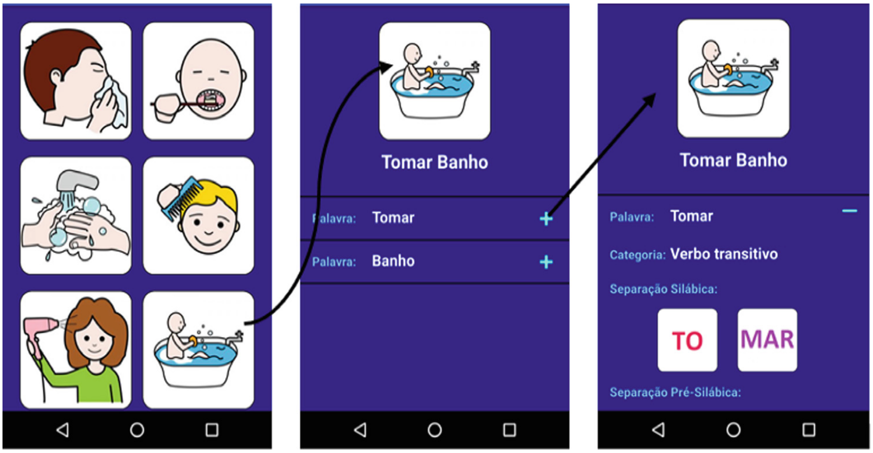
The ADLs adopted on the boards that are part of the corpus are: Food (containing a variety of foods, both solid and liquid, that may be part of the child's diet), Mobility (containing the most common types of transport), Control Clothing (trousers, blouse, T-shirt, sneakers, and other clothing and accessories), personal hygiene (brushing, combing, bathing, etc.), Feelings (fear, joy, sadness, etc.), as well as Objects of the School Context (pencil, eraser, pen, etc.) and Parts of the Human Body (head, arm, leg, etc.).

Another feature is the dictionary that contains all the morphological characteristics of words, such as grades, genders, and grammatical classes that can be used in the application. It also provides syllabic separation (addressing syllabic and pre-syllabic phases) with voice output referring not only to letters, but also to syllables. Still, it associates the words with the images (these are as intuitive and as colorful as possible, to attract the children's attention). In cases of autistic spectrum it is important to have a support for their verbal communication, mainly to favor the linguistic development that the child can possess.

The images present in the system have been downloaded from the ARASAAC (Aragonese Portal of Augmentative and Alternative Communication), a portal that offers graphic materials free of charge to facilitate the communication of people who have some type of special need.

The third functionality will be for the literacy of the children with ASD always using the corpus of ADLs.

The system works as follows: the user chooses one of the features (which have already been mentioned above), and in this category will be all the images related to it. In Fig. 2 this first screen is represented by the number 1. When the user selects that image by means of a click (represented by the arrows in Fig. 2) it will open the words referring to the image. Once you click on the "+" icon, they will open the information about the morphology of the selected word. For instance, if it is a verb, only the category, syllabic and pre-syllabic separator will appear. If it is a word like a noun or an adjective, for example, the category, grade, gender, syllabic and pre-syllabic separation will appear. The user has the option to listen to all the information from the name of the selected image, such as each syllable or letter displayed on the screen, just by clicking on it.



Tomar Banho

Palavra: Tomar +

Palavra: Banho +

Tomar Banho

Palavra: Tomar −

Categoria: Verbo transitivo

Separação Silábica:

TO MAR

Separação Pré-Silábica:

Fig. 2. ACA application screens

In Fig. 2, for example, representation through a picture of a person taking a shower may help children with ASD to have less resistance to sensory activities, such as taking a shower, brushing their teeth or hair, etc.

5 Conclusions

In this work the use of a dictionary in an application called ACA was approached to facilitate the process of memorizing words for children with ASD diagnosis. As a consequence such an application may help such children to increase their vocabulary and also to identify the concrete objects by the correct name.

Another feature of ACA is the alternative communication through ADLs, which has been fully implemented for Android. We also have the literacy function, being developed through the expansion of its linguistic repertoire. Many application developers targeting children on the autistic spectrum forget that it is necessary not only to name objects (because of the difficulty individuals with ASD have with naming) and to provide images of words (which is quite attractive indeed because these children are very visual), but also provide the sounds of the words, helping in the process of issuing the words, so that there is not only need of alternative communication through clues, PECS, etc. Considering that many autistics have great "memory", a software that reproduces words as well as syllabic separation can be very valuable in the literacy process (since sound is as important as picture in this process), thus defining a truly remarkable difference in the ACA Application.

As future works, as soon as authorized, the application will be validated with children of a specialized school in children diagnosed with ASD.

References

1. Schirmer, C.R., Fontoura, D.R., Nunes, M.L.: Distúrbios da Aquisição da Linguagem e de Aprendizagem. Jornal da Pediatria **80**(2), 95–103 (2004)
2. Eigsti, I.M., de Marchena, A.B., Schuh, J.M., Kelley, E.: Language acquisition in autism spectrum disorders: a developmental review. Res. Autism Spectr. Disord. **5**(2), 681–691 (2011)
3. Camargos, Jr. W.: Transtornos Invasivos do Desenvolvimento: 3o Milênio/Walter Camargos Jr e colaboradores. Presidência da República, Secretaria Especial dos Direitos Humanos, Coordenadoria Nacional para Integração da Pessoa Portadora de Deficiência, Brasília (2005)
4. Lima, C.B.: Perturbações do Espectro do Autismo: Manual Prático de Intervenção. Lidel (2012)
5. Santos, R.P.R.: A linguagem em crianças com perturbações do espectro do autismo: Análise Morfossintáctica. Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro, Dissertação de Mestrado (2009)
6. Rondal, J.A.: Teoria de la mente Y lenguaje. Revista de Logopedia, Foniatria Y Audiologia **27**(2), 51–55 (2007)
7. Martos, J., Ayuda, R.: Comunicación y lenguaje: bidireccionalidad em La intervención em niños com trastorno de espectro autista. Rev. Neurol. **48**(Supl2), S58–S63 (2002)
8. Artigas, J.: Em lenguaje em los trastornos autistas. Rev. Neural **28**(Supl2), S118–S123 (1999)
9. Bergeson, T., et al.: Los aspectos pedagógicos de lós trastornos del espectro autista. [s.n.], Espanha (2003)
10. Foscarini, A.C., Passerino, L.M.: Escalando possibilidades de comunicação e interação em crianças com autismo não oralizadas. In: 5º Congresso Brasileiro de Comunicação Alternativa, Gramado (2013)
11. LIMs – Laboratórios de Investigações Médicas - Laboratório de Neurociências do Instituto de Psiquiatria da Faculdade de Medicina da Universidade de São Paulo (2013)
12. Kanner, L.: Autistic disturbances of affective contact. Nerv. Child **2**, 217–250 (1943)
13. Silberberg, N., Silberberg, M.: Hyperlexia: specific word recognition skills in young children. Except. Child. **34**, 41–42 (1967)
14. Ganz, J.B., Hong, E.R., Gilliland, W., Morin, K., Svenkerud, N.: Comparison between visual scene displays and exchange-based communication in augmentative and alternative communication for children with ASD. Res. Autism Spectr. Disord. **11**, 27–41 (2015)
15. El-Soussi, A.H., Elshafey, M.M., Othman, S.Y., Abd-Elkader, F.A.: Augmented alternative communication methods in intubated COPD patients: does it make difference. Egypt. J. Chest Dis. Tuberc. **64**(1), 21–28 (2015)

16. Eigsti, I.M., Benetto, L., Dadlani, M.B.: Beyond pragmatics: morphosyntactic development in autism. J. Autism Dev. Disord. **37**, 1007–1023 (2007)
17. Sigman, M., Capps, L.: Niños y niñas autistas: una perspectiva evolutiva. Ediciones Morata, Madrid (1996)
18. Manookin, M.B.: A formal semantic analysis of autistic language: the quantification hypothesis. Department of Linguistics, Brigham Young University (2004)
19. Mota, C., Bravo, P.: Autismo e Comunicação. Aveiro (2007)
20. Seung, H.K.: Linguistic characteristics of individuals with high functioning autism and Asperger syndrome. Clin. Linguist. Phon. **21**(4), 247–259 (2007)
21. Wing, L.: The Autistic Spectrum – New Updated Edition. Robinson, London (1996)
22. Rodrigues, L.C.C.B., Tamanaha, A.C., Perissinoto, J.: Atribuição de estados mentais no discurso de crianças do espectro autístico. Revista da Sociedade Brasileira de Fonoaudiologia **16**(1), 25–29 (2011)
23. Graebin, C.: Critérios pedagógicos, ambiente educacional, programa curricular e os aspectos didáticos: critérios relevantes na avaliação de software educacionais. Novas Tecnologias na Educação **7**(1) (2009)
24. Gomes, C.G.S.: Ensino de leitura para pessoas com autismo. Appris Editora e Livraria Eireli-ME (2015)
25. Fonseca, M.E.G., Ciola, J.D.C.B.: Vejo e Aprendo: Fundamentos do Programa TEACCH: O ensino estruturado para pessoas com autismo. Book Toy, Ribeirão Preto (2014)
26. Lauriti, N.C., Molinari, S.G.S.: Perspectivas da alfabetização. Paco Editorial (2014)
27. Veras, C.M.P., Ibiapina, S.R.: Ambiente aquático como cenário terapêutico ocupacional para o desenvolvimento do esquema corporal em síndrome de down. Revista Brasileira em Promoção da Saúde **23**(4) (2010)
28. Pinto, L.M., Pereira, R.A.B., Fabri, A.F.: Desempenho ocupacional em atividades de vida diária de pessoas com desnutrição crônica internadas em enfermarias de clínica médica. Cadernos de Terapia Ocupacional da UFSCar **21**(2) (2013)
29. De Vasconcelos, T.B., Cavalcante, L.I.C.: Avaliação das atividades de vida diária em crianças: uma revisão da literatura. Revista de Terapia Ocupacional da Universidade de São Paulo **24**(3) (2013)

Querying an Ontology Using Natural Language

Ana Marisa Salgueiro, Catarina Bilé Alves, and João Balsa(✉)

Faculdade de Ciências, BioISI – Biosystems and Integrative Sciences Institute,
Universidade de Lisboa, 1749-016 Lisboa, Portugal
a.marisa@campus.ul.pt, fc43103@alunos.fc.ul.pt,
jbalsa@ciencias.ulisboa.pt

Abstract. In the context of the development of a virtual tutor to support distance learning courses, this paper presents an approach to solve the problem of automatically answering questions posed by students in a natural language (Portuguese). Our approach is based on three main pillars: an ontology, a conversion process, and a querying process. First, the ontology, was built to model the knowledge regarding all aspects of the course; second, we defined a way of converting a natural language question to a *SPARQL* query; finally, the *SPARQL* query is executed and the result extracted from the ontology. Focusing on the second pillar mentioned above (the conversion of a NL question), we explain the whole process and present the results regarding a set of preliminary experiments.

Keywords: Question answering · Ontology querying
Dependency parsing · Unity

1 Introduction

The work we present is a component of a bigger project that corresponds to the development of a virtual tutor to support distance learning courses. One of the features of this virtual tutor is the ability to answer questions, that students pose in Portuguese, and that are related to the organization and content of a particular course. In order to enable this ability, we conceived an approach that is based in three main pillars: an ontology, a conversion process, and a querying process. First, we modeled all the necessary knowledge using an ontology. Second, we defined a process that, given a question in Portuguese, converts it to a query in a suitable language (*SPARQL*). Finally, the query is executed and the result (or results) presented to the student. In this paper we will explain the methodology we designed in order to perform the conversion process that corresponds to the second pillar mentioned above, which, as far as we know, has not yet been done for the Portuguese language.

We will start, in Sect. 2, by referring some related work; then, in Sect. 3, we will briefly describe the knowledge modeling task. In Sect. 4, we present the basis

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 164–169, 2018.
https://doi.org/10.1007/978-3-319-99722-3_17

of the conversion process, followed by, in Sect. 5, the corresponding implementation. In Sect. 6, we present the results obtained so far; and, finally, we conclude with some perspectives for future work.

2 Related Work

The Virtual Tutor's concept that is subjacent to our work relies mainly in an ontology-based knowledge model, an efficient translation mechanism for information retrieval by non-experienced users and an automated querying process. It has been shown that knowledge modeling based on ontologies has many advantages in comparison to relational data models [1]. Ontologies are semantically richer than database (conceptual) schemas, are always linked to the concept of language and thus more cognitively similar to Natural Language Processing. Besides, an ontology can also be used to specify domain knowledge and to define links between different types of semantic knowledge. Related to our work's domain there is a case study presented by [2] in which the HERO (Higher Education Reference Ontology) ontology's competency is tested using *SPARQL* queries. Translation strategies includes several natural language interfaces as is the ONLI (Ontology-based Natural Language Interface), which proposes the use of an ontology model in order to represent both the syntactic questions structure and the questions context [3], or LODQA (Linked Open Data Question Answering) system mainly characterized by the use of query patterns leading the interpretation of the user NL query and its translation into a formal graph query [4]. Nevertheless, little has been accomplished using Portuguese and, to our best knowledge, there isn't no direct use of a Portuguese natural language interface as we propose here.

3 Knowledge Modeling

In these preliminary experiments, the modeled knowledge mainly concerns the organization of the course and not its specific scientific knowledge. At this stage, the Virtual Tutor will be an interactive guide for the student to navigate the Curricular Unit Program (PUC). The PUC displays the themes, the objectives of the curricular unit, the competences to be developed, the learning resources, as well as the teacher's expectations regarding their participation. Therefore, the virtual tutor knowledge base is a logic-based ontological model providing a formal semantics and reasoning services.

The ontology was built in Protégé [5], a widely used editor in the scientific community for the development of OWL (Web Ontology Language) models [6]. Within a semantic model implementation, all information is identified using "triples" of the form "subject-predicate-object" (s-p-o). The classification of entities (subjects and objects) and the establishment of relationships (predicate) between those entities in the ontology model was made upon interpretation of the course's PUC. As an example, it was possible to identify entities such as *Student, Teacher, Course, Thematic Subject, Bibliography, Evaluation* and establish triples such as *Course1 hasSubject Subject2.*

One form of querying an ontology about its structure an content is using the *SPARQL* query language [7]. Due to space limitation, we won't go into details here, but some examples will be presented in the next section.

4 Natural Language Querying and Conversion

Although the virtual tutor allows the input of free text questions, only some types of questions will produce a relevant output. There are two main challenges here: the first is to find, in the question, the terms that should be used in the *SPARQL* query and that somehow match the entities defined in the ontology; the second, is to define the structure of the query itself.

For the first, the main resource is the use of a dependency parser [8], the results of which are used to identify the terms that are relevant. Check Fig. 1 for an illustrative example.

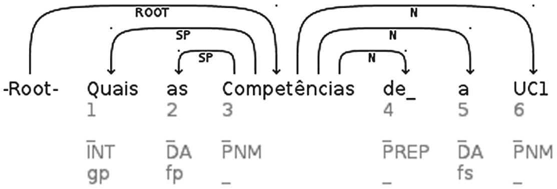


Fig. 1. Dependency parsing for the sentence "Quais as competências da UC1?"

Regarding the second issue, for now all queries have the same structure, that corresponds directly to the (s-p-o) triple pattern.

Concerning the typical question categories, we cover for now some instances of *Definition* questions (*O que é um e-fólio?*); *Yes/No* questions (*Há exame na Unidade Curricular 1?*); *Factual* questions (*Qual a classificação do e-fólio?*); *List* questions (*Quais as competências da Unidade Curricular 1?*). Complex questions, as is the "Why" category, are beyond our scope. Of notice, in Portuguese grammar factual questions can often be posed as a quantifier "Quanto" or a time period "Quando" question.

5 Implementation

Quick communication between the avatar and the student is a requirement in nowadays tutoring applications. One of the application goals is to guarantee fluent conversation and that the correct information is shown to the user. For that, we created a simple application that receives a simple question in natural language and tries to present a conclusive and correct reply (see Fig. 2).

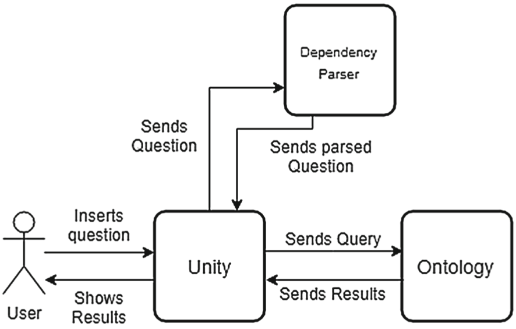
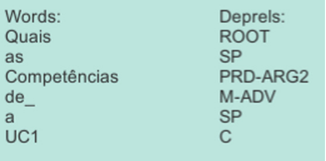


Fig. 2. Diagram of the system

5.1 Analysis of the Question Made by the User

The application was developed in Unity[1], a tool for developing interactive applications, that is used to implement the graphical animated component of the virtual tutor. The translation of the question received by the application to a SPARQL query requires the analysis of every word in the sentence. For example, the sentence "Quais as Competências da UC1?" is decomposed by word, and every word will have an denomination that will help in the construction of the query to ask to the ontology, which is called *deprel* (dependency relation). The *deprel* denomination will help us build the correct query so we can get the correct return value from the ontology. The correct identification of the *deprel* of every word in the sentence is provided by a dependency parser consulted through a webservice made available by NLX Group at FCUL (Faculdade de Ciências da Universidade de Lisboa).



Words:	Deprels:
Quais	ROOT
as	SP
Competências	PRD-ARG2
de_	M-ADV
a	SP
UC1	C

Fig. 3. Example of the analysis of the sentence

5.2 The Creation of a *SPARQL* Query

The results obtained by the division and denomination of the sentence, words and deprel, are an asset to the construction on the *SPARQL* query. This will be analyzed and can give us an idea of which word is the adequate one to be in the query. Following the previous example, we have the all the words and their *deprels* (Fig. 3), and the query will be constructed following the pattern:

[1] https://unity3d.com/.

SELECT ?x ?y ?z WHERE { ?x a tv:PRD-ARG2. ?y rdfs:domain tv:PRD-ARG2. ?x ?y tv:C. OPTIONAL { ?x rdfs:isDefinedBy ?z}}

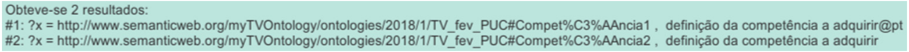
Notice that the terms ‘PRD-ARG2’ and ‘C’ will be substituted by their equivalent word in the final query sent to the ontology.

SELECT ?x ?y ?z WHERE { ?x a tv:Competências. ?y rdfs:domain tv:Competências. ?x ?y tv:UC1. OPTIONAL { ?x rdfs:isDefinedBy ?z}}

Note that this was just an example pattern. The query created might have variations in order to meet the requirements imposed by the user on the question made.

5.3 Getting the Results from the Ontology

As mentioned earlier, Unity is the main program used for implementation. Unity uses C# as base language but other services are used by the application. For instance, the ontology is connected by the *dotNetRDF* plugin that allows the C# code to access the ontology, make a question and get the result. For now, the result is always given in triplets. After that it is shown and counted the number of results that are obtained. Sometimes the result might be a list. For example, the result received for the sentence given earlier is shown in Fig. 4.



```
Obteve-se 2 resultados:
#1: ?x = http://www.semanticweb.org/myTVOntology/ontologies/2018/1/TV_fev_PUC#Compet%C3%AAncia1 ,  definição da competência a adquirir@pt
#2: ?x = http://www.semanticweb.org/myTVOntology/ontologies/2018/1/TV_fev_PUC#Compet%C3%AAncia2 ,  definição da competência a adquirir
```

Fig. 4. Results obtained by querying the ontology

6 Results and Discussion

Regarding the three main pillars it is possible to conclude that the ontology constructed is a very functional semantic representation of the PUC, allowing us to retrieve the information that students need and answer the most frequent questions. The NL analysis showed that it is possible to consistently extract the linguistic triples that match the ontology triples (Subject-Predicate-Object) and convert into *SPARQL*. We also found different query patterns for different question categories. The most reliable patterns were the “Qual/Quais?” and the definition category “O que é?”. Although all the patterns were constructed based on a grammatically correct NL query, the dependency parser is permissive to a freer text, similar to the speaker formalism rather than writing formalism (for example, “Gostaria de saber a nota do trabalho”), not compromising the correct lexemes annotation. The main achievement was the automatization of the NL-SPARQL conversion.

7 Future Work

Although the essence of the methodology we propose is defined and implemented, there are still some issues to deal with, namely: covering additional question types of questions; improving the entity identification process (possibly through the use of information from a named entity recognition process); allowing synonym identification; and, a more thorough evaluation of the system performance. Besides, it will also be necessary to improve the conversion algorithm in order to deal with a wider range of situations (structural diversity of the dependency parser results).

Acknowledgements. This work is supported by UID/MULTI/04046/2013 centre grant from FCT, Portugal (to BioISI – Biosystems & Integrative Sciences Institute) and by PTDC/IVC-PEC/3963/2014 project grant (VIRTUAL TUTORING the virtual tutor as learning mediating artifact in online university education) also from FCT. We would also like to thank our colleagues from the NLX (http://nlx.di.fc.ul.pt/) research group, for their support on using the dependency parser service.

References

1. Munir, K., Anjum, M.S.: The use of ontologies for effective knowledge modelling and information retrieval. Appl. Comput. Inform. **14**(2), 116–126 (2018). https://doi.org/10.1016/j.aci.2017.07.003. ISSN 2210-8327
2. Zemmouchi-Ghomari, L., Ghomari, A.R.: Translating natural language competency questions into SPARQL queries: a case study. In: The First International Conference on Building and Exploring Web Based Environments, pp. 81–86 (2013)
3. Paredes-Valverde, M.A., Rodríguez-García, M.Á., Ruiz-Martínez, A., Valencia-García, R., Alor-Hernández, G.: ONLI: an ontology-based system for querying DBpedia using natural language paradigm. Expert Syst. Appl. **42**(12), 5163–5176 (2015)
4. Shaik, S., Kanakam, P., Hussain, S.M., Suryanarayana, D.: Transforming natural language query to SPARQL for semantic information retrieval. Int. J. Eng. Trends Technol. **41**, 347–350 (2016)
5. Munsen, M.A.: The Protégé team: the protégé project: a look back and a look forward. AI Matters **1**(4), 4–12 (2015)
6. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: OWL web ontology language reference. Technical report, W3C. http://www.w3.org/TR/owl-ref/(2004)
7. Harris, S., Seaborne, A.: SPARQL 1.1 query language. W3C recommendation, W3C. http://www.w3.org/TR/2013/REC-sparql11-query-20130321/ (2013)
8. De Carvalho, R., Querido, A., Campos, M., Pereirsa, R., Silva, J., Branco, A.: CINTIL DependencyBank PREMIUM - a corpus of grammatical dependencies for portuguese. In: Proceedings, LREC2016 - 10th Language Resources and Evaluation Conference, pp. 1552–1557 (2016)

Automatically Grading Brazilian Student Essays

Erick Fonseca[1,2](✉), Ivo Medeiros[2], Dayse Kamikawachi[2], and Alessandro Bokan[2]

[1] Institute of Mathematics and Computer Science, University of São Paulo,
Av. Trabalhador São-carlense, 400, São Carlos, SP, Brazil
erickrfonseca@gmail.com

[2] Letrus, Rua dos Pinheiros, 870, São Carlos, SP, Brazil
ivopdm@letrus.com.br, daysesimon@gmail.com, alessandro.bokan@gmail.com

Abstract. Automated Essay Scoring (AES) is the NLP task of evaluating prose text, still scarcely explored in Portuguese. In this work, we present two AES strategies: the first with a deep neural network with two recurrent layers, and the second with a large number of handcrafted features. We apply our methods to evaluate essays from the ENEM exam with respect to five writing competencies. Overall, our feature-based system performs better in the first four, while the neural networks are better in the fifth one, which is also the hardest to grade accurately. In the aggregated score, our best model achieves a Quadratic Weighted Kappa of 0.752 and a Rooted Mean Squared Error of 100.0 when compared to human judgments, with scores ranging from 0 to 1000.

Keywords: Automated Essay Scoring · Recurrent neural networks
NLP in education

1 Introduction

Automated Essay Scoring (AES) is the NLP task of evaluating and scoring prose text [3]. It is cheaper and faster than having humans evaluating essays, besides providing a deterministic approach, while human reviewers might score essays differently. On the other hand, it lacks full text understanding and may be more vulnerable to cheating, as students may perceive and exploit the factors that determine the scores produced by an AES system.

While there is some reasonable critique against them [11], AES systems for English have been successfully used in schools and large scale exams [3,13]. For research, a well established benchmark is the Hewlett ASAP (Automated Student Assessment Prize) dataset[1].

Here, we describe our research on the development of an AES system for argumentative essays written by Brazilian high school students. More specifically,

[1] Available at https://www.kaggle.com/c/asap-aes.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 170–179, 2018.
https://doi.org/10.1007/978-3-319-99722-3_18

we used the evaluation metrics from the ENEM exam[2], which are also popular in schools.

We pursued two directions for AES: deep neural networks, which have been shown to achieve state-of-the-art results in the literature [5]; and feature engineering based systems, which can benefit from domain knowledge, are usually faster to train and provide a more transparent feedback.

2 Related Work

The problem of AES can be modelled in different ways. Essays can be classified into categories such as *good* and *bad* [8]; ranked from the highest to the lowest scoring [2,14]; or have their grade directly estimated by a regressor [4,5].

Before the recent rise in deep learning, feature engineering already had considerable success. Features usually encode orthography, grammar, style, adherence to the prompt and argumentative abilities. They commonly include word, sentence and paragraph count, word length, spelling and grammar mistakes, syntactic constructions, counts of POS tags and of certain keywords [8,14].

Deep learning approaches usually involve convolutional or recurrent neural networks to create a vector representation of the essay, which is given to an output layer to compute its score. Taghipour and Ng [12] experiment taking the average of outputs from a convolutional neural network (CNN) and from long short-term memory (LSTM) networks, both run over the whole essay. Dong and Zhang [4] use a hierarchical setup in which a first CNN layer generates sentence vectors, and then a second one reads them to output an essay vector. Dong et al. [5] improve over the latter by running an LSTM to obtain sentence representations, and using an attention mechanism to aggregate all intermediate states.

There are few AES works for Portuguese. Amorim and Veloso [1] also work with ENEM essays, and develop a feature-based regressor to grade them. They use some features found in other works for English, and try some new ones specific for the expected style in the ENEM exam. However, they evaluate their model in a different corpus from ours, with a different score range, making a direct comparison infeasible.

3 Essay Dataset

For our experiments, we used a dataset of 56,644 essays, divided in training/validation/test splits of 50,980/2,832/2,832 (corresponding to 90%/5%/5%). They were written in an online platform, and graded by human professionals with respect to the five ENEM competencies:

1. Adherence to the formal written norm of Portuguese.

[2] *Exame Nacional de Ensino Médio*, which serves as an entrance exam for most public universities in Brazil.

2. Conform to the argumentative text genre and to the proposed topic (prompt), developing a text using knowledge from different areas.
3. Select, relate, organize and interpret data and arguments in defense of a point of view.
4. Usage of argumentative linguistic structures.
5. Elaborate a proposal of intervention to solve the problem in question.

Some examples of topics used in the essays are: violence against women, democracy and elections, child publicity, migration in Brazil, prejudice and tolerance, among others.

Each competency (for short, referred to as C1–C5) is graded from 0 to 200, with intervals of 40. The total essay score is the sum of the competency scores, thus ranging from 0 to 1000. The scores in our dataset have a slightly rightward skewed normal distribution, as shown in Fig. 1, also found in [1].

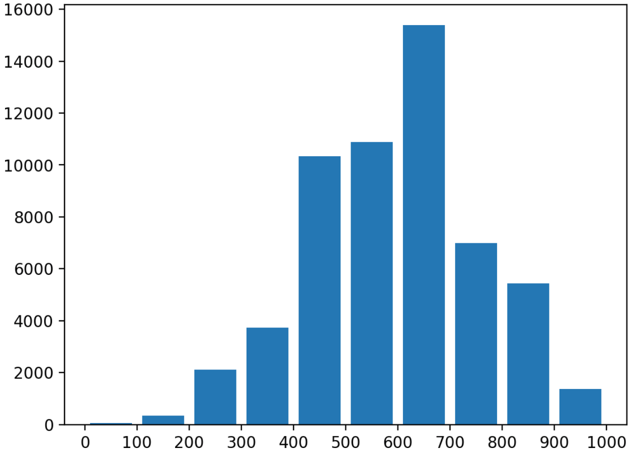


Fig. 1. Distribution of essay scores in our dataset

Table 1 presents some statistics of the dataset. We see that sentences are usually somewhat long, with a median of 28 tokens. Most essays are within 200 and 400 tokens and have four paragraphs.

ENEM determines that essays receive a grade of zero under some circumstances, such as having less than a given number of handwritten lines, not discussing the requested topic or disrespecting human rights. We do not use such essays in our dataset, since they could introduce noise—an essay might have perfect grammar, but have a score of zero on C1 because it does not follow the prompt. Specifically, a score of zero on C2 automatically results in zero for the whole essay, and so we have no essays with zero on C2.

Table 1. Statistics of the dataset. Token counts including punctuation.

Statistic	Median	Mean	Standard deviation
Tokens per sentence	28	32.0	18.2
Tokens per paragraph	72	76.0	34.4
Tokens per essay	323	329.2	101.4
Sentences per paragraph	2	2.4	1.3
Sentences per essay	10	10.3	4.3
Paragraphs per essay	4	4.3	1.0

4 Deep Neural Network

We used a hierarchical neural architecture similar to the one from [5], with two recurrent neural network layers. The first one reads word vectors and generates sentence vectors, which are in turn read by the second one to produce a single essay vector.

Both recurrent layers use bidirectional LSTM (BiLSTM) cells. A BiLSTM is basically two LSTMs, one reading the sequence from left to right and the other reading it from right to left. At each time step (each token in the first layer or each sentence in the second one), the hidden states of both LSTMs are concatenated, and the resulting vector of the layer (sentence or essay vector) is obtained as the mean of all hidden states.

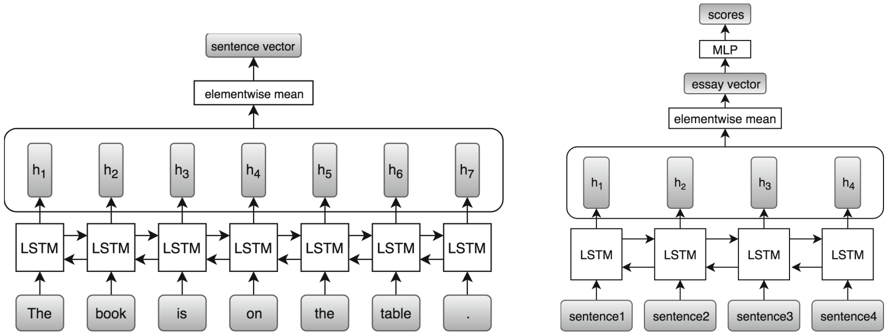
As word vectors, we use word embeddings followed by a projection layer (without bias or activation function). Embeddings were trained with Glove [10] on a corpus of over 560 million tokens, composed of news, literary books and Wikipedia. Vectors for out-of-vocabulary words are sampled from a uniform distribution from -0.05 to 0.05; all vectors are adjusted during training. Dropout is applied to projected word vectors, sentence vectors and the essay vector.

The final essay vector goes through an output layer with five units and sigmoid activation (scores are normalized to the interval $[0, 1]$). The network architecture is illustrated in Fig. 2.

Essentially, we train five regressors in tandem with a common representation for the whole essay. Instead of aggregating LSTM hidden states by mean, we also tried max pooling and concatenating max and mean pooling, both of which yielded worse results.

We also tried a CNN followed by max and mean pooling to aggregate sentence vectors and compute the essay vector, as in [5]. However, we found results substantially worse in comparison with a second BiLSTM. The model training objective is to minimize the sum of the mean squared error in each competency in the training set:

$$L(x) = \sum_{i=1}^{5} (y_i - \hat{y}_i)^2 \tag{1}$$


sentence vector
elementwise mean
h_1 h_2 h_3 h_4 h_5 h_6 h_7
LSTM LSTM LSTM LSTM LSTM LSTM LSTM
The book is on the table .
scores
MLP
essay vector
elementwise mean
h_1 h_2 h_3 h_4
LSTM LSTM LSTM LSTM
sentence1 sentence2 sentence3 sentence4

(a) Structure of the first LSTM layer

(b) Structure of the second LSTM layer and score computation

Fig. 2. Structure of our neural model.

where y_i is the reference score in the i-th competency for the input x, and $\hat{y}_i$ is the score predicted by the model. Thus, at each iteration of the training algorithm, weights are updated according to the gradients of the five competencies.

5 Feature Engineering

While deep learning methods have achieved considerable success in the literature [5], feature based systems are still worth investigating for AES. They provide a more transparent explanation of their decisions and are faster to train, although they may suffer from preprocessing overhead when extracting features for running a trained model. On top of that, AES in Portuguese is still scarcely explored, making it more important to compare different methods for the task. Thus, we trained one regressor for each competency with features extracted from the data.

Our preprocessing is relatively simple: we only tokenize the texts and run the nlpnet POS tagger [6] on them. Ideally, we would like to run a syntactic parser as well, in order to extract richer features. However, we found that a parser trained on a corpus of grammatically well written sentences performs very poorly on essays which often have malformed sentences. We thus leave syntactic analysis as a future work.

Some of our features consider a "valid" vocabulary. This vocabulary is composed of the Unitex DELAF wordlist[3] plus some words found in the essays that we identified as correct but were not present in this resource. Conversely, OOV (out-of-vocabulary) words are those not present in the valid vocabulary.

We extracted features in five major categories: generic count metrics, presence of specific expressions, token n-grams, POS n-grams and POS counts. They are listed as follows.

[3] Available at http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html.

Count Metrics. Most of these features are commonplace in the literature and extract some basic statistics about the text. They are: number of commas, number of characters, number of paragraphs, number of sentences, sentences per paragraph ratio, average sentence length (in tokens), valid vocabulary size (i.e., number of word types in the valid vocabulary), number of word tokens (tokens excluding punctuation), average word length (in characters), number of OOV word types, number of OOV word tokens, OOV type to token ratio, number of non-stopword tokens appearing in the prompt, number of non-stopword tokens appearing in the prompt support texts, mean word corpus-frequency and lowest word corpus-frequency.

The last two consider the frequency with which words appear in a large corpus. We counted words in the same corpus we trained our embeddings to estimate their frequency. Their rationale is that less common words (with lower frequency) indicate a more refined vocabulary.

Specific Expressions. Some groups of words and expressions are expected to appear (or to be avoided) in good essays. They are social agents (such as the government, media, family, law enforcement agencies, schools, etc.), connectives (expressions that connect the discourse, such as *portanto*, *por conta disso*, *nesse sentido*), propositives (expressions that indicate some proposal, such as *precisa-se*, *para que haja*, *é fundamental*) and oralities (words and expressions used in colloquial speech, such as *coisa*, *meio que*, *você*). The last class is the only one expected to have a negative impact in the score. Our lists for these groups have respectively 99, 105, 104 and 62 entries.

The features defined are: has any social agents, number of social agents, connectives in paragraph begin, connectives in sentence begin, connectives in the middle of a sentence, total number of connectives, number of different connectives, number of oralities, number of propositives.

Token n-Grams. We check for the presence of n-grams highly correlated with essay score. We compile a list of such n-grams searching the training data for the ones that appear in between 5 and 50% of the essays, and compute the Pearson correlation between presence of an n-gram in the text and its scores. We keep n-grams with a correlation equal or greater than 0.1 in any competency score, with $1 \leq n \leq 4$, totaling 250 entries. For a richer analysis, we consider an additional token `<s>` in the sentence start, signaling that some n-grams are expected to appear there.

POS n-Grams. We extract a similar list of POS tag n-grams, with $2 \leq n \leq 4$, and check their presence in essays. We have a total of 347 entries.

POS Counts. We count the occurrences of each POS tag in the text. We use the raw counts and the counts normalized by essay length, totaling 52 features (with a set of 26 tags).

In total, we have a pool of 681 feature values, but not all of them are relevant to each of the ENEM competencies. Since we train five regressors separately, we also perform feature selection for each of them. For each competency score, we only keep features with a Pearson correlation of at least 0.1 in the training data—this value was optimized checking performance in the validation data.

6 Essay Grading Results

Here we report the results of our deep neural networks and our feature-based method. In the first one, each half of the BiLSTM had 75 hidden units; word embeddings had 300 dimensions, projected down to 150. The learning rate was set constant at $5 \cdot 10^{-4}$, using the Adam [7] optimizer, and dropout probability of 50%. We trained for only two epochs, with batches of 8 essays. Training longer did not improve performance, and larger batches yielded slightly worse results.

Training sentences were capped at a maximum of 100 tokens (affecting less than 0.7% of the sentences) and essays capped at 51 sentences (affecting less than 0.03% of the essays). At test time, there is no length limit. We experimented different versions of the network: one using a monodirectional LSTM in the second layer, and another with a hidden layer between the essay vector and the sigmoid layer.

For our feature engineering approach, we trained gradient boosting and linear regression models from sklearn [9]. We tuned the hyperparameters of the former checking performance in the validation split. It has 200 estimators, with $\eta = 0.05$, maximum tree depth of 4, and can use all features available.

We measure performance in root mean squared error (RMSE) and quadratic weighted kappa (QWK). QWK is a common evaluation metric for AES, reported in most works in the literature. It measures the agreement between two raters, typically varying from 0 (only random agreement) to 1 (complete agreement). It considers discrete grades, and thus we convert our regressor outputs to the closest multiple of 40 within the range $[0, 200]$ for each competency to match the gold standard scores. However, QWK views scores as unordered values, treating any error equally regardless of magnitude; i.e., answering either 200 or 80 has the same impact if the correct answer was 40. RMSE, instead, is more suited to evaluate regressors (recall that we use MSE as our neural loss function), and penalizes more heavily scores that deviate too much from the human evaluation.

Tables 2 and 3 present our results on the test set, a baseline system that always outputs the training data average scores[4], and results from [1]. We stress that the latter were obtained in a much smaller corpus and are not directly comparable to ours, but some trends are interesting to examine. In Table 2, we also include the results of statistical significance tests. We tested whether the error distribution of the best results from each approach (feature-based or neural network-based) varied significantly in each competency and in the aggregate score, using a two-tailed T-test.

All our models overcome the baseline by a large margin. Feature engineering with Gradient Boosting achieved the best results in C1–4 and the aggregated score, while the neural networks had better results in C5. This indicates that our features model the problem reasonably well for the first four competencies, but the neural networks are better at capturing subtleties of C5—which happens to be the hardest one to predict according to the RMSE.

[4] The baseline always has a QWK of zero because of the definition of the metric, which expects some variation in the results.

Table 2. Rooted mean squared error on the test data. Lower values are better. † denotes that the error differs significantly from the best one in the other approach with $p < 0.5$, and ‡ with $p < 0.1$.

Model	C1	C2	C3	C4	C5	Total
Gradient boosting	**25.81**‡	**26.02**‡	**27.40**†	**28.34**‡	41.49	**100.00**†
Linear regression	26.10	26.37	27.75	28.42	42.07	101.53
Neural network (NN)	27.75	26.58	27.51	29.26	**38.85**†	100.59
NN, monodirectional LSTM	27.72	26.77	27.82	29.09	40.03	101.85
NN, extra hidden layer	28.14	27.18	27.99	29.38	39.48	102.20
Average baseline	38.26	33.53	34.72	39.47	55.27	160.42

Table 3. Quadratic weighted kappa on the test data. Higher values are better.

Model	C1	C2	C3	C4	C5	Total
Gradient boosting	**0.676**	**0.511**	**0.508**	**0.619**	0.577	**0.752**
Linear regression	0.667	0.499	0.493	0.615	0.564	0.747
Neural network (NN)	0.615	0.503	0.500	0.580	**0.636**	0.750
NN, monodirectional LSTM	0.615	0.487	0.490	0.592	0.623	0.745
NN, extra hidden layer	0.584	0.435	0.467	0.552	0.635	0.738
Average baseline	0.000	0.000	0.000	0.000	0.000	0.000
Amorim & Veloso [1]	0.315	0.268	0.231	0.270	0.139	0.367

Another interesting observation is that models perform better on the aggregated score than in each competency separately. This happens because models often score one competency too high and another one too low. The same is observed in the results from [1].

Interestingly, our neural networks did not have any information about the essay prompt, and still performed only slightly worse than the feature models in C2 (which indicates whether the essay discusses the requested subject and has the expected genre, an argumentative essay). The networks can effectively detect the latter, but not the former. A deeper analysis of performance in C2 is desirable, especially if there are data available with the two requirements (subject and genre) evaluated separately.

Comparing our results with the ones from [1], we see that C1 was consistently easier to evaluate. However, their performance on C5 falls considerably lower than on other competencies, unlike ours. This indicates that we have better features for C5, as well as a strong neural model for it.

We also present the RMSE of our Gradient Boosting model broken down by score range in Table 4. Only the aggregated score performance is shown, and we can see that it correlates well with the frequency shown in Fig. 1. The error is very high in the range 0–100, more than twice that of the range 900–1000, which is the second highest one. However, the test set only has four essays in the range 0–100, which makes this statistic not very representative.

Finally, we advocate in favor of RMSE as a metric for AES instead of the more popular QWK. As Tables 2 and 3 show, the latter may give false impressions about the data, such as C5 being easier to evaluate than C2 and C3.

Table 4. Gradient boosting aggregate RMSE broken down by score range

	0–100	**100–200**	**200–300**	**300–400**	**400–500**
RMSE	423.70	147.38	154.06	115.21	96.36
	500–600	**600–700**	**700–800**	**800–900**	**900–1000**
RMSE	82.10	74.10	90.89	131.18	176.61

7 Conclusions

We have presented two approaches for AES in Portuguese, focusing on ENEM essays. We have shown that both achieve similar levels of performance, with the feature based one having better results in the first four competencies while deep neural networks perform better on the fifth. We also raised the issue of performance metrics that should better measure the errors of AES systems.

As future work, besides refining our approaches, we intend to try hybrid methods and adapt our systems to different text genres. More importantly, we intend to use AES results to point out essay writing deficiencies in students, especially with the feature based approach, which provides a more interpretable feedback. In doing so, we can automatically recommend activities to improve specific student weaknesses, such as usage of connectives or essay structuring.

References

1. De Amorim, E.C.F., Veloso, A.: A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In: Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 94–102 (2017)
2. Chen, H., He, B.: Automated essay scoring by maximizing human-machine agreement. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1741–1752 (2013)
3. Dikli, S.: An overview of automated scoring of essays. J. Technol. Learn. Assess. **5** (2006)
4. Dong, F., Zhang, Y.: Automatic features for essay scoring - an empirical study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1072–1077 (2016)
5. Dong, F., Zhang, Y., Yang, J.: Attention-based recurrent convolutional neural network for automatic essay scoring. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 153–162 (2017)

6. Fonseca, E.R., Rosa, J.L.G., Aluísio, S.M.: Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. J. Braz. Comput. Soc. **21**(2) (2015)
7. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. ArXiv e-prints, December 2014
8. Larkey, L.S.: Automatic essay grading using text categorization techniques. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1998 (1998), pp. 90–95. https://doi.org/10.1145/290941.290965
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
10. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014). http://www.aclweb.org/anthology/D14-1162
11. Perelman, L.: When "the state of the art" is counting words. Assess. Writ. **21**, 104–111 (2014)
12. Taghipour, K., Ng, H.T.: A neural approach to automated essay scoring. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1882–1891 (2016)
13. Williamson, D.M.: A framework for implementing automated scoring. In: Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education (2009)
14. Yannakoudakis, H., Briscoe, T., Medlock, B.: A new dataset and method for automatically grading ESOL texts. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 180–189 (2011)

Analyzing Actions in Play-by-Forum RPG

Artur de Oliveira da Rocha Franco[1], José Wellington Franco da Silva[1],
Vládia Célia Monteiro Pinheiro[2], José Gilvan Rodrigues Maia[1(✉)],
Fernando Antonio de Carvalho Gomes[1], and Miguel Franklin de Castro[1]

[1] Universidade Federal Do Ceará, Fortaleza - Ceará, Brazil
{arturoliveira,gilvanmaia}@virtual.ufc.br, wellington@crateus.ufc.br,
{carvalho,miguel}@ufc.br
[2] Universidade de Fortaleza (UNIFOR), Fortaleza - Ceará, Brazil
vladiacelia@unifor.br

Abstract. Interactive Storytelling (IS) technologies are enabling richer experiences for electronic games. Current computational IS models are based on investigations about how games are planned by game designers and actually played by the audience. Unfortunately, most research efforts are limited to the structured data for obtaining insights about IS models. This paper presents a study aimed at determining which actions modeled by Role-Playing Games (RPG) are more important for actual gameplay and how these actions are related. For doing so, we first extracted textual data from existing adventures found in a play-by-forum RPG portal. Such gameplay data is written in Portuguese and reflect natural gameplay without observer intervention. Our analyses from a natural language processing perspective provide valuable insights for IS models in reducing the gameplay chasm between electronic and tabletop RPG.

Keywords: Role Playing Games · Interactive Storytelling
Natural language processing

1 Introduction

Electronic games emerged as new media closely related to Computer Science. Gaming industry has a strong influence on major research in Computer Science, including *Computer Graphic* (CG) and Audio Processing. Demands for technologies continue to grow, especially to produce and manage audiovisual contents, game mechanics, and storytelling [9]. In particular, the *Massive Multiplayer Online Role Playing Game* (MMORPG) genre demands a huge amount and variety of content to be produced. MMORPG titles are inspired by classical, tabletop *Role Playing Game* (RPG), a kind of game in which players interpret characters living adventures in a fictional world in which actions obey the rules of a "system" while the narrative is driven by a "game master". In their turn, MMORPG titles resort to computer-based, simulated environments maintained by servers which provide players with shared remote access to such content-rich and complex games, which are comprised by: game levels; news; rule systems;

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 180–190, 2018.
https://doi.org/10.1007/978-3-319-99722-3_19

scenery elements, including puzzles and history; roads, cities and ecosystems; non-player character behavior; textures, sound, vegetation and particle effects.

That said, developing MMORPG titles challenged developers' ability to manage such a myriad of resources, especially in terms of *quality*. For example, Project *Rathena*[1] which curates contents for the MMORPG *Ragnarok Online*, had $1,782$ different creatures in 2017 [23]. *World of Warcraft*, in its turn, had more than $1,400$ interest locations, $5,300$ creatures, $7,600$ missions, and more than 2 million words written in English in 2008. This title continues to expand even at the time this paper was written. *Procedural Content Generation* (PCG) techniques are quite handful to tackle the challenge of content generation regarding MMORPGs [9] despite concerns regarding quality and diversity of generated content.

Interactive Storytelling (IS) has gained special attention since narratives play a key role in many modern AAA games [22]. Concerning narrative in games, most solutions include decision making models such as fine state machines, grammar plots, planning and emotion models. These models are based on clear aesthetic concepts both in the narrative context and in the application [11,22]. However, these models must be specified in order to encompass the actions of history, without compromising the aesthetic quality known as agency [17], which expresses the level of freedom given to the player viewer to alter the plot.

Analyzing data extracted from digital games is an increasingly common practice for obtaining insights useful for electronic sports, marketing, and more, importantly, understanding the game itself and its actual gameplay [6,8]. Parberry and Doran, for example, investigated four MMORPG titles in order to conceive a quest generator [3]. These authors grouped quest motivations and the actions available for accomplishing them into 9 and 20 classes, respectively. Using their framework, it is possible to generate game quests based on predefined elements in the game world but restricted to choosing a set of actions are needed and in which order these actions must be performed in order to accomplish the quest. Actions in an electronic game are effect of either button presses, for example, attacking or casting a spell. Other actions are an indirect effect, for example, of moving a character to a given area, so the character itself will collect information automatically. That said, implementing such actions is a challenging task since these must cope with conditions and results. In some titles, these actions are abstract and directly exposed to players in terms of dialogue boxes and similar menus, so users are aware of such actions in the game. This, however, may affect immersion and the quality of other resources in the game. Tabletop RPG, in its turn, does not require this actions to be implemented in software since players just need to interpret them and simply adopt the game system's rules in order to determine each action's outcome.

As also pointed out by [15], the work by Jarberry and Doran was broadly disseminated as a model for electronic RPG. Jeong, Cho, and Kang extend the aesthetic qualities into new categories and distribute them in two dimensions:

[1] https://rathena.org/.

material resources and interpersonal relations [10]. Specific extensions were also made to some game titles [16].

Unfortunately, most analyzes serving as building blocks for narrative in games are made on pre-structured and repetitive data provided by *games designers*, thus limiting the interaction with the world in two layers, corresponding to the technological limitations of the media and to the limiting of mechanics by designers, respectively. Agency is thus a result of these limitations, as well as CG specifies limitations for visual elements especially in representing fluids, fire particles and textures, among others [9]. Therefore, it is important to identify and understand the elements which endow agency without harming the fun in order do design better interactive stories [17]. Interacting with the game by means of Natural Language gives greater freedom to content creators to change the gameplay experience and the players to interact and interpret.

Tabletop Role Playing Game (TRPG) are RPG in which users literally play characters and the story is controlled by its description. It is safe to assume that TRPG do not have clear agency and interactivity constraints, as the game is played by using natural language. However, TRPG do not employ attractive computer-supported media resources. Moreover, adopting natural language also makes TRPG a fairly complex study matter.

The main goal of this investigation is to analyze TRPG gameplay in Portuguese oriented by aesthetic and practical concepts from the iconic *Dungeons & Dragons* (D&D) system, is which is related to the origin of TRPG itself and has an open version released in the 2000's. D&D inspires many game systems played in internet forums, the so-called *Play-by-Forum* (PBF). PBF RPG has a more narrative appeal to TRPG since users communicate asynchronously, usually on a weekly basis. Textual gameplay data are analyzed based on features from D&D, thus extracting dimensions of actions [3] in order to test our hypothesis that PBF RPG can provide key insights on gameplay aesthetics, with the potential to prototype part of the creation process in the sense of building electronic RPG titles more similar to TRPG.

The remainder of this text is organized as follows: first we discuss concepts, applications and aesthetics from IS at Sect. 3; our study case is detailed in Sect. 3; analyses and results are presented in Sect. 4; finally, remarks about this work and future research opportunities are discussed in Sect. 6.

2 Interactive Storytelling

IS are based on or adapted from classical concepts of narrative such as the Neo-Aristotelian theory of narrative, thus retaining the idea of action units [1,17]. The main aesthetic qualities for IS are *agency*, *immersion* and *transformation* [17]. Finally, the plot type and application determine narrative requirements.

Practical models for IS appeared in the mid-1970s with *Tale-Spin* [18]. Designed to recreate fables, this model resorts to *automated planning* method for dealing with a conflict. Lebowitz introduced a complex system of interpersonal relationships and emotions in his IS model for soap operas [14]. Most modern IS systems usually aim to feed a formal model with attributes and *actions* operating over these attributes in order to solve story-related problems [15,22,27] via planning [7]. Recent works typically resort to Hierarchical Task Network (HTN) planning and its variations for handling concurrency issues despite the concerns of solving a *NP-hard* problem [27]. Actions clearly play a key role in IS. Therefore, understanding how actions are used in TRPG seems to be a particularly fruitful approach to develop a new generation of IS systems.

3 Case Study

Metadata from actions and other elements from D&D system were extracted from "*Andargor Home*"[2], a public database available both in *eXtensible Markup Language* (XML) e SQLite. Such information reflects how the game was *designed*. On the other hand, we also analyzed *actual gameplay text data* from a web forum called "*Fórum da Jambô Editora*"[3]. Collecting these data samples manually is a boring, time-consuming, and error-prone task, so we developed a customized web crawler in order to tackle this challenging task.

Our web crawler was implemented using Python. We resort to the *Beautiful Soap*[4] for making HTTP calls to the website, thus parsing the resulting web pages so our crawler can navigate into the forums. Forums pages are organized in a hierarchical manner, i.e., starting pages for each forum contains web links to other forums. For each adventure forum, most of the links are parts of the adventure gameplay. Moreover, there is a link for a forum used for RPG system definition character and setup, plus another forum for general discussion.

The web crawler extracted 55 MB of *Hypertext Markup Language* (HTML) data by visiting 980 pages with 8, 902 posts from 68 authors over 12 finished adventure forums. Most execution time is spent performing HTTP requests. This is an important source of research material since it reflects RPG found in the formal market with large runs and that is related to D&D[5].

3.1 Andargor Database

Andargor gathers data about D&D 3.5, including tables for canonical system data, which defines 6 main attributes per character: **Strength** (STR), **Dexterity** (DEX), **Constitution** (CON), **Intelligence** (INT), **Wisdom** (WIS) and **Charism** (CHA). Data on the combat system includes equipment, magic, maneuvers and skill tests. The latter emulate actions and feats within the game, a portion of which are presented as obligations in certain situations. Most skills are triggered according to the player's actions and choices. Skills can be deployed as a sequence of actions that usually involves an interpretive appeal from players, or at least mastery from the characters played in the game [5,8].

[2] http://www.andargor.com/.

[3] http://forum.jamboeditora.com.br/.

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[5] http://www.rpgnoticias.com.br/o-mercado-brasileiro-de-rpg.

A RPG has a low cost to be produced and it is relatively easy to implement, therefore it can serve as a prototype [21]. Concerning its electronic counterpart and IS, an RPG also brings a series of notes on actions and useful elements, both of which served as a reference for the digital genre [8]. Therefore, we can observe a potential within the skills found in the D&D system and, more importantly, their resemblance to the computational models whose present simpler structured actions. We, therefore, propose to revisit the actions and motivations of Doran and Parberry in order to approximate electronic RPG to TRPG.

3.2 Forums Database

Narratives are analyzed under two perspectives. The first is based on actions, as described in Subsect. 3.1. The other is based on *Story Grammars* (SG) classifying story events in 6 categories [26]: *Setting* (S), where the setting of the story is presented; *Initiating Events* (IE) are the initial events after the explanation; *Internal Response* (IR) are the intrinsic responses from characters, i.e., this is how characters are affected and will generate action plans; *Attempt* (A) is the representation of the act, often a result of internal response planning; *Consequence* (C) is the event that shows the consequence of an A; and, finally, *Reaction* (R) that is caused by the event, is usually used at the end of the even. Only 10 of the 12 forums have adventures since two were terminated prematurely. Two of these 10 adventures were classified per sentence extracted for analysis by Cogroo, as explained below.

3.3 Tools and Techniques

HTML content was extracted and processed using Beautiful Soap which interprets the Document Object Model. Then we resort to *Natural Language ToolKit*(NLTK), a Python library for natural language processing tasks, accessed by Cogroo[6]. The latter is a Java interface for analyzing texts in Portuguese. More specifically, we access Cogroo as a LibreOffice module to identify and classify text based on its morphosyntax [24].

Moreover, we analyzed statistical relationships between characters' skills and attributes using *association rules* from Andargor database. These rules were evaluated under parameters defined as *support(X, Y) = count(X, Y)/N*, $confidence(X \rightarrow Y) = support(X,Y)/support(X)$ and *lift(X) = confidence(X, Y)/sup(X)*, where X and Y are absolute frequencies of two entities. We also employed verification by counting and correlation between data samples using Pearson coefficient [13].

$$r(Y, X) = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}} \tag{1}$$

Data from forums were analyzed in two phases: descriptive analysis and *Natural Language Processing* (NLP) analysis. In the first phase we use counting of

[6] http://cogroo.sourceforge.net.

relevant words from the text samples and compare adventures between themselves. For doing so, each word undergoes *Lemmatizion* and *Stemming* [12], thus finding a canonical form in a dictionary and a recommendation of a radical based on the word's structure. This process results in a common word, even when the word is improperly spelled. We then analyze the word counting using descriptive statistics, and the similarity relationship between adventures by means of the Pearson coefficient. See Eq. 1, where X and Y are two attribute vectors (columns) from the same samples (rows), presented in a correlation matrix, the coefficient lies within the interval $I = [-1, 1]$.

The second phase uses a bag-of-words (BoW), i.e., vector with word counts found in the text, generated from nouns, verbs, and adjectives extracted from the text in order to perform NLP analysis. The other words were discarded due to their little importance for the interpretation of the context, as the stopwords, formed by prepositions and pronouns of possession. We adopted the stopword set from NLTK, which in turn was extracted from the PostgreSQL repository[7]. Sentences were represented using BoW in order to classify and predict their category according to the 6 classes found in SG.

This classification task aims to estimate SG based on word presence and frequency. The following classification procedures were used in our experiments: *Gaussian Naive Bayes* (GNB) and its alternative Multinominal Naive Bayes (MNB)[8], which is more suited to handle text; *Multilayer Perceptron* (MLP); and *Support Vector Machine* (SVM) [2]. Finally, we analyze and interpret SG data according to their corresponding actions described by the D&D system.

4 Analyses

Our experimental analysis adopted the Spyder *Integrated Development Environment* (IDE) version 3.2.4[9], in which we reuse libraries from *Scikit-Learn* [19]. Results are presented throughout the following subsections. First, we develop our analysis of data samples extracted from the Andargor Database, then we explore textual gameplay data from internet forums.

4.1 Andargor Database

Results obtained concern the relationship between the 42 skills found in Andargor database. Skill proficiency is not taken into account, but in the presence of the same skill in relation to the others. Comparisons were made applying the same association rules with values above 0.2, confidence above 0.9, and until four rules. This resulted in 991 associations.

We obtained the following results: Listening (WIS), Diplomacy (CHA), Feel motivation (WIS) implies in hiding (DEX); Diplomacy (CHA), Listen (WIS)

[7] https://anoncvs.postgresql.org/cvsweb.cgi/pgsql/src/backend/snowball/stopwords/.

[8] https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html.

[9] https://pythonhosted.org/spyder/.

Table 1. Correlation matrix for canonical attributes from D&D in 681 samples

	STR	DEX	CON	INT	WIS	CHA	CR
STR	1	−0.3	0.59	0.54	0.56	0.61	0.79
DEX		1	0.12	0.03	0.14	0.12	0.19
CON			1	0.63	0.66	0.66	0.66
INT				1	**0.89**	**0.86**	0.67
WIS					1	**0.87**	0.73
CHA						1	0.74
CR							1

and feel motivation imply each other as well as; knowledge (INT), Intimidate (CHA), Search (INT), Hide (DEX), Feel Motivation (WIS), Escape Art (DEX), Concentration (CON) imply each other. This suggests a link between mental skills, the frequency of WIS and CHA is observed. All creatures were analyzed by the correlation of attributes, see Table 1, which denotes the high ratio of mental attributes CHA, INT, and WIS, plus a weak or moderate correlation of physical attributes. These attributes also have a strong relation to the creatures' fighting power and their danger specified by the Challenge Rate (CR) attribute, with the exception of DEX. In fact, there are powerful creatures with high DEX, but there are also powerful creatures with low DEX as large animals and giants that may present great danger but are slow and clumsy.

4.2 PBF Data

The texts from forums were analyzed by the categories of SG. Two forums were analyzed, forum 1 (F1) "Back Alliance Expedition"and forum 2 (F2) "Touhou RPG", under Cogroo's Shallow Parser treatment, extracting 738 sentences, of which 531 are valid sentences of analysis, the others are markings and expressions whose count does not infer semantic value as many prepositions that often establish a regency relation that involve topology of the elements of the sentence [20]. Finally, the sentence classification was obtained from Bag-of-Words (BoW), which adopts a matrix representation. Each line corresponds to a sentence obtained from the parsing process. In their turn, each column corresponds to a word resulting from the lemmatization and stemming steps performed on top of NLTK and Cogroo, respectively. As discussed in Sect. 3, we adopted four different algorithms to predict the classifications: GNB, MNB, MLP, and SVM. We obtained as a result the set-up in the rankings ranging from 46% to F1 and 40% to F2, which is a considerable advantage under the random process which should be around 16.66%.

In addition to sentence classification, a similarity comparison was made between the words contained in the adventures that express the skills, in the Fig. 1(a), presenting a high correlation between adventures with high or medium values, but never negative. Adventures 7 and 8 do not have enough data for analysis, they appear in the forum as closed prematurely and therefore have no adventure data. The analyzed adventures F1 and F2 have indexes 4 and 9 respectively in the Fig. 1(a).

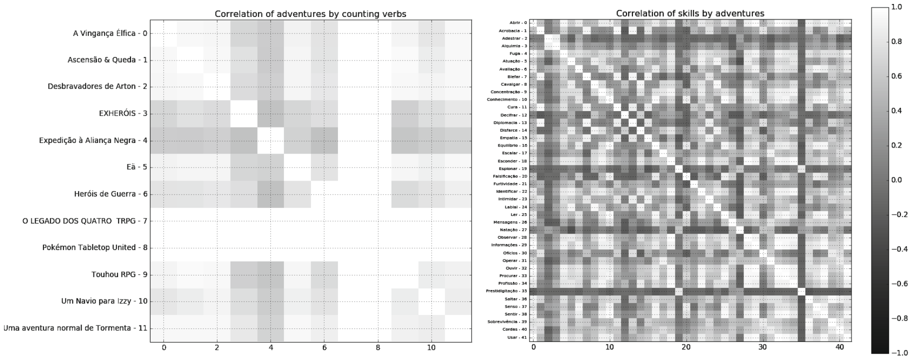


Fig. 1. Correlation matrices for forums, by skills (left) and for skills, by forums (right).

Skills are evaluated in close proximity by their distributions through the forums. This can be seen in the Fig. 1(b). A correlation is observed in a large set of actions, highlighting “Open”, “Escape”, “Messages”, “Watch”, “Search”, “Open”, “Hide”, “Survival” and “Sense” which have high correlation and are present in F1 and in F2. From these we have the Subsect. 4.2 table that summarizes the data crossing the adventures of F1 and F2. Noteworthy the frequency of the skills in the **A** and **C** attributes of the SG, conclude that this is because the skills present core actions within the narrative. In particular, the scenes described that involve actions of escape and hiding place (Table 2).

SG classes were analyzed under support association rules above 30% of the data and with confidence above 90%, thus producing the following rules: **IE** implies **S** and **A**; **A** is an implication of **S**, **IR**, b, **R**. Sentences from a post almost certainly have some correlated action; and Occurrences of **A**, **C** and **IE** imply in **S**. This reflects the logical structure expected from analysis; **A** is the action-related category which is usually central, either as a consequence of character planning after IR and IE, or as a predecessor of **C** and **R**.

5 Related Work

Due to the production demands, data provided for digital game reviews are structured but also limited by technology. Therefore, the analysis by other works are based on well-defined actions. This constraint is somewhat relaxed in RPG: players are free to improvise using natural language as a mean. As already mentioned, Doran and Parberry [3] are an outstanding reference in analysis. Models that procedurally build rules and actions for digital games are complex. Let us

Table 2. Table relating skills by SG in common with adventures F1 (◇) and F2 (●). Where S - Settings, IE - Initiating Events, IR - Internal Response, A - Attempts, C - Consequence, R - Reaction

	S	IE	IR	A	C	R
Abrir (*Open*)	●	●		●, ◇	◇	●
Fugir (*Escape*)	●, ◇	◇		●, ◇	●	●
Mensagem (*Message*)				●, ◇	◇	
Observar (*Observe*)	●	●		●, ◇		
Ouvir (*Listen*)	●	●, ◇	●	●, ◇	●, ◇	
Procurar (*Search*)		◇		●,◇	◇	◇
Sentir (*Sense*)	●, ◇	●, ◇	●	●, ◇	●, ◇	●, ◇
Usar (*Use*)	◇	●		●, ◇	●, ◇	●, ◇
Esconder (*Hide*)	●				●, ◇	◇
Sobrevivência (*Survive*)		●, ◇		◇	◇	
Conhecimento (*Knowledge*)			●			●
Senso de direção (*Directional sense*)	●, ◇					

take the *Ludocore* model [25] as an example. *Ludocore* resorts to logic for building rules of digital games, starting from the creation and construction of rules about states, events and consequences. However, *Ludocore* establishes restrictions on its uses because designers still struggle to specify and understand the consequences of actions [25].

At the time this paper was written, we could not find other works dealing with the analysis of RPG texts with NLP for digital games. Fairchild presents analyses about written of RPG under aspects of reading, handling and circulation [5], providing an overview about writing styles both of players and the material provided by designers. Eliasson's work shows how it is possible to make an application using NLP in digital games with data obtained from analyzes [4]. More specifically, Eliasson proposed a model that creates abstracts for missions following the work by Doran and Parberry.

6 Concluding Remarks

This paper investigates how attributes and actions can be approached in electronic RPG for obtaining experiences similar to TRPG by analyzing PBF RPG texts. This challenge has not been sufficiently explored in terms of analytical approaches. NLP methods combined with multivariate statistical techniques provided a promising set of reusable attributes we believe are useful for a better RPG gameplay design in electronic games. Our analytic method is able to recognize relationships among actions, skills and their distributions. Moreover, our results support the validity of the narrative model proposed by Trabasso [26].

To the extent of our knowledge, this is an unprecedented study on the D&D system under the light of spontaneous textual gameplay descriptions obtained without interference from researchers. Besides, it is worthy to mention that we extracted such a rich, working database which is useful for future investigations. Future works include: (a) extend our analysis to additional adventure sets and to other RPG systems; (b) perform narrative evaluation considering events and facts, thus increasing the complexity of analyses, such as centrality of events; and (c) re-create and evaluate IS models for RPG based on our results.

References

1. Aristóteles: Arte Poética - Aristóteles Texto Integral, vol. 151. Martin Claret, Rua Aegrete, 62, Bairro Sumaré, São Paulo - SP (2004)
2. Coppin, B.: Inteligência artificial. Grupo Gen-LTC (2015)
3. Doran, J., Parberry, I.: A prototype quest generator based on a structural analysis of quests from four MMORPGS. In: Proceedings of the 2nd International Workshop on Procedural Content Generation in Games, p. 1. ACM (2011)
4. Eliasson, C.: Natural language generation for descriptive texts in interactive games (2014)
5. Fairchild, T.M.: Leitura de impressos de RPG no Brasil: o satânico e o secular. Ph.D. thesis, Universidade de São Paulo (2007)
6. Feitosa, V.R., Maia, J.G., Moreira, L.O., Gomes, G.A.: GameVis: game data visualization for the web. In: XIV Simpósio Brasileiro de Jogos e Entretenimento Digital, pp. 70–79 (2015)
7. Fikes, R.E., Nilsson, N.J.: Strips: a new approach to the application of theorem proving to problem solving. Artif. Intell. **2**(3–4), 189–208 (1971)
8. Franco, A.O., Maia, J.G., Neto, J.A., Gomes, F.A.: An interactive storytelling model for non-player characters on electronic RPGS. In: XIV Simpósio Brasileiro de Jogos e Entretenimento Digital, pp. 52–60 (2015)
9. Hendrikx, M., Meijer, S., Van Der Verden, J., Iosup, A.: Procedural content generation for games: a survey. ACM Trans. Multimedia Comput. Commun. Appl. (2011)
10. Jeong, B.G., Cho, S.H., Kang, S.J.: Procedural quest generation by NPC in MMORPG. J. Korea Game Soc. **14**(1), 19–28 (2014)
11. Kybartas, B., Bidarra, R.: A survey on story generation techniques for authoring computational narratives. IEEE Trans. Comput. Intell. AI Games **9**(3), 239–253 (2017)
12. Larson, R.R.: Introduction to information retrieval. J. Am. Soc. Inf. Sci. Technol. **61**(4), 852–853 (2010)
13. Lattin, J.M., Carroll, J.D., Green, P.E.: Analyzing Multivariate Data. Thomson Brooks/Cole, Pacific Grove (2003)
14. Lebowitz, M.: Creating characters in a story-telling universe. Poetics **13**(3), 171–194 (1984)
15. Soares de Lima, E., Feijó, B., Furtado, A.L.: Hierarchical generation of dynamic and nondeterministic quests in games. In: Proceedings of the 11th Conference on Advances in Computer Entertainment Technology, p. 24. ACM (2014)
16. Machado, A., Santos, P., Dias, J.: On the structure of role playing game quests. Revista de Ciências da Computação **12** (2017)

17. Mateas, M., Stern, A.: Interaction and narrative. Game Des. Read. Rules Play Anthol. **1**, 642–669 (2006)
18. Meehan, J.R.: Tale-spin an interactive program that writes stories. In: IJCAI, vol. 77, pp. 91–98 (1977)
19. Pedregosa, F., et al.: SCIKIT-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
20. Pütz, M., Dirven, R.: The construal of space in language and thought, vol. 8. Walter de Gruyter (1996)
21. Ramalho, R.M.A.G.L.: Narrativa e jogos digitais: Lições do RPG de mesa. SBGames (2006)
22. de Oliveira da Rocha Franco, A., Maia, J.G.R., de Carvalho Gomes, F.A.: A programming framework for autonomous NPCS. In: Game Engine Gems 3. CRC Press (2016). Chapter 19
23. Santos, A.M.M., Franco, A.O.R., Maia, J.G.R., Gomes, F.A.C., Castro, M.F.: A methodology proposal for MMORPG content expansion analysis. In: XVI Simpósio Brasileiro de Jogos e Entretenimento Digital (2017)
24. Shaalan, K., Hassanien, A.E., Tolba, F. (eds.): Intelligent Natural Language Processing: Trends and Applications. SCI, vol. 740. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-67056-0
25. Smith, A.M., Nelson, M.J., Mateas, M.: LUDOCORE: a logical game engine for modeling videogames. In: Computational Intelligence and Games (CIG) IEEE Symposium on 2010, pp. 91–98. IEEE (2010)
26. Trabasso, T., Van Den Broek, P.: Causal thinking and the representation of narrative events. J. Mem. lang. **24**(5), 612–630 (1985)
27. Ware, S.G., Young, R.M., Harrison, B., Roberts, D.L.: A computational model of plan-based narrative conflict at the fabula level. IEEE Trans. Comput. Intell. AI Games **3**(6), 271–288 (2014)

Technical Implementation of the Vocabulário Ortográfico Comum da Língua Portuguesa

Maarten Janssen(✉) and José Pedro Ferreira

CELGA-ILTEC, University of Coimbra, Coimbra, Portugal
maartenjanssen@uc.pt

Abstract. The recent Portuguese language orthographic agreement (AOLP90) specifies that the new spelling rules are implemented in an official spelling dictionary (VOC). VOC, released in 2017, is the first common spelling dictionary valid in all Portuguese-speaking countries. AOLP90 allows for some national-level spelling variation, defined in a national spelling dictionary (VON) for each country, containing the nationally-representative words and national-level variants. This combination of a single official spelling with national variation cannot be handled in a traditional set-up for lexical data. This article describes how the lexicon is practically implemented in the VOC database. We start by presenting the nature of AOLP90, the requirements for VOC, and the lexical database. We then analyze the technical implications of orthographic variation in a pluricentric context and present the solutions and practical implementation adopted in VOC. We finish by presenting the pluricentric management system designed for this purpose, devised to cater for decentralized, but compatible management of the lexical database.

Keywords: Spelling dictionary · Computational lexicography
Portuguese as a pluricentric language

1 Introduction

The spelling of the Portuguese language is, like for instance French and Dutch, set in a legally binding document to which all official documents have to adhere. Despite various attempts to have a single legal definition for all Portuguese speaking countries, until recently Portuguese orthography was dealt with in a bicentric setting: there was a legal spelling for Brazil, drawn up in 1943, and there was a legal spelling for Portugal, drawn up in 1945, the latter of which was followed not only in Portugal, but also in all other Portuguese-speaking countries.

This work was developed in the context of the project POCI-01-0145-FEDER-006986, co-financed by Portugal through the Fundação para a Ciência e a Tecnologia and by the European Union's ERDF through the program COMPETE – POCI.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 191–200, 2018.
https://doi.org/10.1007/978-3-319-99722-3_20

In 1990, a common document was finally agreed upon [1], called the *Acordo Ortográfico da Língua Portuguesa* (AOLP90). Following, as usual in orthographic reforms for any language [2,3], a lengthy and convoluted process, the document was legally implemented and put into effect gradually only in 2009.

AOLP90 consists of a set of rules defining how to write the words of Portuguese. An official part of the agreement is that the practical implementation of those rules is detailed in an official spelling dictionary, called the *Vocabulário Ortográfico Comum da Língua Portuguesa* (Common Spelling Dictionary of the Portuguese Language - VOC) [4]. As the name suggests, VOC is a common spelling dictionary valid in all Portuguese-speaking countries. Since, at the same time, AOLP90 allows for some national-level spelling variation, the CPLP countries introduced the notion of national spelling dictionary (VON - *Vocalário Ortográfico Nacional*), containing the nationally-representative words and, in some instances, national-level variants. This combination of a single official spelling with national variation cannot be handled in any traditional set-up for lexical data. This article describes how the lexicon is practically implemented in the VOC database. The next section first gives some general background information about the VOC lexicon.

2 Vocabulário Ortográfico Comum da Língua Portuguesa

Spelling dictionaries are special function dictionaries, determining how to apply the spelling rules of a given language and having a normative function [5]. VOC is the practical implementation of the AOLP90 spelling rules, defining explicitly, based on those rules, what the orthography of concrete words should be for Portuguese. VOC is organized under the guidance of the International Institute for the Portuguese Language (IILP), a body of the *Comunidade de Paises da Língua oficial Portuguesa* (CPLP - Community of Portuguese Speaking Countries), which officially recognized and published it. The technical maintenance of the database itself is carried out by a central team, organized by the CELGA-ILTEC institute of the University of Coimbra.

Contrary to earlier attempts, AOLP90 and VOC take a pluricentric approach, in which all national varieties of the CPLP are assumed to have an equal standing, and allowed to be treated on a par. VOC would in this regard be the intersection of the national VON for each of the eight countries of the CPLP. The elaboration of the VON is hence the responsibility of each country, with a central team having the responsibility of ensuring compatibility and homogeneous criteria [6].

Although all the words in VOC are legally correct in all countries of the CPLP, AOLP90 explicitly allows for some degree of variation in spelling, reflecting in some orthographic contexts the way words are pronounced differently in different countries. Therefore, each VON specifies on the one hand which of the words in VOC are considered a core part of each national vocabulary, and on the other hand, for those words that allow spelling variation, which of the spelling options are recommendable for that country.

Each VON is in principle based on two pillars: on the one hand, the lexicographic memory of the country, that is, the words included in the reference dictionarie(s) of the country, and on the other hand the frequency of each word in a corpus of the national variety. The corpus used for this purpose, the CPLP corpus [7], is itself a pluricentric corpus, and was created for the purpose of VOC due to the lack of sufficiently large existing corpora for the various countries. Since there were no reference spelling dictionaries for any of the CPLP countries other than for Portugal and Brazil, the lexicographic memory was only used directly for those two countries, with the remainder of the VON being mostly corpus-based.

Both the corpora, the lexicographic memory and other, secondary, sources were only used as guidance by the national lexicographic teams - words could be added and removed autonomously by each team. For this purpose, a number of management tools were adopted or purposefully devised.

2.1 Open Source Lexical Information Network

VOC is implemented in a lexical management platform called the Open Source Lexical Information Network (OSLIN) [8]. OSLIN is designed as a relational database, with tables for basic lexicographic notions and relations between them. The core of the database consists of two tables: one containing the lexical entries, and the other the word-forms that are the inflected forms of those lexical entries. Each lexical entry contains a citation form and a word-class, as well as additional information such as the syllabic structure of the word and its inflectional paradigm. Each inflected form contains the ID of the lexical entry it belongs to, its orthography, and a code indicating which inflectional form it is. Together, these two tables define the correct spelling and inflection for every word in the lexicon and provide most of the formal information associated with each entry.

Around these two core tables, additional tables can be added. For VOC, there are three main tables of interest. The first is a table relating lemmas to each other. This table contains the ID of the two related lemmas, plus a code to indicate the type of relation they have. This is used for two purposes. Firstly, it is used to link morphologically related entries, e.g. the adverb *rapidamente* (fast) to the related adjective *rápido* (fast), or the quality noun *beleza* (beauty) to its root adjective *belo* (beautiful) [9]. But more relevantly for the current article, it is also used to relate perceived variants: (1) entries that are perceived to represent forms of the same word (e.g. *impacte* and *impacto*, two valid equivalent forms of *impact* in several countries); (2) words for which the AOLP90 allows variation, such as the translation of *anonymous*, which, due to its pronunciation, is spelled *anônimo* in Brazil, but *anónimo* in other countries; (3) and even some rare instances of purely orthographic variants that were not dealt with in the context of the AOLP90, such as *berinjela* and *beringela*, respectively the Brazilian and European forms for *eggplant*.

The second table of interest for VOC pertains to unadapted loan words, such as the French *abat-jour*, which is often used as such in Portuguese texts, but which does not abide to the phonotactic rules of the Portuguese spelling system. These words are part of the Portuguese vocabulary, but don't follow Portuguese spelling rules. They should be written in italics in Portuguese texts, and often do not have the regular morphology in Portuguese, such as regular inflection, although they can sometimes be inflected according to the rules of the language they come from. For this reason, they are kept separate from the rest of the lexicon in VOC, and provided, where available, with a recognized adapted spelling (*abajur*), and a lexical equivalent in Portuguese (*quebra-luz*).

The third table of interest in VOC is a table of toponyms. It is not uncommon for official spelling resources to contain information on place-names, which often obey a specific lexicographic micro-structure [10]. VOC contains the official spelling for all place-names with administrative relevance for each VON, as well as all names for countries and capital cities in the rest of the world. The spelling of these toponyms is officially fixed in VOC, but the places they denominate are fixed by the various responsible entities in the individual countries, and the official code assigned to them by those organizations is kept in the VOC lexicon, in line with international directives where applicable (e.g. ISO 3166).

3 Dealing with National Variation

Due to its history, the base design of OSLIN deals with national variation in the same way most traditional dictionaries do: by choosing a base variant, and putting a mark on all words that do not belong to that variant marking them as ex-varietal, in the sense of not belonging to that variety or being less favoured. Since the Portuguese database was built in Portugal, prior to VOC the European Portuguese variant was taken as the base.

But such a contrastive design would hardly be pluricentric: there is no mark for words specific to European Portuguese, so it would not be possible to describe another variety fully from the same database, since the words specific to Portugal would not be marked as ex-varietal. Also, the mark is supposed to indicate a single variety, whereas words often belong to several varieties, but not all.

An additional problem is that ex-varietal labels often do not intend to mean what they do at face value: in a traditional contrastive lexicographic work, *bué* would typically be marked as a word from Angola, since that is its perceived or documented origin. But that does not automatically mean it is not correct or not used in Portugal, not that is used only in Angola and not in Mozambique as well, nor even that it is used or correct in Angola at this time (more on this in Sect. 3.2). It would be possible to mark all words not belonging to all varieties with a mark, and allow lists of varieties, as well as negative indications (e.g. "not used in Brazil"). But such an approach would be in our view inelegant for a system that intends to be pluricentric, and also barely computer-readable: it would be necessary to parse the marks before being able to determine whether a word belongs to a given variety under view, and, ultimately, whether it is representative of a given national variety, integrating its VON.

An additional issue is that it is not merely the question of whether a word *belongs* to a given variety or not: although any given word is always legally acceptable in any variety, some of them are not representative of the phonology of that variety, represent very infrequent variants, or are simply not used in a given country. And as such, those words should not be included in its official VON, which intends to be a representation of the lexicon of a given variety. All this complexity cannot be elegantly captured by a label, however complex.

The next section describes the solution adopted in the VOC database in OSLIN in order to solve these tensions within the design of the VOC lexicon.

3.1 VOC Index

The idea behind the implementation of a pluricentric approach to the lexicon in OSLIN developed for VOC is very simple. When considering a variety, there is scale of acceptability: from words that are in the official VON, to forms that are deemed not acceptable in that variety. That is exactly what VOC implements: each entry in the lexicon's database is adorned with a number, ranging from 1 to 5, indicating how acceptable it is in each variety, where 1 indicates it is explicitly listed in the VON, and 5 that it is fully unacceptable in that variety.

So in its core VOC consists of an index table with nine columns: the first indicating the ID of the word, and the other columns the acceptability index for each of the eight countries participating in the project. At the database level, there are no separate entry lists for each country. This reflects the understanding that a given word does not *belong* to a given country or variety, but rather to the language as a whole, and that a VON should as a general rule be perceived as usage and representativeness guides for a given country, and not as a binding, finite list of what is acceptable in its political space, although in certain contexts that may be a VON's intended usage [5].

A VON for a country consists hence of the selection of all the items in the index with an acceptability index below a given threshold in the column corresponding to that country. The acceptability codes are given in Fig. 1.

1 A word explicitly registered in a primary source for the VON
2 A word of unrestricted use in the country, but not explicitly registered in a primary source
3 A word not recommendable in the VON due to usage considerations
4 A word not representative of the county's lexicon due to country-specific AOLP90 variation choices
5 A word fully unacceptable in the VON

Fig. 1. Acceptability codes in the VON index

The ID column consists of two parts: firstly, an indication of the table the entry belongs to, i.e., whether it is an entry in the general lemma list, a toponym, or an unadapted loan word. And secondly, the ID of the word in that table. All usage information about the entry is defined in the country/VON columns.

The values were initially computed from primary sources, both lexicographic and corpus-based, and after that they were made compatible with OSLIN and joined with the VOC index table. Initially, all the entries had a token value *2* for each of the VON. If registered in a primary source deemed representative for a given country, they were presented as candidates to a national team, and, if validated by the team through iterative manual and semi-automatized lexicographic validation processes, their value for that country was stored as *1*.

The value 3 is not intended to capture what is basically meant by the country usage labels in dictionaries: that a word is etymologically from, culturally related to, or most frequent in a specific variety. Etymological and cultural concerns play no role in VOC, and pure frequency differences are dealt with in VOC by an actual frequency indication. Rather, the value 3 is meant for those words that have different spellings, or different variants of the same word in different countries (and as such should be used in the form appropriate for that country), but that variation is not due to AOLP90. A good example is the translation for the word *register*, for which in Brazil they exclusively use the form *registro*, while in other countries the form in use is *registo*. This is the most demanding value to fill in, as automation based on existing lexicographic works is hard to acquire, given the limitations (see Sect. 3) in ex-variety labels used in traditional lexicographic works. It is also the most subjective value, since it provides a way to national teams to manually disqualify words for their national variety for not necessarily objective reasons.

A dedicated table and management system was developed to deal with issues that are specific to AOLP90, pertaining to orthographic contexts where variation is predicted in the spelling rules themselves. Examples are words containing the consonant sequences <*ct*>, <*cp*>, <*cç*>, in pairs such as *sector* vs *setor*, and the aforementioned cases of phonology-induced variation, such as *anônimo* vs *anónimo*. Each of those cases was treated as a variable, and each of its possible values linked to its corresponding entry in VOC and checked against all the available sources for each country, filling in preliminary values. These variables were grouped in word families to ease systematicity. For instance prefixed words should systematically present the same values of acceptability as their morphological base - so if the form *seccionar* is accepted in a given country, but *secionar* is no, *resseccionar* should likewise be acceptable, but *ressecionar* not. There are exceptions to such rules, especially when it comes to derivations and their base: *seção* is more common in Brasil than *secção*, but *seccionar* is much more common than *secionar*. Therefore, index values were not assigned to each related entry in each VON, and for such cases the value *4* in VOC's index had to be assigned on a pondered individual basis by the members of each VON team.

Notice that the value 5 seems to be in contradiction with the official status of VOC: any word acceptable in one variety should be legally acceptable in all varieties, and hence no word should be fully unacceptable in a single variety. The exception pertains to incompatible grammatical terminologies, which originate asystematically country-specific word classes (see Sect. 4.2). The value 5 is also used as a value to exclude entries from VOC entirely: due to the fact that VOC relies on various external sources, those sources sometimes contain entries that should not be considered correct in Portuguese in any variety (see Sect. 3.2). Such words are kept for the sake of more straightforward maintenance.

Due to the set-up of VOC, there are further constrains on the index, that are periodically verified by a set of scripts. An example is the automatic check that there are no words that have only a value *2*, *3*, or *4* in every column. It would in principle not be problematic to have words with only *2* values, since those would be words that are acceptable, but not explicitly listed in any of the VON. However, since VOC is a meta-dictionary combining the VON from the various countries, and hence in order for a word to be part of VOC, it in principle has to be included in at least one of the VON, and hence have a *1* in at least one of the columns. Words that are not acceptable (with no values under *3*) in any of the countries can never occur, and should either be removed from the database entirely, or marked *5* in every column instead, to explicitly mark them as non-acceptable in the Portuguese language.

The VOC index contains an entry for all of the records in all of the three main tables in the database, even for words that are almost by definition part of the VON for every country like high-frequency words such as the verb *ser*. On top of those, there is one class of entries in the index that functions slightly differently: inflected forms. There are several classes of inflected forms that are country-specific, and those are included in the VOC index. However, contrary to lemmas, toponyms, or non-adapted loan words, there is an implicit assumption that any inflected form not listed is always acceptable. And since inflected forms are never included in the VON, nor are they currently subject to frequency concerns - they are only at times against the AOLP90 rules for a variety. As such, entries for inflected forms can only have the value 2 or 4.

3.2 Lexicographic Memory

As mentioned in the previous section, VOC is in part based on the lexicographic memory of each country. But since this lexicographic memory consists of traditional dictionaries using a monocentric perspective, creating a pluricentric database out of them was not straightforward: for words marked as specific to a given country, it had to be manually decided how to interpret that label.

A good example is the word *bué* (ADV) in DLPC [11], which has the tag *Angol.* to indicate it is a word *from* Angola. Since DLPC is a Portuguese dictionary, intending to represent the European variety, it should never be used as a reliable source for the Angolan variety. But whether the tags should be interpreted as indicating it is not correct or used in Portugal is not clear from the tag, and at least in this example it is not the case: it is a word frequently used in Portugal, as can be seen from its high frequency in Portuguese corpora and the fact that some Portuguese dictionaries do not have any tag for the word, while other dictionaries indicate it as specific to both Angola and Portugal.

Another issue with the lexicographic memory is the fact that the spelling in the source dictionaries does not always match the spelling in VOC, and such words need to be mapped to their current spelling. Many such cases involve word that have been changed with AOLP90, e.g. *contrafacção* (counterfeiting), which is no longer a correct spelling in any country of the CPLP. But it also involves errors or proposed spellings, such as the DLPC-proposed spelling *croissã* for *croissant*, which contradicts VOC's orthographic criteria. This mapping is for a good part manual work. Since those words are no longer correct, they should not be part of VOC. However, since removing a word completely also removes the often intricate process of deciding whether they are still correct or not, and makes mapping entries against corpora and other sources harder, such words are kept in the database, using the value *5* to indicate they shouldn't be part of VOC.

4 Practical Implementations

4.1 Online Interface

The VOC lexicon is officially hosted at the site of the CPLP[1], but also included in other sites, most prominently the Portal da Língua Portuguesa[2]. The default way in which those interfaces behave is to guess the preference of the user (based on locale and country of the request IP) or ask for it explicitly, and display the spelling recommendation specific to the selected VON. However, it is also possible to have access to the entire VOC lexicon.

Since there are no explicit databases for each VON, the display of a given VON is handled using the VOC index: in the display of a VON, only those words that have a 1 or a 2 in the column for that VON are displayed, since anything above that should be perceived as not recommended in that VON. Furthermore, in the alphabetic listing, those words that are explicitly part of a given VON (i.e., that have a value *1*), are presented with a filled in square in front of them, similar to the treatment in other online dictionaries such as the *Diccionari Enciclopèdic* for Catalan [12], that uses the same method to indicate which words are included in the official Institut d'Estudis Catalans dictionary [13].

4.2 Complicated Cases

The system adopted in VOC to explicitly list the status of each word in each variety with a number is very flexible and expressive, leading to a system that is easy to computationally exploit, but also easy to maintain. However, that does not mean there are no cases that are complicated to deal with in this design.

A first complication is the case of partial homographs: words that allow more than one spelling, whether due to AOLP90 or not, but for which one or more of those are homographous with another word that does not allow spelling variation. A good example is the word *fato*: on the one hand, it is the Brazilian corresponding variant spelling for *facto* (fact), used in other countries. But it is also the European Portuguese word for *suit*, which can never be written as *facto*.

Another difficult case has to do with the word classes for words, the name of which is determined by the official terminology, which is independent of VOC, and not harmonized throughout the CPLP. There are three different official

[1] http://voc.cplp.org.

[2] http://www.portaldalinguaportuguesa.org.

terminologies: the one from Brazil (*Nomenclatura Gramatical Brasileira, NGB*), the one from Portugal (*Dicionário Terminológico, TL*), and the terminology formerly used in Portugal (Nomenclatura Gramatical Portuguesa, NGP), which is still in effect in all countries of the CPLP except for Portugal and Brazil. Common nouns are called *substantivo* according to NGB and NGP, but have been changed to *nome* in the more recent TL. But more problematic are the close-classed words for which the difference is not only in naming, but also in the scope of each of the classes, which renders them incompatible across countries. For such rare cases an independent entry was created for each terminology, and a value *5* assigned to it in VON not following that specific official terminology.

5 Conclusion

In this paper, we have shown how the use of a simple index that keeps track of the acceptability status of each word in each national variety can solve the problem of having a single database in which it is possible to store and organize national vocabularies for each of the countries CPLP countries, both the official vocabulary and the list of acceptable words in that variety, while correctly dealing with the fact that there are words in VOC that preferentially have a different orthography in the different countries.

The same strategy can also be extended to full dictionaries, by providing the same kind of acceptability and frequency information for each meaning of each word, and to lexicographic works for other pluricentric languages. Like in the case of word-forms, not the entire range of acceptability measures would apply to word senses, with in principle only 2 and 3 being possible values, since meanings are never included in a VON (although it would be possible to think of lists of core meanings for a variant), and they never go against the AO.

References

1. Marquilhas, R.: The portuguese language spelling accord. Writ. Lang. Lit. **18**(2), 275–286 (2015)
2. Johnson, S.: Spelling Trouble? Language, Ideology and the Reform of German Orthography. Information and Interdisciplinary Subjects Series. Multilingual Matters Ltd, Bristol (2005)
3. Coulmas, F.: Writing reform: conditions and implications. In: Writing Systems: An Introduction to Their Linguistic Analysis, pp. 241–263. Cambridge University Press (2003)
4. Ferreira, J.P., Correia, M., de Almeida, G.B., eds.: Vocabulário Ortográfico Comum da Língua Portuguesa. In: Instituto Internacional da Língua Portuguesa/Comunidade dos Países de Língua Portuguesa, Praia, Cape Verde/Lisbon, Portugal (2017)
5. Buchmann, F.: Spelling dictionaries. In: Durkin, P. (ed.) The Oxford Handbook of Lexicography. Oxford University Press, Oxford (2016)

6. Ferreira, J.P., Janssen, M., de Almeida, G.B., Correia, M., de Oliveira, F.M.: The common orthographic vocabulary of the portuguese language: a set of open lexical resources for a pluricentric language. In: Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, pp. 1071–1075 (2012)
7. Janssen, M., Kuhn, T.Z., Ferreira, J.P., Correia, M.: The CPLP corpus, a corpus of portuguese as a pluricentric language. In: XVIII EURALEX International Congress, Ljubljana, Slovenia (2018)
8. Janssen, M.: Open source lexical information network. In: 3rd International Workshop on Generative Approaches to the Lexicon, Geneva, Switzerland (2005)
9. Janssen, M.: Affix selection and deadjectival nouns: a data-driven approach. In: Humphries, G. (ed.) Enlish Language, Literature, and Culture: New directions in research. Bielo-Bialska, Poland (2008)
10. Styles, T.: Place-name dictionaries. In: Durkin, P. (ed.) The Oxford Handbook of Lexicography. Oxford University Press, Oxford (2016)
11. Casteleiro, J.M. (ed.): Dicionário da língua portuguesa contemporânea: 1. A-F. Academia das Ciências de Lisboa, Verbo (2001)
12. Enciclopèdia.cat: Diccionari enciclopèdic (2018)
13. Institut d'Estudis Catalans: Diccionari de la llengua catalana. Institut d'estudis catalans (2007)

Identifying Intensification Processes in Brazilian Sign Language in the Framework of Brazilian Portuguese Machine Translation

Francisco Aulísio dos S. Paiva[1], José Mario De Martino[1](✉), Plínio A. Barbosa[2], Pablo P. F. Faria[2], Ivani R. Silva[3], and Luciana A. Rosa[3]

[1] School of Electrical and Computer Engineering, University of Campinas, Campinas, Brazil
martino@fee.unicamp.br

[2] Institute of Language Studies, University of Campinas, Campinas, Brazil

[3] School of Medical Sciences, University of Campinas, Campinas, Brazil

Abstract. Brazilian Portuguese (BP) to Brazilian Sign Language (Libras) machine translation differs from traditional automatic translation between oral languages specially because Libras is a visuospatial language. In our approach the final step of the translation process is a 3D avatar signing the translated content. However, to obtain an understandable signing is necessary that the translation takes into account all linguistic levels of Libras. Currently, BP-Libras translation approaches neglect important aspects of intensification of words and construction of plural nouns, generating inappropriate translations. This paper presents a study of the intensification of adjectives, verbs and nouns' pluralization, with the aim of contributing to the advance of automatic sign language translation. We apply a hierarchical clustering method for the classification of Libras' intensified signs considering the modification of manual parameters. The method allows for the identification of distinct intensification patterns and plural marking in BP sentences. The results of our study can be used to improve the intelligibility of the signing avatar.

Keywords: Brazilian sign language · Machine translation
Signing avatar

1 Introduction

In Brazil, there are two official languages: Brazilian Portuguese (BP) and Brazilian Sign Language (Libras). A BP-Libras Machine Translation (MT) system can promote the inclusion of deaf people in the society. Daily tasks, such as reading books or accessing news on the Internet, could be facilitated for deaf people by such a translation system. However, there is a great challenge to a reasonably comprehensive translation. Considering that Libras is a natural visuospatial language, we included signing by a 3D avatar as the final step of our BP-Libras MT system.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 201–210, 2018.
https://doi.org/10.1007/978-3-319-99722-3_21

In this context, we are developing an intermediate language to control the avatar. In our intermediate language, each sign is represented by a gloss, a BP word written in capital letters, to distinguish Libras' signs from BP words. For example, the BP sentence "João gosta de Maria" (João likes Maria) is transcribed as "JOÃO GOSTAR MARIA". Ferreira-Brito [7] and Felipe [5] pointed out the main characteristics of a gloss-based representation for Libras. For instance, verbs are always represented in the infinitive; two words are separated by a hyphen when they are required to represent a single sign (e.g.: COMER-MAÇÃ (eat apple); the @ is used to sign the non-representation of gender (e.g.: AMIG@ = "amiga" and "amigo" (friend)).

The key idea of our system is that given a BP text, a gloss transcription is generated by a set of morphosyntactic and semantic translation rules. Figure 1 shows the link between the glosses and the avatar we are developing.

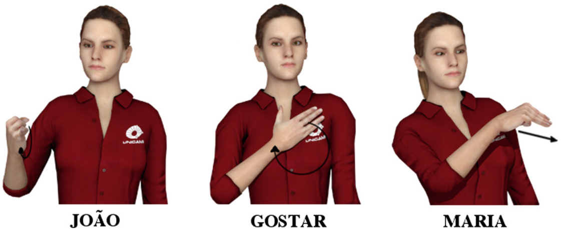


Fig. 1. Glosses linked to Libras signs

We identify in the literature a lack of works dealing with the intensification processes in the context of MT for Sign Languages. According to [3], intensification is the linguistic expression we use to scale a quality, either up or down. Adverbs of intensity modify adjectives, verbs or others adverbs. Some examples are "muito" (very), "pouco" (a bit). In Sign Languages, intensification modifies the realization of the signs, because it involves the modification of manual and non-manual expressions [13].

This paper presents a study on the intensification of adjectives, verbs and on the pluralization of nouns, with the aim of contributing to the advance of automatic sign language translation and to the study of Libras' prosody. The manuscript is organized as follows. In Sect. 2, we present works related to the description of sign language intensification and BP-Libras MT systems. Section 3 describes our corpora and methodology. Section 4 presents results and the implication for inference of morphosyntactic and semantic rules. In Sect. 5, we present the conclusions and perspectives for future works.

2 Related Work

According to Felipe [6], the intensification in Libras happens by means of the repetition of the corresponding sign or other type of change in movement, such as, through slower (or tense) movement or through an amplification of the gesture. Quadros and Karnopp [12] also indicate that non-manual expressions (facial expressions, head and torso movement) act for phonological building adverbs of intensity. To transcribe the phrase "muito bonito" (very beautiful) into glosses, Felipe [6] transcribes it as "BONIT@muito" and Quadros and Karnopp [12] as "BONITO+". In this paper, we gloss it as "BONITO-MUITO" to indicate that two words are represented by a single sign in Libras. In the following, we present some studies involving intensification that will be discussed in the light of works on BP-Libras automatic translation.

Wilbur, Malaia and Shay [13] investigated the intensification and degree of modification in adjectives in ASL. The authors cite that the specific sign for "VERY" is considered as "signed English" and rejected as an intensifier. In fact, they observed that when intensifying the signs the following characteristics were used: increased tension of movements, addition or extension of a gesture, delay of the start of the movement and non-manual modifications (face, head, body). In addition to these features, Xavier [14] and Xavier and Barbosa [15] showed that Libras uses hand duplication for intensification and pluralization of nouns. We have adopted the convention that plural marking will be denoted by right adding -PL to the gloss. These modifications in the parameters of Libras are important for natural signing.

There are some BP-Libras MT systems, such as HandTalk (v. 2.3.1.4) [8], ProDeaf (v. 3.6) [11] and VLibras (v. 3.2.0) [1]. In order to verify their functionality with intensification and pluralization, we tested the following phrases: (1) "muito bonito" (very beautiful); (2) "muitos carros" (many cars); (3) "amar muito" (to love so much).

ProDeaf and VLibras perform two signs for these phrases, that is, "MUITO" and "BONITO", "AMAR" and "MUITO", "MUITO" and "CARRO". HandTalk does the same for the verb "amar", but ignores "muito" for the adjective while replacing it with "VÁRI@" (several) for the noun "carros". This indicates that ProDeaf, VLibras and HandTalk basically use signed Portuguese strategies, since they did not use the modified signing involved in the Libras intensification process.

In a previous study, we identified how intensification is signed in Libras, that is, we identified the modifications of the manual parameters and the non-manual expressions involved in this process [10]. In that study, we have found that, for non-manual expressions, their frequency of use is higher when compared to the neutral form. Furthermore, we concluded that, to model a realistic avatar, it is necessary to incorporate the following modifications: wider gesturing, use of two hands, inflated cheeks, frowned eyebrows and tightening of eyes.

In the present study, extending the findings of [10], we focus on manual parameters and propose a hierarchical clustering of intensified signs by observing the changes from the neutral form, in order to identify possible distinct classes of intensification. We also formalized morphosyntactic rules for dealing with intensification processes in Libras.

3 Materials and Methods

In this section, we present: (1) BP-Libras parallel corpus; (2) morphosyntactic processing; (3) Libras intensification corpus; (4) Hierarchical cluster analysis.

3.1 BP-Libras Parallel Corpus

The BP-Libras parallel corpus contains the BP-to-Libras translation of a school science textbook developed by De Martino et al. [4]. The translation process involved collaboration between BP speakers proficient in Libras and deaf individuals with reading skills in BP and with Libras as their first language. This work yielded a material consisting of a set of videos with the sentences of the book associated with Libras' glosses.

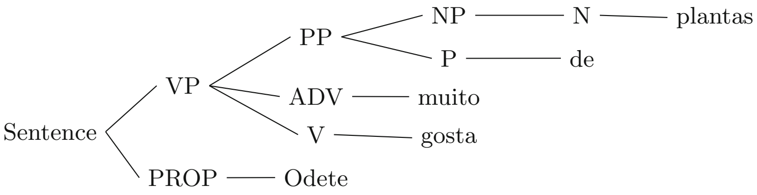
3.2 Morphological and Syntactic Processing

For morphosyntactic analysis of the parallel corpus, we are using the following tools: (1) DELAF-BP dictionary developed by Muniz [9] and NILC (Interinstitutional Center for Computational Linguistics USP-São Carlos-SP). The dictionary has circa 880,000 entries. Each entry consists of a BP word, its lemma, grammatical category and its inflection. (2) A lemmatizer developed by Maziero[1], from the NILC group, which annotates words with lemmas and grammatical categories. (3) PALAVRAS - a syntactic parser for Portuguese developed by Bick [2]. The basic steps of our algorithm are:

(1) **Given the input text:** Odete gosta muito de plantas (Odete likes plants very much).

(2) **Morphological analysis:** Lemmatize and assign part-of-speech tags from DELAF-BP returning two lists with the results, as follows:
Lemmas: [odete, gostar, muito, planta]
Tags: [['N+Pr fs'], ['V Y2s', 'V P3s'], ['ADV+Int'], ['N fp']]
The positions in the lists are element-wise associated with each other. Therefore, we have the following results:
Odete is **Pr**oper **N**oun; **F**eminine; **S**ingular. **Gostar** is **V**erb, **Y** (imperative), **2s** (2 person singular) or **P**resent tense, **3s** (3 person singular). **Muito** is ADV+Int (intensity adverb); **Plantas** is **N**oun; **F**eminine; **P**lural.

(3) **Syntactic analysis:** The sentence is used as input to the parser that returns the following syntactic tree:



[1] http://conteudo.icmc.usp.br/pessoas/taspardo/LematizadorV2a.rar.

With these pieces of information from both the lemmatizer and the parser, we are able to model the behavior of adverbs, quantifiers and plurals in the corpus sentences. From this, we formalize and implement BP-Libras translation rules. Furthermore, these rules can be implemented in other Portuguese lemmatizers and parsers, such as spaCy[2] modules for Portuguese.

3.3 Libras Intensification Corpus

For our study on intensification, we have collected a specific corpus. The corpus of 140 declarative sentences was built by combining 70 key-signs x 2 conditions (neutral and intensified). The sentences were produced by a deaf informant fluent in Libras, and we analyzed 30 adjectives, 30 nouns, 10 verbs combined with the adverb/determinant "muito(s)" (very), with the aim of identifying the changes in the manual parameters. The list of all Libras' signs analyzed is presented below.

Verbs: CAMINHAR, CORRER, CAIR, LEVANTAR, CHORAR, SORRIR, SABER, APRENDER, PERGUNTAR, AMAR.

Adjectives: PREOCUPADO, ORGULHOSO, ESTÚPIDO, CIUMENTO, HUMILHADO, BRAVO, DESCONFIADO, FEIO, CARINHOSO, AGITADO, HONESTO, EDUCADO, MODESTO, CHATEADO, TRISTE, DECEPCIONADO, CORAJOSO, CONCENTRADO, CONSOLADO, ABORRECIDO, CALMO, ANSIOSA, TÍMIDO, SIMPÁTICO, SAUDÁVEL, ALEGRE, APAIXONADO, ADMIRADO, PREGUIÇOSO, APAVORADO.

Nouns: COMIDA, PRODUTO, VIOLÃO, SAPATO, ANIMAL, CAMISA, PESSOA, ARROZ, FEIJÃO, CONHECIMENTO, FRIO, TRANSITO, AÇÚCAR, VACA, CAIXA, POSTE, MOTO, CARRO, CASA, ÁRVORE, CHUVA, CALOR, LEITE, NEVE, ÁGUA, SAL, MAÇÃ, AREIA, DINHEIRO, LÁPIS.

We take into account the following modifications: (1) sign repetition; (2) duplication of hands; (3) amplification of the gesture; (4) movement tension; (5) adding a sign to represent intensification or plural.

3.4 Hierarchical Cluster Analysis

For the analysis of the corpus in Sect. 3.3, we used a clustering method to classify the signs:

(1) For each sign, we mark with 1 the use of one of the five modifications given above or 0 otherwise. Table 1 exemplifies this strategy.
(2) Clusters were produced for the following classes of signs: adjectives; adjectives x verbs; nouns. With the complete Table 1, we calculated the correlation between lines (i.e., between signs). Using the Pearson coefficient with

[2] SpaCy - Available in https://spacy.io/models/pt.

correlations ranging from −1 to 1, where −1 represents perfect negative correlation, 0 represents that there is no correlation, and 1 represents perfect positive correlation.

(3) The calculations were performed using the hierarchical classifier (hclust) of the R Statistics Sofwtare taking as input the distances based on the correlation between the representations of the signs.

Table 1. Example of sign modifications for three key signs

Key sign	Repetition	Duplication	Amplification	Sign add	Tension
AGITADO (agitated)	1	0	0	0	1
ANSIOSO (anxious)	0	0	0	1	1
CARINHOSO (amorous)	1	0	0	0	1

4 Results and Discussion

4.1 Statistical Analysis by Hierarchical Clustering

We first evaluated adjectives. Our first analysis indicates that the intensified signs can be classified into two major groups: signs with and without movement tension. The classification reveals that movement tension is the relevant modification to split the tree in two major groups. The following modification for subgrouping is amplitude of sign, as shown in the dendrogram of Fig. 2.

Figure 2 shows adjectives in different valence polarities (positive and negative). It is worth noting the grouping of signs of negative valence formed by the signs ORGULHOSO, ESTÚPIDO, CIUMENTO, HUMILHADO, BRAVO, DESCONFIADO, and FEIO. These signs are intensified by using both movement tension and amplification of the sign. On the other hand, positive-valenced signs, such as SIMPÁTICO, SAUDÁVEL, ALEGRE, APAIXONADO, were intensified with no tension, but with amplitude of the gesture and duplication of hands. However, in general, a mixture of signs of both valences occurs in each grouping. This indicates that the manual parameters are more determinant for the classification than the valences.

The dendogram shows how to intensify the signs. For example, TÍMIDO is intensified as follows: tense, no amplification of gesture, no repetition, with duplication and no sign added. When analyzing the videos of adjectives and verbs, we observed that they were intensified with the same features. Then, we performed a second analysis by putting together adjectives and verbs.

In the dendrogram of Fig. 3, we notice that the pointed verbs are inserted in the already existing groups. In fact, SORRIR groups with ADMIRADO, previously isolated. The verb SABER was withdrawn from the analysis because it was

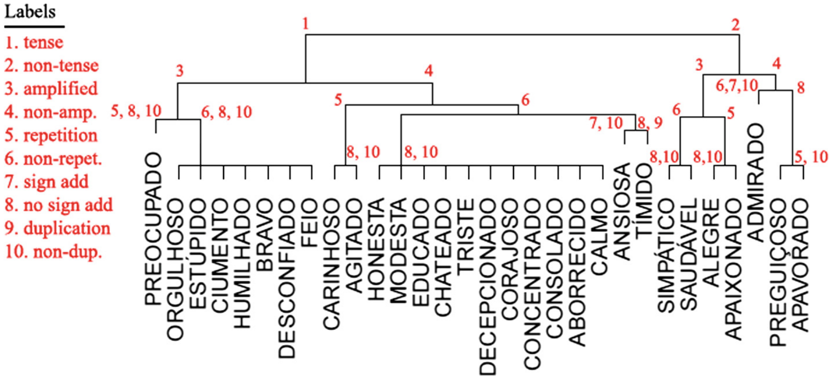


Fig. 2. Dendrogram of the adjectives

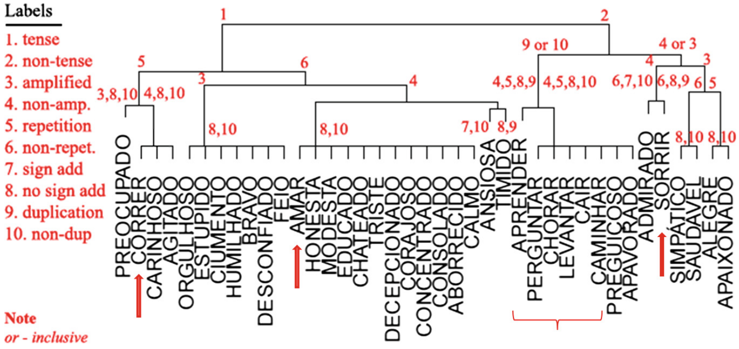


Fig. 3. Dendrogram of the adjectives and verbs

not intensified correctly. The other verbs are correlated with adjectives using the same modification parameters. Thus, we argue that the dendrograms indicate the set of parameters that should be implemented in the avatar to represent intensification of both adjectives and verbs.

The third analysis was perfomed with nouns. By analyzing the videos we observed that the plural is realized with the following characteristics: sign repetition, duplication of hands, amplification of the gesture, addition of the sign "VÁRI@" (several) or addition of a specific sign to represent more elements. The dendrogram of Fig. 4 shows that there are more nouns that use the sign repetition feature or gesture amplitude. And there are few signs that need the help of an auxiliary sign (such as the VÁRI@). We argue that such observations about pluralization should also be accommodated in the plural implementation of BP-Libras MT systems.

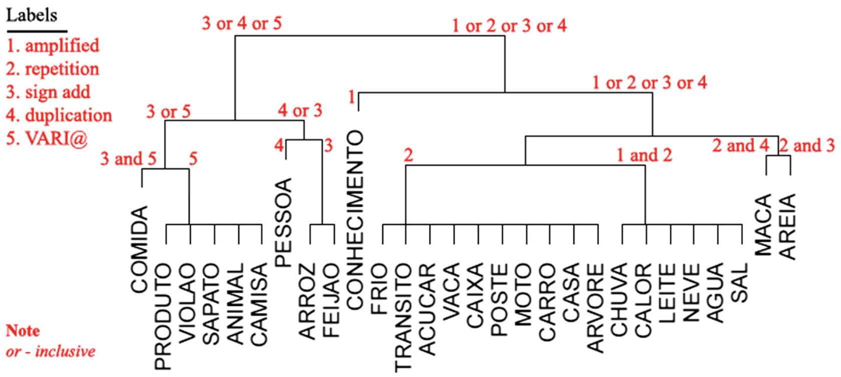


Fig. 4. Dendrogram of the nouns

4.2 Implications for BP-Libras Machine Translation

We formalized and implemented four translation rules on intensification.

(R1) **Adverb and verb**: given a text, the parser identifies the verb and the adverb, so we implement the junction of the two words, indicating a single sign in Libras. Example:
BP text: João gosta muito de Maria. (João likes Maria very much)
Libras' glosses: JOÃO GOSTAR-MUITO MARIA.

(R2) **Adverb and adjective:** the parser identifies the adjectival phrase and exchanges their positions to join them. In BP, the adverb "muito" comes before the word that is being intensified. Example[3]:
BP text: Odete é muito educada. (Odete is very polite.)
Libras' glosses: ODETE EDUCADO-MUITO.

(R3) **Class 1 of nouns:** we call class 1 nouns that are pluralized only with the use of a manual parameter, such as repetition. In this case, we exclude the determinant "muitos" and add the -PL notation indicating the plural. Example:
BP text: Odete gosta de muitas plantas. (Odete likes many plants.)
Libras' glosses: ODETE GOSTAR PLANTA-PL.

(R4) **Class 2 of nouns:** we call class 2 those nouns that need the auxiliary sign "VÁRI@" to represent the plural. Generally, it is signed after the noun. Our rule excludes "muitos" and adds "VÁRI@" sign after the noun. In addition, the noun sign remains in the singular, as there is no change in itself. Example:
BP text: Odete gosta de muitos animais. (Odete likes many animals.)
Libras' glosses: ODETE GOSTAR ANIMAL VÁRI@.

These examples show the importance of using a parser. To form the gloss representing the intensification sign in Libras it is necessary to identify the word

[3] There are other rules involved, such as the exclusion of prepositions (such as "de") and linking verbs (such as "é").

to which the adverb is associated. For example, in the sentence "Maria ficou muito triste com a notícia" (Maria was very sad about the news), the parser will identify that "muito" is associated with "triste" and not with "ficou" through the adjectival phrase (ADJP (ADV muito) (ADJ triste)). Formalized rules are sufficiently general for other sentences with similar syntactic structures. Other examples:

(1) Adjectives
BP text: Maria é inteligente e muito modesta. (Maria is intelligent and very modest.)
The word "muito" is related only to "modesto" then we generate the gloss "MODESTO-MUITO".

(2) Nouns
BP text: Odete tem muitos sapatos e muitos carros. (Odete has many shoes and many cars.)
In this sentence, rules R3 and R4 are applied together. As we saw in Fig. 4, the noun "SAPATO" needs the sign "VÁRI@" to designate the plural, and the noun "CARRO" is pluralized only with the repetition of the sign, producing the gloss "CARRO-PL".

This item indicates that our morphosyntactic rules work well even for sentences containing nouns of different pluralization forms. It is worth mentioning that the rules are in agreement with the corpus' sentences and also with the Libras' grammar. Lastly, each of these glosses are signed by the avatar, and the hierarchical classification indicates important characteristics that it should perform.

5 Conclusions and Future Work

The study presented in this paper supports the formalization of rules for a BP-Libras MT system. We have shown that for ensuring naturalness in Libras signing, signed Portuguese strategies must be avoided, in favor of the full use of the sign modifications associated to intensification processes discussed here. In addition, the findings are useful for modeling a 3D avatar that signs Libras' sentences in a more natural and intelligible way.

It is worth emphasizing that the hierarchical clustering of the signs identifies the classes amongst adjectives, verbs and nouns, given possible manual parameters for intensification of signs. In the future, we will increase the number of signs and add facial expressions for clustering. The intention is to check the possibility of additional or different grouping by combining manual and non-manual expressions. The analysis indicated that the manual parameters are more determinant than the valence to classify the adjectives. With a larger corpus, it would be possible to apply a sentiment classifier and analyze the interaction between the parameters and their positive and negative valences.

Currently, we are developing an evaluation protocol to assess the quality of our translations by deaf individuals.

Acknowledgements. This work was supported by Capes/ SDH/ MCTI No. 59/2014 – Proc. # 88887.091672/2014-01.

References

1. Araújo, T.M.U.D.: Uma solução para geração automática de trilhas em língua brasileira de sinais em conteúdos multimídia. 2012. 197 f. Tese (Doutorado em Engenharia Elétrica e de Computação) - Universidade Federal do Rio Grande do Norte, Natal (2012)
2. Bick, E.: The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Århus University Press, Århus (2000)
3. Bolinger, D.: Degree Words. Walter de Gruyter, Berlin (1972)
4. De Martino, J.M., Costa, P.D.P., Benetti, Â.B., Rosa, L.A., Kumada, K.M.O., SILVA, I.R.: Building a Brazilian Portuguese - Brazilian sign language parallel corpus using motion capture data. In: The 12th International Conference on the Computational Processing of the Portuguese Language, Tomar. Proceedings Workshop on Corpora and Tools for Processing Corpora Workshop, Tomar, pp. 56–63 (2016)
5. Felipe, T.: A relação sintático-semântica dos verbos e seus argumentos na LIBRAS. 1998. 143 f. Tese (Doutorado em Linguística) - Centro de Letras e Artes, Universidade Federal do Rio de Janeiro, Rio de Janeiro (1998)
6. Felipe, T.A., Monteiro, M.S.: Libras em Contexto: Curso Básico (Libras in context: Basic Course). WalPrint Gráfica e Editora, Rio de Janeiro (2007)
7. Ferreira-Brito, L.: Por uma gramática de línguas de sinais. Tempo Brasileiro, Rio de Janeiro (1995)
8. HandTalk. https://www.handtalk.me/. Accessed 4 Apr 2018
9. Muniz, M.C.M.: A construção de recursos linguístico-computacionais para o português do Brasil: o projeto Unitex-PB. 2004. 92f. Dissertação (Mestrado em Cências de Computação e Matemática Computacional), Universidade de São Paulo, São Carlos (2004)
10. Paiva, F.A.S., Will, A.D., De Martino, J.M., Barbosa, P.A., Benetti, A.B.: Incorporating non-manual expressions into a realistic signing avatar. In: Congreso Iberoamericano de Tecnologías de Apoyo a la Discapacidad, 2017, Bogotá. Memorias Congreso Iberoamericano de Tecnologías de Apoyo a la Discapacidad. Editorial Escuela Colombiana de Ingeniería, Bogotá, pp. 551–558 (2017)
11. ProDeaf. https://web.prodeaf.net/. Accessed 04 Apr 2018
12. Quadros, R.M., Karnopp, L.B.: Língua de sinais brasileira: estudos lingüísticos. Artmed Editora, Porto Alegre (2004)
13. Wilbur, R.B., Malaia, E., Shay, R.A.: Degree modification and intensification in american sign language adjectives. In: Aloni, M., Kimmelman, V., Roelofsen, F., Sassoon, G.W., Schulz, K., Westera, M. (eds.) Logic, Language and Meaning. LNCS, vol. 7218, pp. 92–101. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31482-7_10
14. Xavier, A.N.: Uma ou duas? Eis a questão! Um estudo do parâmetro número de mãos na produção de sinais da língua brasileira de sinais (libras). 2014. 178f. Tese (Doutorado em Linguística), Universidade Estadual de Campinas, Campinas (2014)
15. Xavier, A.N., Barbosa, P.: Com quantas mãos se faz um sinal? Um estudo do parâmetro número de mãos na produção de sinais da língua brasileira de sinais (libras). Todas as Letras-Revista de Língua e Literatura **15**(1), 111–128 (2013)

Using a Discourse Bank and a Lexicon for the Automatic Identification of Discourse Connectives

Amália Mendes(✉) and Iria del Río

University of Lisbon, Center of Linguistics-CLUL, Lisbon, Portugal
{amaliamendes,igayo}@letras.ulisboa.pt

Abstract. We describe two new resources that have been prepared for European Portuguese and how they are used for discourse parsing: the Portuguese subpart of the TED-MDB corpus, a multilingual corpus of TED Talks that has been annotated in the PDTB style, and the Lexicon of Discourse Markers for Portuguese (LDM-PT). Both lexicon and corpus are used in a preliminary experiment for discourse connective identification in texts. This includes, in many cases, the difficult task of disambiguating between connective and non-connective uses. We annotated the PT-TED-MDB corpus with POS, lemma and syntactic constituency and focus on the 10 most frequent connectives in the corpus. The best approach considers word-form+POS+syntactic annotation and leads to 85% precision.

1 Introduction

While annotation levels such as POS, lemmatization, and syntactic relations have been consistently addressed for English and other languages with good results in terms of resource availability and tool development, work on the higher levels of text and discourse is still scarce, even for English. In the case of the Portuguese language, resources and tools for semantics and discourse are few, and are frequently only available for one variety of Portuguese, Brazilian or European Portuguese [3].

To be able to address discourse parsing, it is important to count on linguistically informed data that will provide the necessary input for the automatic identification of text spans explicitly connected by a discourse marker (DMs) or implicitly related through a discourse relation (also referred to as rhetorical sense), such as cause, justification, condition, elaboration, instantiation. In this paper, we describe two new resources that have been prepared for European Portuguese and how they are used for discourse parsing. One such resource is the Portuguese subpart of the TED-MDB corpus, a multilingual corpus of TED Talks that has been manually annotated in the PDTB style [18] with some adaptations required by the multilingual character of the corpus and by the specific genre (prepared speech) [34]. Another resource is the Lexicon of Discourse Markers for Portuguese (LDM-PT), that was compiled from grammars and corpora, and that provides information on a set of 222 DMs in European Portuguese.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 211–221, 2018.
https://doi.org/10.1007/978-3-319-99722-3_22

With the goal of building an automatic system for discourse parsing in European Portuguese, we performed a first experiment focused on the automatic identification of discourse connectives. The LDM-PT lexicon provided the list of candidate connectives, and the discourse-annotated corpus was further labeled for POS and parsed with detailed syntactic categories. We then evaluated the level of ambiguity of the identification task and we investigated which linguistic information contributed more to the recognition of discourse connectives. These results shed light on the linguistic features that are especially helpful for the automatic identification task.

The paper is organized as follows: we review related resources on discourse and discourse processing, especially for Portuguese, in Sect. 2. We present the TED-MDB corpus in Sect. 3 and the LDM-PT lexicon in Sect. 4, before addressing the automatic DM's identification in Sect. 5.

2 Related Work

Discourse parsing involves discourse connective identification (distinguishing between connective and non connective uses if required), the delimitation of the arguments of the discourse relation, and discourse relation labeling. As discourse connectives can frequently express several rhetorical relations, sense disambiguation is another required step.

There are several discourse-annotated corpora in different theoretical frameworks. The PDTB [18] style of annotation has been applied to other languages besides English, such as Turkish [33], Chinese [35], Czech [26], and applied to English and French speech data [6]. For Brazilian Portuguese, several corpora have been annotated in the RST and CST frameworks (CSTNews, CorpusTCC, Rhetalho, Summ-it) [1,14].

Lexicons of DMs are even rarer than discourse-annotated corpora. The German lexicon DiMLex [29] and the French lexicon LEXCONN [24] are two of the first initiatives. DIMLex includes 275 connectives and provides information on orthographic variants, non-connective readings, focus particle, syntactic category and, more recently, discourse relations [27]. LEXCONN describes 328 connectives and provides a syntactic category and the set of discourse relations that apply to each connective, based on SDRT. Both lexicons have inspired the development of recent lexicons for other languages, such as the Italian lexicon LiCO (173 connectives) [9]. For Spanish, the DPDE, an online dictionary of Spanish discourse markers with 210 entries, covers both written and spoken data and provides a definition, together with detailed information on each connective, such as register, prosody, formulae and comparable markers [4]. The dictionary provides a Portuguese semi-equivalent to the Spanish particles. Recently, the design of a Czech lexicon of DMs that exploits the Prague Dependency Treebank was presented in [16]. Several lexicons have been converted to the DIMLex format to integrate Connective-lex, a multilingual lexicon of discourse markers [8,31].

Although all these lexical resources address discourse related devices, the type of unit that they contain can vary considerably. DIMLex, LEXCONN and most of the lexicons cover discourse connectives, while the DPDE targets mainly discourse markers in speech with pragmatic and interactional meaning [7]. Even when focusing exclusively on discourse connectives, the lexicons may be restricted to the more typical categories (conjunctions, prepositions and adverbial phrases) or include a larger set of expressions that fulfill a cohesive function in a specific context. This occurs frequently in cases where the units were extracted from a discourse bank. For instance, the PDTB includes Alternative Lexicalizations [21,22] and the Prague Discourse Treebank includes secondary connectives (and free connective phrases) [25], that fall outside the traditional categories associated to discourse connectives.

There are different approaches to discourse parsing, from rule-based methods [13] to machine learning techniques [19]. For English, [20] extracted explicit discourse connectives in the PDTB and disambiguated their senses. Other work in sense identification includes [11], as well as the CoNLL Shared Task (http://www.cs.brandeis.edu/~clp/conll15st/). Lopes [12] reports an experiment for fully automatic identification of multilingual lexica including Portuguese. For most languages other than English, work on discourse parsing is either scarce or non-existent. However consistent work has been developed in discourse processing for Brazilian Portuguese: the corpora annotated with discourse information have lead to manual and automatic discourse annotation in the RST and CST frameworks (RST Toolkit, DiZer, CSTParser) [1,14]. To our knowledge, no such resources exist for the European variety of Portuguese. Hence, our goal is to contribute with resources and tools for the development of state-of-the-art discourse parsers for this variety.

3 The Discourse Bank: TED-MDB

The TED-Multilingual Discourse Bank (TED-MDB) is a corpus of TED talks transcripts involving six languages (English, German, Polish, Portuguese, Russian and Turkish), annotated for discourse relations [34]. Two of the talks have been aligned and can be queried on the TextLink portal[1].

TED talks are prepared presentations delivered to a live audience. The transcripts are prepared according to the norms of written language (e.g., they include punctuation) and are translated to various languages by volunteers, and revised. An XML version of the transcripts in all languages is available at the WIT3 website [5]. The TED-MDB corpus contains 6 talks annotated in the PDTB style of annotation: discourse relations that are either explicitly marked by a discourse connective or that can be inferred from the context are labeled. These relations may hold at the inter-sentential or the intra-sentential level. In TED-MDB, both explicit and implicit relations are labeled at the inter-sentential level, while only explicit relations are annotated at the intra-sentential level.

The annotation of an explicit relation labels the discourse connective, its two arguments and its sense. TED-MDB follows the PDTB 3.0 relation hierarchy, which has 4 top-level senses (Expansion, Temporal, Contingency, Contrast) and

[1] http://ec2-18-219-79-53.us-east-2.compute.amazonaws.com:8000/ted_mdb/.

their second- or in some cases third-level senses [32]. We give an example of an explicit inter-sentential (1) relation. The discourse connective is underlined, the first argument is rendered in italic, and the second argument in bold. An example of an implicit inter-sentential relation is given in (2): in this case, there is no overt connective and the annotation provides a connective that expresses the inferred relation. As in PDTB, TED-MDB considers non prototypical devices that assure coherence in the text. Such elements are labeled Alternative Lexicalizations and one such example is given in (3). The original English transcript is provided in the examples.

1. *Ela disse-me que algumas delas não correspondiam à sua marca, às suas expectativas.* Na verdade **uma das obras de tal modo não correspondia à sua marca, que ela tinha-a posto no lixo no seu estúdio**. (She told me that a few didn't quite meet her own mark for what she wanted them to be. One of the works, in fact, so didn't meet her mark, she had set it out in the trash in her studio) [Expansion:Instantiation] (TED Talk no. 1978)
2. *esta companhia tem a visão direcionada para o que eles chamam de "o novo Novo Mundo".* (Implicit = porque) **São quatro mil milhões de pessoas da classe média que precisam de comida, de energia e de água.** (this company has their sights set on what they call "the new New World." That's four billion middle class people demanding food, energy and water.) [Contingency:Cause:Reason] (TED Talk no. 1927)
3. *muitos desses amputados do país não usavam as suas próteses.* A razão, como vim a saber mais tarde, era que **o encaixe das próteses era doloroso por não ser um encaixe perfeito**. (many of the amputees in the country would not use their prostheses. The reason, I would come to find out, was that their prosthetic sockets were painful because they did not fit well) [Contingency:Cause:Reason] (TED Talk no. 1971)

4 The Lexicon: LDM-PT

The Lexicon of Discourse Markers (LDM-PT) [15] provides a set of lexical items in Portuguese that have the function of structuring discourse and ensuring textual cohesion and coherence at intra-sentential and inter-sentential levels [10]. Each discourse marker (DM) is associated to the set of its rhetorical senses (also named discourse relations or coherence relations), following the PDTB 3.0 sense hierarchy (Webber et al., 2016).

Discourse connectives are taken in the lexicon as elements that do not vary regarding inflection, express a two-place semantic relation, have propositional arguments and are not integrated in the predicative structure. This includes conjunctions, adverbs and adverbial phrases, but also prepositions and alternative lexicalizations, as defined in the PDTB (see Sect. 3). The DMs were taken from several sources: grammars; corpus-driven lists for the main POS, such as conjunctions and prepositions; manual contrastive approach between English and Portuguese, based on the parallel Europarl corpus (the manual identification of connectives based on a contrastive language analysis calls attention to other lexical strategies that express coherence relations between text spans); and, mainly, the automatic extraction of the DMs that are labeled as connectives in the Portuguese part of the TED-MDB corpus.

As a result, the lexicon mainly reflects the decisions taken in the treebank in what concerns which rhetorical senses are associated with a connective. In the TED-MDB treebank, the intrinsic values of the DM are included, and values that may be triggered by adjacency between sentences and by the lexical content of the clauses are excluded. When the contexts leads to infer an additional sense, the explicit DM is labeled with its prototypical sense and an implicit relation is added to describe the sense that is inferred from the context, as in the PDTB [23]. One such example in TED-MDB is provided below: the explicit coordinate conjunction (underlined) is labeled with the sense Expansion:Conjunction (4) and an additional implicit DM (underlined and in parentheses) accounts for the inferred sense Contingency:Cause:Result (5).

4. *Estas iniciativas criam um ambiente de trabalho mais móvel* e **reduzem a nossa pegada imobiliária**. (TED talk 1927) (These initiatives create a more mobile work environment and reduce our housing footprint.)

5. *Estas iniciativas criam um ambiente de trabalho mais móvel e* (portanto) **reduzem a nossa pegada imobiliária**. (These initiatives create a more mobile work environment and consequently reduce our housing footprint.)

The lexicon includes both continuous (*porque* 'because', *então* 'then', *na verdade* 'in fact') and discontinuous units (*por um lado... por outro lado* 'on the one hand... on the other hand', *tal como... também* 'just as... so too'), and this information is part of the features of the XML structure. The typology is more detailed than the one found in the treebank: the connectives are divided in primary connectives, secondary connectives and alternative lexicalizations. The latter were described in 3. The distinction between primary and secondary connectives follows the proposal of [25]. Primary connectives are prototypical discourse connectives such as conjunctions, prepositions, adverbs and adverbial phrases. Secondary connectives are devices with a lesser degree of lexicalization, where one element (usually a deitic) is typically replaceable: *antes disso* 'before that', *da mesma maneira* 'in the same way', *nessa altura* 'at that time'.

The lexicon provides information on restrictions on the mood of the clause introduced by the DM and on its tense. For each discourse connective/sense pair, one or more English near-synonyms are listed. They are extracted, when applicable, from the DiMLex-en, compiled from data from the PDTB (Stede et al., 2017). Each entry of the lexicon provides a corpus example and information on the source of the example. Contrary to DIMLex, there is no feature in LDM-PT that identifies possible non connective uses of the DMs. The XML version of the lexicon was converted to the DIMLex format and is integrated in a multilingual resource [31] through a web app (at Connective-Lex.info) [8].

5 Automatic Identification of Connectives

5.1 The Ambiguity of Discourse Connectives

Argument identification is the first step of discourse parsing and has a central role in building quality discourse representations [28]. We understand argument identification as in [11] that is, the identification of the different elements that compose a discourse relation (explicit or implicit and inter or intra- sentential): potential discourse markers and arguments.

In many cases, words that have a cohesive function in texts may also have non connective functions, that is, they are ambiguous [30]. As we mentioned in Sect. 4, the lexicon does not provide any information on those cases. For instance, the adverb *assim* 'in such a way' modifies the pronoun in (6) and does not perform a cohesive function at the discourse level. However, it is indeed a connective when connecting two sentences in (7) with a Result sense. Another very frequent case of ambiguity are coordinating conjunctions, that connect lower-level phrases such as nominal phrases (8)[2], or high-levels constituents, such as clauses and sentences (4). Only the latter cases are to be included in a discourse annotation task.

6. Isto tem que ser feito com grande precisão, mas se o conseguirmos, se conseguirmos construir esta tecnologia, se a colocarmos no espaço, poderão ver algo *assim*. (This has to be done very precisely, but if we can do this, if we can build this technology, if we can get it into space, you might see something like this.) TED Talk no. 1976
7. Eles acreditam que o ASG tem o potencial de criar impacto em riscos e receitas, *assim*, incorporar o ASG no processo de investimento é fundamental ao seu dever de agir no melhor interesse dos membros do fundo... (They believe that ESG has the potential to impact risks and returns, so incorporating it into the investment process is core to their duty to act in the best interest of fund members...) TED Talk no. 1927
8. As companhias *e* os investidores não são os únicos responsáveis pelo destino do planeta. (Companies and investors are not singularly responsible for the fate of the planet.) TED Talk no. 1927

5.2 Identification of Connectives

To pursue the identification of connectives, we used a data-driven approach that exploits the information encoded in LDM-PT and in the Portuguese section of the TED-MDB corpus.

As a first step, we extracted all the explicit discourse relations in the corpus and we identified the explicit connectives with their sense (PDTB 3.0 sense hierarchy). There are 275 instances of explicit connectives. These connectives

[2] Nominalizations (e.g., the destruction of the city) can be considered as equivalent to clauses and part of the discourse level, as in the PDTB (although few such cases are actually annotated), so that coordinating conjunctions connecting nominalizations would have to be identified as discourse connectives.

Table 1. Distribution of word-forms, connectives and non-connectives in the corpus for the ten most common connectives.

Word-forms	Connectives	NonConnectives
569	224 - 39%	345 -61%

correspond to 42 different word-forms with 886 cases in the corpus. Therefore, only a 31% of the possible candidates are effectively working as connectives in our data.

The ten most common connectives (by lemma) in the corpus are: *e* (and), *mas* (but), *para* (for/to), *se* (if), *quando* (when), *porque* (because), *depois* (after), *por* (for/because), *ou* (or), *então* (then). They account for 81% of the total cases (569 word-forms, 224 connectives, 345 non-connectives). Considering this fact and that we were performing a preliminary experiment, we restricted our analysis to these ten connectives.

In our list of ten connectives, we have six conjunctions, two prepositions and two adverbs. It is interesting to note that conjunctions account for 69% of the total connectives in the corpus. In fact, a single conjunction, *e* (and), accounts for a 32% of the total occurrences of connectives in the corpus. However, only a 37% of the occurrences of the word *e* have a discourse connective function. All these aspects are relevant for testing the ambiguity of connectives.

As a second step, we automatically annotated the PT-TED-MDB corpus with lemma, POS and syntactic information. For POS and lemma, we used the Portuguese module of Freeling [17]. Freely available Portuguese parsers are scarce. We tested different options and we chose the constituency representation of the parser PALAVRAS [2] because its syntactic trees contain rich linguistic information[3]. To investigate the contribution of different linguistic features to the identification task, we first defined three levels of linguistic information: word-form of the connective; POS and lemma of the connective; word-form, POS, lemma and syntactic information involving the connective and its context. We then applied a rule-based method that makes use of these levels of linguistic information, and we measured precision (and, in some cases, recall) in the identification of connectives and non-connectives in the corpus. We describe our results in the following paragraphs.

(1) Word-form.
In this approach, we consider that each word-form that can be a connective is effectively working as a connective, and we measure precision for the identification of connectives and non-connectives.
As expected, word-form is not enough to identify connectives accurately. Word-forms corresponding to the ten most common connectives are effectively connectives in a 39% of their occurrences in the corpus. That is: considering that any word that can be a connective is working as such, we obtain a precision of 39% in the identification of connectives and a 0% of

[3] Also, dependency analysis is not available in the upload interface of PALAVRAS.

precision in the identification of non-connectives (because all occurrences are considered connectives).
For some connectives ambiguity is low. For example, *quando* (when) works as a connective in 94% of its occurrences. However, this connective represents only a 6% of the total use of connectives. In other cases, there is a higher level of ambiguity, as in the case of the most common connective *e* (and), mentioned above.

(2) Word-form + lemma + POS.
In this approach, we used the morphological information encoded in the LDM-PT corpus and the POS and lemma from Freeling to discriminate the connectives. Adding POS and lemma slightly improves precision: from 39% to 41% in the identification of connectives, and from 0% to 100% in the identification of non-connectives. Recall is 100% for connectives and 9% for non-connectives since, as in the previous approach, we consider most of the candidates as connectives. These results make sense considering the fact that connectives are words with low POS ambiguity. Indeed, we can see an improvement for word-forms with more than one POS (that are more or less equally frequent). This is the case of the connective *se* (if), which can be a conjunction (if) or a clitic pronoun.

(3) Word-form + lemma + POS + syntax.
In order to add syntactic information as a new layer, we used the constituency representation of the parser PALAVRAS (constraint grammar). This is the approach with the best performance. Using syntactic information, general precision increases to 85% for connectives and to a 99% in the identification of non-connectives, with a recall of 99% for connectives and a 89% for non-connectives. We experiment a slight decrease in recall for connectives and a high increase for non-connectives.
Syntactic information is especially relevant for connectives that can link different types of structures, like coordinated conjunctions. It is important to remember that the most common connective in the corpus is the copulative conjunction *e*, which accounts for 32% of all the connective cases. On the other hand, this conjunction is fairly common in the corpus, with 237 occurrences as word-form. Of these 237 occurrences, only 89 are connectives (37%) - the conjunction *e* works as a discourse connective when it links clauses (as in (4)) or sentences.

Using syntactic information from PALAVRAS' output, we can identify all the cases where *e* is linking clauses/sentences. Following this approach, we got an 89% of precision and a 100% of recall identifying the connective uses of this conjunction. Since conjunctions account for a 83.5% of the total connectives in the corpus, the use of syntactic information highly improves the results.

Connectives that are used in specific constructions could be identified with simpler approaches, like pattern matching. It is the case of the prepositions *por* (because/for) and *para* (for/to). Those connectives have a unique POS, and they work as connectives in a very specific construction: when they introduce infinitive subordinated clauses (*para fazer isso* (to do so)). For these uses, it would be enough to identify the cases where the preposition is followed by an infinitive/adverb+infinitive. This simple approach, however, would not be enough for conjunctions like *e* (and) or *mas* (but), that can introduce multiple types of structures and which can be located far from the verb when they introduce clauses. Defining a clause with a surface pattern can be difficult and introduce a lot of errors.

6 Conclusion

We have presented work on discourse processing for Portuguese, based on LDM-PT, a new lexicon of DMs for Portuguese and on the Portuguese part of the multilingual treebank TED-MDB. Both resources account for a wide range of syntactic categories: conjunctions, prepositions, adverbs and adverbial phrases, but also alternative lexicalizations that carry a cohesive function in texts.

Both lexicon and corpus are used in a preliminary experiment for discourse connective identification in texts. This includes, in many cases, the difficult task of disambiguating between connective and non-connective uses. We annotated the PT-TED-MDB corpus with POS, lemma, using Freeling, and syntactic constituency using the PALAVRAS parser. We focus here on the 10 most frequent connectives in the corpus, and in some cases, also the most ambiguous ones between connective and non connective uses. We test the results of adding layers of annotation in our identification task. Using word-form+POS information only provides an increase in precision from 39 to 41, performing better only in cases where a word-form has more than one POS category. The approach that considers word-form+POS+syntactic annotation leads to 85% precision on the identification of connectives. Syntactic information for complex sentences, with coordinated or subordinated clauses, has a high impact in the identification of conjunctions working as connectives.

In the future, we plan to extend this approach to all the connectives in our corpus, experimenting also with a dependency representation. We want to explore the identification of connectives in nominalization structures, accounted for both in the PDTB and in the TED-MDB. Taking the discourse processing further will lead to the task of sense attribution for each discourse relation.

Acknowledgments. This work was partially supported by national funds through FCT - Fundação para a Ciência e a Tecnologia (under the project PEst-OE/LIN/UI0214/2013), and some of its developments were implemented in the scope of the COST Action TextLink – Structuring Discourse in Multilingual Europe3.

References

1. Aleixo, P., Pardo, T.A.: CSTTool: um parser multidocumento automático para o português do brasil. In: Proceedings of the IV Workshop on M.Sc. Dissertation and Ph.D. Thesis in Artificial Intelligence - WTDIA, pp. 140–145 (2008)
2. Bick, E.: The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Arhus, Århus (2000)

3. Branco, A., et al.: The Portuguese Language in the Digital Age/A Língua Portuguesa na Era Digital. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29593-5
4. Briz, S.P.B., Portolés, J.: Diccionario de partículas discursivas del español (2003). http://www.dpde.es
5. Cettolo, M., Girardi, C., Federico, M.: WIT3: web inventory of transcribed and translated talks. In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT), vol. 261, p. 268 (2012)
6. Crible, L.: Discourse markers and (dis)fluency across registers : a contrastive usage-based study in English and French. Ph.D. thesis, Louvain (2007)
7. Cuenca, M.J., Marín, M.J.: Co-occurrence of discourse markers in catalan and spanish oral narrative. J. Pragmat. **41**, 899–914 (2009)
8. Dombek, F.: Connective-lex.info - a web app for a multilingual connective database. Bachelor thesis, Potsdam (2017)
9. Feltracco, A., Jezek, E., Magnini, B., Stede, M.: Lico: A lexicon of Italian connectives. In: Proceedings of the 3rd Italian Conference on Computational Linguistics, Napoli, Italy (2016)
10. Halliday, M., Hasan, R.: Cohesion in English. Longman, Harlow (1976)
11. Lin, Z., Ng, H.T., Kan, M.Y.: A PDTB-styled end-to-end discourse parser. Nat. Lang. Eng. **20**(02), 151–184 (2014)
12. Lopes, A., et al.: Towards using machine translation techniques to induce multilingual lexica of discourse markers. http://arxiv.org/abs/1503.0914 (2015). Accessed 15 Jan 2016
13. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge (2000)
14. Maziero, E., Pardo, T.A.: CSTPARSER - a multi-document discourse parser. In: Proceedings of the PROPOR 2012 Demonstration, pp. 1–3 (2012)
15. Mendes, A., del Rio, I., Stede, M., Dombek, F.: A lexicon of discourse markers for portuguese - LDM-PT. In: Proceedings of LREC 2018 (2018)
16. Mírovský, J., Synková, P., Rysová, M., Poláková, L.: Designing CzeDLex - a lexicon of Czech discourse connectives. In: Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (2016)
17. Padró, L., Stanilovsky, E.: Freeling 3.0: towards wider multilinguality. In: Proceedings LREC 2012 (2012)
18. PDTB group: the penn discourse treebank 2.0 annotation manual. Technical report Institute for Research in Cognitive Science, University of Philadelphia (2008)
19. Pitler, E., Nenkova, A.: Using syntax to disambiguate explicit discourse connectives in text. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 13–16. Association for Computational Linguistics, Stroudsburg, PA, USA (2009)
20. Pitler, E., Nenkova, A.: Using syntax to disambiguate explicit discourse connectives in text. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 13–16. Association for Computational Linguistics (2009)
21. Prasad, R., et al.: The penn discourse treebank 2.0. In: LREC (2008)
22. Prasad, R., Joshi, A., Webber, B.: Realization of discourse relations by other means: alternative lexicalizations. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 1023–1031. Association for Computational Linguistics (2010)
23. Rohde, H., Dickinson, A., Clark, C., Louis, A., Webber, B.: Recovering discourse relations: varying influence of discourse adverbials. In: Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pp. 22–31 (2015)

24. Roze, C., Danlos, L., Muller, P.: LexConn: a French lexicon of discourse connectives. Revue Discours (2012)
25. Rysová, M., Rysová, K.: Secondary connectives in the prague dependency treebank. In: Hajičová, E., Nivre, J. (eds.) Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pp. 291–299. Uppsala, Sweden (2015)
26. Rysová, M., et al: Prague Discourse Treebank 2.0 (2016)
27. Scheffler, T., Stede, M.: Adding Semantic relations to a large-coverage connective lexicon of German. In: et al., N.C. (ed.) Proceedings of LREC 2016 (2016)
28. Soricut, R., Marcu, D.: Sentence level discourse parsing using syntactic and lexical information. In: Proceedings of NAACL 2003, vol. 1, pp. 149–156. Association for Computational Linguistics, Stroudsburg, PA, USA
29. Stede, M.: DiMLex: a lexical approach to discourse markers. In: Exploring the Lexicon - Theory and Computation, Edizioni dell'Orso, Alessandria (2002)
30. Stede, M.: Discourse Processing. Morgan & Claypool Publishers, San Rafael (2011)
31. Stede, M., Scheffler, T., Dombek, F.: Connective-lex.info. Potsdam University (2017). http://connective-lex.info
32. Webber, B., Prasad, R., Lee, A., Joshi, A.: A discourse-annotated corpus of conjoined VPs. In: Proceedings of the 10th Linguistics Annotation Workshop, pp. 22–31 (2016)
33. Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., Çakıcı, R.: Turkish discourse bank: porting a discourse annotation style to a morphologically rich language. Dialogue Discourse **4**(2), 174–184 (2013)
34. Zeyrek, D., Mendes, A., Kurfalı, M.: Multilingual extension of PDTB-style annotation: the case of ted multilingual discourse bank. In: LREC (2018)
35. Zhou, Y., Xue, N.: PDTB-style discourse annotation of Chinese text. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1, pp. 69–77. Association for Computational Linguistics (2012)

The Other C: Correcting OCR Words in the Presence of Diacritical Marks

Sérgio Luís Sardi Mergen[1](✉) and Leonardo de Abreu Schmidt[2]

[1] Departamento de Linguagens e Sistemas de Computação,
Universidade Federal de Santa Maria (UFSM), Santa Maria, RS, Brazil
mergen@inf.ufsm.br

[2] Programa de Pós-Gradação em Informática - PPGI (UFSM),
Santa Maria, RS, Brazil
lschmidt@inf.ufsm.br

Abstract. We propose a lexicon based method whose purpose is correcting a word recognized by an OCR engine (a classifier). This post-processing method was originally designed to be used for language models that support diacritical marks, such as Portuguese. Since these special marks can be confused with noise by the classifier, wrong predictions can be derived if only the top hypothesis per glyph of the original image is preserved. To cope with this, our method uses a filtering strategy to select the best hypotheses for each glyph, which are used to produce candidate queries. A best query is selected in terms of confidence rate and edit distance to the word. A similarity search method over the best query suggests a correction. Experiments show the method improves prediction accuracy considerably for Portuguese words correction.

Keywords: OCR · Similarity search · Classifier

1 Introduction

The OCR problem has gain a lot of attention in the past decades. Several works proposed different techniques to enhance accuracy [6]. Since pixel-based computation is very time consuming, one common approach is to decompose glyphs into a set of features, such as the presence of closed loops or lines intersections. Classifiers predict the characters based on the informed features. Pre-processing and pos-processing steps are also commonly used. The pre-processing involves cleaning and correcting glyphs [4]. The post-processing involves correcting individual characters or whole words hallucinated by the classifier [10].

In this paper we focus on the post-processing problem of word correction. The correction is required when the classifier produces words with typographical errors, i.e. spurious words produced when a correct word is disrupted by the addition, deletion or substitution of characters. The word correction problem can be handled using lexicons (dictionaries formed by admissible words). The correction is a valid word (taken from the lexicon) that is most similar to the query (the

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 222–230, 2018.
https://doi.org/10.1007/978-3-319-99722-3_23

word recognized by the classifier), where similarity is measured by the number of edit operations that transforms the query into the valid word. This approach is particularity interesting for Portuguese words correction because of the intense use of diacritical marks, like ‘Ç’. Classifiers may introduce typographical errors by confusing a diacritic with noise (turning ‘faça’ into ‘faca’), or with a part of another character shape (turning ‘faça’ into ‘fava’).

We call the attention to the fact that classifiers produce a list of hypotheses per glyph, that is, confidence rates of the glyph being a specific character. Only the highest ranked hypothesis is used to form the recognized word, which is then submitted to the lexicon search correction stage.

We propose an approach that copes with the uncertainty of the OCR classifier by accepting hypotheses (characters) with lower confidence rates as well. In general terms, the approach works as follows: first, a filtering stage defines the lists of hypotheses per glyph. Then, a query expansion technique creates several candidate queries, by crossing all lists. Finally, the queries are submitted to a lexicon search stage. One of those queries is the best in terms of distance to the original image (to be determined during search time) and confidence rate (determined by the classifier). It is up to the best candidate query to provide a single answer to the original problem.

This paper is organized as follows: Sect. 2 discusses works whose purpose is correcting the words produced by an OCR engine. Section 3 introduces our method, describing the filtering strategies, the query expansion algorithm and the similarity search over the lexicon. Section 4 presents practical results. Finally, Sect. 5 brings our concluding remarks.

2 Related Work

There are several works regarding post-processing correction of OCR recognized words. We devise two main strategies: based on a lexicon and based on a probability model.

Lexicon based approaches aim at correcting a word using the most similar words taken from a dictionary (the lexicon). Similarity is usually measured with edit distance, that is, the minimum cost in terms of the edit operations required to transform one word into another.

Tesseract, one of the most popular OCR engines, uses many different lexicons with top ranked words of different categories [13]. A ranking system based on character features chooses the output from these categories. The work of [14] uses an intrinsic lexicon by levering from Google’s web search results. Candidate words are taken from the summary of the top ranked results. Edit distance is then used to select the best candidate for a given recognized word. In [11], the correction follows an orthogonal direction. Instead of using a recognized word as a query, the system uses frequent correct words as queries (focus words). Given a focus word (taken from a lexicon), the purpose is to find their recognized typographical variations that are within a predefined edit distance. The variations are then replaced with the corresponding focus word.

Lexicon based solutions are interesting if a comprehensive lexicon is available. It is not suited for low density languages, where the lexicon (if any) is not representative enough [7]. It is also limited for the recognition of words not usually found in a lexicon, such as named entities. For such cases, probability models can be used.

A probability model is usually built over the concept of n-grams, where the purpose is to predict the probability of an element appearing after a sequence of n elements. The model can be built as a finite state (with n as the maximum path) where the transitions indicate the probability of going from one state to the next (a Markov model). Elements can be characters, for the correction of single characters, or words, for the correction of complete words.

With respect to word correction, error models are used to assign a cost for characters insertion, deletion or substitution. One recent work shows that the insertion and deletion of a space enables spotting words that should be separated and merged, respectively, in cases where the OCR failed to properly separate the words [7]. In [5], the error model uses individual weights for character substitution. Smaller weighs are assigned to pairs of characters where the OCR normally mistakes one for the other. The weights can be adjusted as new statistics on errors become available. The best sequence of characters that form a word can also be predicted as the maximum likelihood path on a Markov model [3]. This path can be found using Viterbi Algorithm, a Dynamic Programming solution. The work of [9] also employs a finite state automaton combined with an error model. Interestingly, the automaton is modelled with the complete list of OCR hypotheses, which is similar in spirit to our approach. The difference is that we use the different hypotheses to improve accuracy in a lexicon based method.

3 The Correction Method

The method proposed is designed to be used as a final part of an OCR engine, which we refer to as the classifier. We assume the original image was already correctly decomposed into a collection of glyphs $g \in G$, where each glyph corresponds to a character c, and that the classifier was trained over a language model C composed by output classes (characters) c.

Given the image, the classifier returns the complete list of hypotheses concerning every glyph $g \in G$ and every class $c \in C$. A hypothesis comes on the form of a normalized score whose value ranges from zero to one, where greater values mean higher confidence rates. In what follows we show how our method uses the hypotheses in order to find one single answer as the correction.

3.1 The Filter Stage

The filter stage reduces the number of candidates per glyph g. Assume there is a function filter(g) that returns a sub-list of all possible hypotheses for g. The function applies two filtering steps: Top-k and threshold.

- **Top-k**: keeps only the k candidates per glyph with the higher score.
- **Threshold**: keeps only the candidates whose score lies above a predefined threshold.

The threshold strategy can be used to fine-tune the prediction, by accepting hypotheses whose confidence is reasonably high. On the other hand, the top-k strategy is intended to reduce the search space during the query expansion.

The two filters can be used in conjunction to create a balance between an excessive and an insufficient number of candidates. A number of different strategies can be devised. For instance, to assure all glyphs are assigned to a class, one can define a strategy that sets k to *one* and the threshold to *zero*. A more conservative approach could use a high threshold. A more relaxed approach could select more than one candidate regardless of their score. Section 4 discusses the strategy that achieved the best accuracy for the case study.

3.2 The Query Expansion Stage

This stage is responsible for creating candidate queries to be submitted to the similarity search stage. The candidates are intended to enhance the word prediction by amplifying the spectrum of possible answers.

Algorithm 3.1 explains the query expansion process. Assume $g.H$ gives the set of valid hypotheses for a glyph g. The function $addCharacter$ assigns a hypothesis h to each character of a query q and updates the query score. The score is the summed probability of the characters that are part of a query. The probabilities are taken from the classifier hypotheses (The function $score(g, h)$ gives the confidence rate that a glyph g corresponds to the hypothesis h). Queries with the highest ranked are formed by the most likely characters, according to the classifier.

```
create_Queries($G$)
for each $g \in G$ do
    $Q_2 \leftarrow \emptyset$
    for each $h \in g.H$ do
        for each $q \in Q_1$ do
            $addCharacter(q, h, score(g, h))$
            $Q_2 \leftarrow Q_2 \cup q$
        end for
    end for
    $Q_1 \leftarrow Q_2$
end for
return $Q_1$
```

Algorithm 3.1: Create queries from the classes that potentially match the glyphs of a word

The number of candidate queries is highly sensitive to the selectivity of the filter and the size of G. It can be determined as $\prod^{G} filter(g_i)$. On average, the number of queries is proportional to $\mu^{|G|}$, where μ is the minimum between k and the average number of characters above the threshold. A small threshold associated with a large k value will most likely produce a space whose search is infeasible.

3.3 The Similarity Search Stage

The similarity search stage runs each query individually. The purpose is to find the words that are most similar to the query. Similarity is measured by edit distance. Hence, the purpose is to find the words that have the minimum edit cost with respect the query.

There are basically two types of similarity search queries: range queries and k-nearest neighbor queries [1]. The former finds objects that are within a maximum distance (the range) from the query object. The latter finds the k objects that are closer to the query.Both strategies can be used to find all objects whose distance to the query is minimum.

Algorithm 3.2 describes how the search occurs using a straightforward approach based on range queries [8]. We start with the minimum possible range (0) and gradually increment the range by one unit until answers are found. Given a range d, candidate queries are executed in relevance order (queries whose confidence rate are higher are execute first). Results are found in a greedy way. Each query contributes with a set of answers A. If the set has at least one answer, the first one is chosen as the result.

```
correct_word($Q$)
$sort(Q)$
for $d = 0$ to $\infty$ do
    for each $q \in Q$ do
        $A \leftarrow similarity\_search(q, d)$
        if $A$ is not empty then
            return $A[0]$
        end if
    end for
end for
```

Algorithm 3.2: Finds a single word from candidate queries based on a minimum distance criteria

The Edit Distance Criteria: A single answer is selected, the best one according to distance and confidence rate. The algorithm prioritizes answers whose distance to the query is minimal. This strategy assumes a valid answer has few (or none) typographical errors with respect to the recognized OCR word. It is indeed the case for commercially available OCR systems, which reach 99% character accuracy [7]. Even assuming errors are local (tend to appear next to each other), [14] states that misrecognized words have an average of less than two errors in terms of edit operations.

The Confidence Rate Criteria: If more than one candidate query lead to answers with the same minimal distance, answers from the highest ranked query are given priority. The reason is simple. Those answers reflect the confidence rate of the classifier. It makes sense to trust these predictions.

Handling Ties: More than one answer may appear as alternatives for the best match if multiple candidate queries have the same maximum score or if the best candidate produces more than one answer. In such cases, the method arbitrarily chooses one. However, ties are not likely to appear. Candidates will seldom present the same score. In addition, multiple answers are possible only if more than one valid word has the same edit cost, which is less common when we consider few typographical errors (low edit costs).

4 Experiments

In this section we consider the problem of correcting words recognized by OCR. For the sake of simplicity, we assume the OCR was able to correctly identify the glyphs that are part of a word. We are interested in analyzing how different hypotheses for a glyph help in the process of word correction.

Since OCR engines are (in most cases) black boxes that hide the classification hypotheses, we built our own classifier in the form of a multi layer perceptron, whose implementation is detailed in [12]. The neural network was set with one hidden layer and an input layer with 784 inputs. The output layer had outputs for all lower-case Portuguese characters, including characters with diacritical marks. Learning rate and number of epochs were set to 0.001 and 500, respectively. The sigmoid was used as the activation function.

The words are formed by 1200×900 glyphs taken from a dataset containing handwritten characters. The glyphs were generated by 55 volunteers as part of a benchmark for recognizing characters in images of natural scenes [2]. Diacritical marks needed to be added manually. Glyphs from 54 volunteers were used for training the neural network. The glyphs from the remaining volunteer were used for testing. All glyphs needed to be preprocessed for grayscale reduction (28×28), cropping and rescaling according to the size of the input layer.

The testing revealed that adding diacritical marks into the neural network model led to incorrect predictions. For instance, without these special characters, only two characters were badly predicted ('e' and 's', as 'c' and 'p', respectively). When the network was trained with the diacritical marks, eight characters were badly predicted, out of the normal ones ('a', 'c', 'e', 'i', 'o', 's', 'u', 'w'), and five out of the special ones ('á', 'ã', 'ê', 'ó', 'õ'). Another interesting aspect is that the character 'à' was considered the best choice for a number of different inputs, such as 'a', 'b', 'e', 'g', 'l', 'm', 'q', just to name a few, with 100% probability. Curiously, the correct class also appeared as a best choice with the same probability.

It is also interesting to remark that 16 characters were correctly classified with 100% probability. Other 17 characters were correctly classified with 100%

probability, but were involved in ties with other characters (mostly with ‘á’). Two characters had their right classes in the second or third places (‘õ’, ‘ã’, ‘ó’), and the remaining characters were poorly predicted (their final positions were 5th, 6th, 10th, 19th, 23th, 30th).

In what follows we show results related to the recognition of words taken from a Portuguese dictionary with over 300.000 entries[1]. Different query-sets were used, each with a hundred words taken from the dictionary, which we assumed were handwritten by our ’testing volunteer’. Words of the same query-set are constrained by a minimum and a maximum length(3–5, 6–10, 11–15, 16–20). Given a word, the classifier produced hypotheses for all characters and our correction method chose a single word from a lexicon as the output. The lexicon is the very own dictionary where the word was originally taken from.

We have done several tests varying the top-k and the threshold. It turned out the results are only rarely affected when top-k is greater than 2. The reason is related to the fact that the classifier finds the correct class with $k = 2$ for most characters, as mentioned earlier. Therefore, we focus on using top-1 and top-2. The first one can be think of as a baseline that just uses the best output produced by the classifier. The threshold was set to 90%, but the same results were obtained when setting the threshold at a maximum.

Figure 1 shows the outcome. The first thing to notice is that accuracy at top-1 is considerably low for small to medium size queries. The reason is mostly related to the fact that many glyphs were mistaken with ‘á’. As a result, incorrect words were found, usually at high distances to the query. The accuracy is higher when using large queries, mainly because a sufficient number of correctly predicted characters were sent to the similarity search module. More importantly, taking the second hypothesis per glyph into account led to meaningful improvements. The wrong character was still selected at top-2 along with the correct character. However, the correct character led to (correct)answers at lower edit distances when compared to the (incorrect) answers found when the wrong character was used instead.

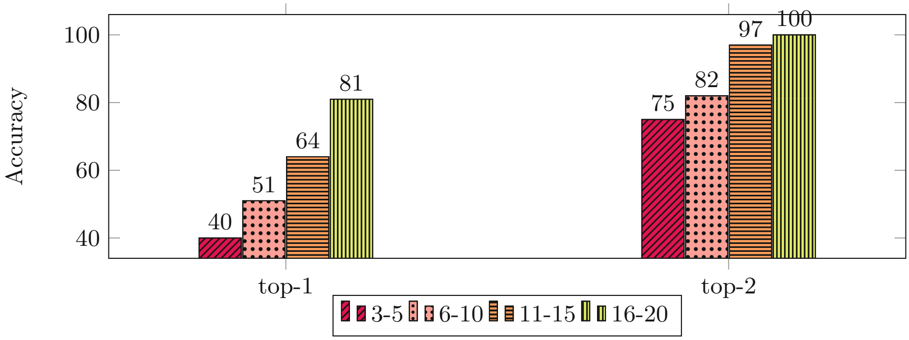


Fig. 1. Accuracy results for four query-sets using top-1 and top-2

[1] Vero - Brazilian Portuguese Spellchecking Dictionary & Hyphenator-2.0.8.

Finding the answer when $k = 2$ is reasonably fast, especially considering the high threshold used in the experiment. The computational cost is related to the number of candidate queries, which is proportional to the size of the query and the filtering parameters, as mentioned in Sect. 3.2. To better understand the computational cost, Fig. 2 shows the number of candidate queries that would be produced by four different queries with 10 glyphs each. Threshold varies from 80% to 100%, with an unlimited k.

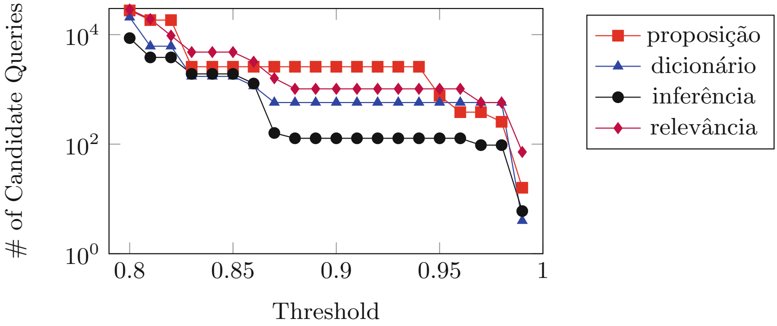


Fig. 2. Query expansion results when varying the threshold

Logically, a lower threshold means more candidates. The relation depends on the glyphs used and the hypotheses produced by the classifier. However, we can see that even at higher thresholds the number of candidates it reasonably high. We remark it is important to define a filtering strategy that preserves the true positives and at the same time reduces the number of candidates so that searching is feasible. The investigation of additional filtering strategies is left as future work.

5 Final Remarks

We have presented a post-processing OCR word correction method that is particularly suited for cases where the classifier is unable to issue accurate predictions for all glyphs that constitute a word. Experiments showed that the problem occurred more heavily when using a language model composed by diacritical marks. Of course, better results could be obtained using different neural networks in an attempt to prevent local minima. However, our purpose was to show that the method has room regardless of the classifier executed in the previous stage, as long as it issues normalized hypotheses.

The method, despite promising, could lead to the generation of a high number of candidate queries in different scenarios, which would make the lexicon search prohibitive. As mentioned before, more restrictive filtering strategies are one way of tackling this problem. Another interesting research topic is merging the query expansion with the lexicon search so that a single query needs to be submitted. This can be accomplished by building a string similarity search method that processes queries where each position accepts multiple disjoint characters.

References

1. Chen, L., Gao, Y., Zheng, B., Jensen, C.S., Yang, H., Yang, K.: Pivot-based metric indexing. Proc. VLDB Endow. **10**(10), 1058–1069 (2017)
2. Coates, A., et al.: Text detection and character recognition in scene images with unsupervised feature learning. In: Document Analysis and Recognition (ICDAR) International Conference on 2011, pp. 440–445. IEEE (2011)
3. Farooq, F., Jose, D., Govindaraju, V.: Phrase-based correction model for improving handwriting recognition accuracies. Pattern Recogn. **42**(12), 3271–3277 (2009)
4. Gupta, M.R., Jacobson, N.P., Garcia, E.K.: OCR binarization and image preprocessing for searching historical documents. Pattern Recogn. **40**(2), 389–397 (2007)
5. Hládek, D., Staš, J., Ondáš, S., Juhár, J., Kovács, L.: Learning string distance with smoothing for OCR spelling correction. Multimedia Tools Appl. **76**(22), 24549–24567 (2017)
6. Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. Pattern Recogn. **37**(5), 977–997 (2004)
7. Kolak, O., Resnik, P.: OCR post-processing for low density languages. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 867–874. Association for Computational Linguistics (2005)
8. Liu, S.G., Wei, Y.W.: Fast nearest neighbor searching based on improved VP-tree. Pattern Recogn. Lett. **60**, 8–15 (2015)
9. Llobet, R., Cerdan-Navarro, J.R., Perez-Cortes, J.C., Arlandis, J.: OCR post-processing using weighted finite-state transducers. In: Pattern Recognition (ICPR), 20th International Conference on 2010, pp. 2021–2024. IEEE (2010)
10. Llobet, R., Navarro-Cerdan, J.R., Perez-Cortes, J.C., Arlandis, J.: Efficient OCR post-processing combining language, hypothesis and error models. In: Hancock, E.R., Wilson, R.C., Windeatt, T., Ulusoy, I., Escolano, F. (eds.) Structural, Syntactic, and Statistical Pattern Recognition SSPR/SPR 2010. LNCS, vol. 6218, pp. 728–737. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14980-1_72
11. Reynaert, M.: Non-interactive OCR post-correction for Giga-scale digitization projects. In: Gelbukh, A. (ed.) CICLing 2008. LNCS, vol. 4919, pp. 617–630. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78135-6_53
12. Schaul, T.: Pybrain. J. Mach. Learn. Res. **11**(Feb), 743–746 (2010)
13. Smith, R.: An overview of the Tesseract OCR engine. In: Document Analysis and Recognition 2007 ICDAR Ninth International Conference on 2007, vol. 2, pp. 629–633. IEEE (2007)
14. Taghva, K., Agarwal, S.: Utilizing web data in identification and correction of OCR errors. In: Document Recognition and Retrieval XXI, vol. 9021, p. 902109. International Society for Optics and Photonics (2014)

Temporal Tagging of Noisy Clinical Texts in Brazilian Portuguese

Rafael Faria de Azevedo(✉), João Pedro Santos Rodrigues, Mayara Regina da Silva Reis, Claudia Maria Cabral Moro, and Emerson Cabrera Paraiso

Pontifícia Universidade Católica do Paraná, R. Imac. Conceição,
1155 - Prado Velho, Curitiba, PR, Brazil
{rafael,paraiso}@ppgia.pucpr.br, jpsanr@gmail.com,
mayara.reis@outlook.com, c.moro@pucpr.br
http://www.pucpr.br

Abstract. Temporal expressions are present in several types of texts, including clinical ones. The current research over temporal expressions has been done by the use of rule-based systems, machine learning or hybrid approaches, in most cases, over annotated (labeled) news texts correctly written in English. In this paper, we propose a method to extract and normalize temporal expressions from noisy and unlabeled clinical texts (discharge summaries) written in Brazilian Portuguese using a rule-based approach. The obtained results are similar to the state-of-the-art researches made with the same purpose in other languages. The proposed method reached a F1 score of 88.92% for the extraction step and, a F1 score of 87.89% for the normalization step.

Keywords: Temporal tagging · Clinical texts · Rules

1 Introduction

It is common to find words which indicate time in text. However, to make these words useful, it is necessary to extract and make them available to other systems [1]. Thus, temporal information extraction has been a topic of interest in recent years [3]. It is important to text processing tasks [16] such as question answering, search, text classification, text summarization among others [22,23,25,33].

One of the basic units for this process is the temporal expression (TE), which examples in Portuguese are: "*06 de abril de 2018*", "*ontem*", "*pós-operatório*", "*manhã da cirurgia*" and "*25/03/2018*". The whole process of temporal information extraction has three steps: temporal tagging, event tagging and temporal relation [29,35]. This work covers only the temporal tagging step [21]. In this step, a TE has to be extracted and normalized to a standard format which allows the TE to be used as an input of a question answering system, a chatbot, or a summarization system, for instance.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 231–241, 2018.
https://doi.org/10.1007/978-3-319-99722-3_24

Clinical texts are rich of temporal information. Applying temporal tagging over this texts is the first step to cope with more complex and useful processes such as the creation of the patient timeline and summarizing reports of a long time chronic patient. Due to pressure conditions in the workplace of health professionals, clinical texts are often not correctly written, containing typing errors, non-standard abbreviations etc. Considering the research of temporal tagging, most of them are done over annotated (labeled) news texts correctly written in English for NLP competitions like TempEval [34] and TempEval-2 [36]. Examples of clinical texts used in NLP competitions are the i2b2 2012 corpus [30] and the Clinical Tempeval corpus [2], but both are also well-formed and labeled (annotated) texts in English.

The main contribution of this paper is to present a method to extract and normalize TEs from noisy (raw) clinical texts written in Brazilian Portuguese. To accomplish this goal, the temporal tagger HeidelTime was used to make temporal tagging of correct and incorrect written TEs (we say noisy TEs). In this research, noisy TEs mean the ones with typing errors, spelling errors or unstandardized abbreviations (e.g. "*36 sem*" which is found in clinical texts meaning "thirty six weeks of pregnancy"). The correct form in Brazilian Portuguese would be "*36 semanas*". To cope with the noisy TEs, a n-gram strategy was used in the unlabelled data available. The method reached a F1 score of 88.92% for the extraction step and a F1 score of 87.89% for the normalization step. The obtained results are similar to the literature [9,10,17].

The rest of the paper is organized as follows. Section 2 presents the basic concepts present in the paper. Section 3 is about the related works. Section 4 presents the proposed method. Section 5 comments the experiments and results. Finally, Sect. 6 presents the conclusions and future work.

2 Basic Concepts

This section presents important concepts for this work, the first one is tagging. It is a process applied in text to mark events, verbs, time, names, places and other entities for many purposes [20,29]. Temporal tagging can be considered as a particular type of Named Entity Recognition (NER). It can be divided in two steps, extraction and normalization. In the extraction step TEs such as "*15-12-2013*" and "*ontem*" (yesterday) are extracted from text. In the normalization step, they are changed to a standard format as "2013-12-15" and "2013-12-14" [29]. In this work, four types of TEs were considered. They were defined by the TimeML standard, specifically, to its TIMEX3 tag [14,18,19,24]. They are:

- Date: refers to a point in time equal or greater than day, like in "*2 de novembro de 2017*" (November 2, 2017) or "2017".
- Time: refers to a point in time smaller than day, for example, parts of the day like "*quinta-feira de manhã*" (Thursday morning).
- Duration: deals with the length of an interval, which can be of different granularities like "*5 horas*" (5 h), "*3 anos*" (3 years) etc. A duration can also specify the point in time where an interval starts and ends.

- Set: serves to describe a set of times, dates, or frequency within a time interval that an event occurs. For example: "*todo sábado*" (every Saturday).

As one TE can be made of more than one word, it is important to distinguish between strict and relaxed matching, both concepts are related to the tagger evaluation. In a strict matching approach, when the TE "*quarta-feira á tarde*" (Wednesday afternoon) exists in the gold standard (test set), it is counted as correct for metric purposes, only if the temporal tagging system recognizes the TE written in the exact way it is in the gold standard. However, in a relaxed matching approach, if the system finds only the word "*quarta-feira*" in the analyzed text, it will be counted as a correct match with the gold standard. According to Strötgen and Gertz [29], the relaxed matching with correct normalization is usually considered as the most important evaluation measure. The results of this work are evaluated using a relaxed matching approach, considering the correct extraction and normalization steps. The next paragraph presents the temporal tagger used in this research.

HeidelTime[1] is a rule-based, multilingual, domain-sensitive temporal tagger developed at Heidelberg University, in Germany. It was designed to extract TEs from documents and normalize them according to the TimeML TIMEX3 annotation standard [14,18,19,24]. HeidelTime was chosen for this research because of its separation between the algorithmic part and the language resources part. The tagger contains manually created resources for 13 languages, Portuguese is one of them. For this specific language, HeidelTime original version only contains resources for the news domain [6,28]. Figure 1 shows these resources.

Figure 2 exemplifies HeidelTime extraction and normalization mechanism. Part (I) shows raw TEs. Part (II) shows the rule (regular expression) triggered in the tagger. Part (III) presents TEs normalized and ready to be used by another system. The element in bold "reUnitAbbrev" (II) is an extraction pattern resource (file) that contains abbreviations. The "*sem*" and "*sems*" abbrevia-

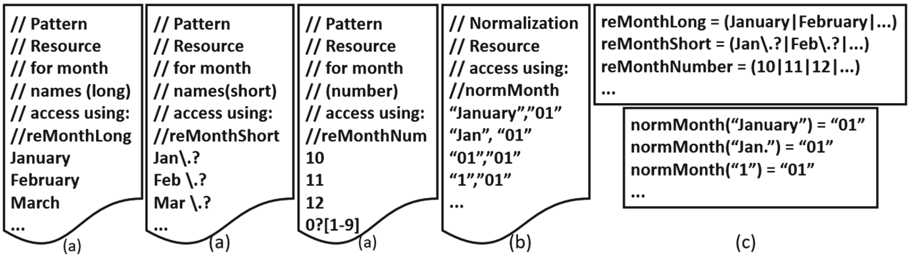
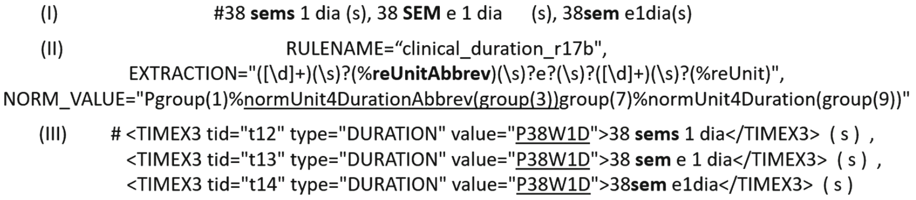


Fig. 1. Pattern resource files for different expressions for month names and numbers (a) and normalization resource file for month expressions (b), (c) represents how their information is used in HeidelTime after being read and translated by HeidelTime's resource interpreter. Adapted from [27].

[1] https://github.com/HeidelTime/heideltime.



(I) #38 **sems** 1 dia (s), 38 **SEM** e 1 dia (s), 38**sem** e1dia(s)

(II) RULENAME="clinical_duration_r17b",
EXTRACTION="([\d]+)(\s)?(%**reUnitAbbrev**)(\s)?e?(\s)?([\d]+)(\s)?(%reUnit)",
NORM_VALUE="Pgroup(1)%normUnit4DurationAbbrev(group(3))group(7)%normUnit4Duration(group(9))"

(III) # <TIMEX3 tid="t12" type="DURATION" value="P38W1D">38 **sems** 1 dia</TIMEX3> (s) ,
<TIMEX3 tid="t13" type="DURATION" value="P38W1D">38 **sem** e 1 dia</TIMEX3> (s) ,
<TIMEX3 tid="t14" type="DURATION" value="P38W1D">38**sem** e1dia</TIMEX3> (s)

Fig. 2. HeidelTime rule functioning example

tions (week and weeks respectively) in (I) were found in this resource "reUnitAbbrev". The underlined element in (II) is a normalization resource (file).

The next section presents papers related to this research.

3 Related Works

According to Kreimeyer and colleagues [11], the majority of NLP systems for capturing and standardizing unstructured clinical information use the rule-based approach (58.92%). The hybrid approach represents 33.92% and 7.16% of the systems use the machine learning approach. The hybrid approach is the combination of the other two. The main approach for the temporal tagging task is also the rule-based one.

Clinical texts usually have peculiar terms, abbreviations and wrong spelling, what turns the classification task harder when compared to the news texts (the commonest field studied in the temporal tagging task). In this case, a rule-based approach presents a more trustworthy result when temporal tagging is the scope. This understanding is presented by Chang and Manning [4], when its SUTime system is presented. It is a deterministic rule-based system for temporal tagging recognition of news-style text written in English, the system is part of the Stanford CoreNLP project. Lee and colleagues [12] present another rule-based strategy, the UWTime system. It is a context-dependent semantic parsing approach to extract and normalize TEs of the news domain written in English.

The current state-of-the-art system for temporal tagging is HeidelTime [27]. It is a rule-based system that was made to deal with multilingual and multi-domain text [28]. HeidelTime was developed to extract and normalize TEs from four domains: news, narrative, colloquial (short text) and scientific. It has been already extended to cope with the news domain in languages such as Spanish [7], Chinese [13] and Italian [15]. It was also extended to the narrative domain in Croatian [26]. Another extension of HeidelTime was presented by Hamon and Grabar [10], which is able to work with clinical narratives in English and French. This last work reached a F1 score of 94.31 for the extraction and normalization tasks in French.

Gupta and colleagues [9] presented an example of machine learning strategy to extract TEs using an artificial neural network. In their system, the classification output labels indicate if the type of the TEs is time, date, duration or frequency. The network is built to check where is the beginning, the middle (inside) and the end (outside) of each word classified as one of the four TE types. To train and test the classifier, news and clinical corpora written in English were combined. The F1 score got by their work was 84.08 in the extraction task.

The currently most used classifier in the temporal information extraction and/or normalization has been the Conditional Random Fields (CRF), especially because it uses a statistical approach that takes the correlation between words (tokens) into account. One example of its use is the work of Moharasar and Ho [17], who proposed a hybrid system which uses the CRF classifier. In their strategy, HeidelTime extracts TEs from clinical texts in English. Along with the HeidelTime extracted TEs, the authors generated lexical features to train a linear chain CRF. The F1 score reached by their work was 79.95 in the extraction task.

An example of extraction and normalization of TEs written in Portuguese is the research done by Costa and Branco [5]. Their work uses a hybrid approach, which combines the use of a classifier to predict each word of a text in three labels: B (begin), I (inside) and O (outside). The classification is combined with other features which are: current token, previous token and following token, position of a white space before the current token and the previous token, document creation time etc. The data used by their work is the translation to Portuguese of the TempEval-2 challenge data originally written in English [36]. Their work neither deal with noisy TEs nor with clinical data. The results of their method are not calculated using F1 score.

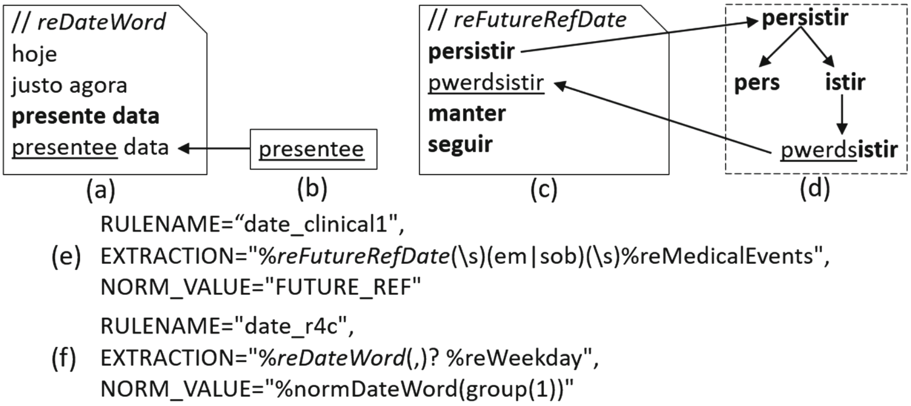
4 Extracting and Normalizing Temporal Expressions

In this section we present our method to extract and normalize TEs in clinical texts written in Brazilian Portuguese. Clinical texts are as noisy as the routine of the health professionals who write them. Examples of noisy TEs found in clinical texts are: (1) "*Revebe alta*" - the correct spelling is "*Recebe alta*" (receives discharge), (2) Different ways to present an hour and minute - "*05 h e 30 min*", "15:07" and "*4h10*". Thus, they need to pass through a preprocessing step.

In the preprocessing step, data is cleaned and prepared. In this step, training and test sets are changed to lowercase, while some typing errors are corrected. The correction is done in two steps: identification of errors and correction of all occurrences of the identified patterns (errors). Examples of the preprocessing are: "*ago/06*)" which was corrected to "*ago/06)*"; "*#retorno a*" that was corrected to "*# retorno a*"; "*3 dia(s)*" which was corrected to "*3 dia (s)*" and "*100 MAÇOS.ANO*" that was corrected to "*100 maços.ano*". These corrections are necessary once HeidelTime does not find a TE like "*3 dia(s)*" (because there is no white-space between the letter "a" and the left parenthesis). Each clinical text was turned into a single sentence because some TEs with two or more words were separated in two paragraphs, one word at the end and the other word at the beginning of different paragraphs (a line break problem).

The processing step starts by manually transforming each TE of the training set into a new rule and/or a new (pattern and/or normalization) resource in

HeidelTime. It is important to highlight that, we only modified the language resources (Portuguese) part of HeidelTime, the algorithmic part (Java code) was not changed. The processing step was done in three different incremental approaches, what supposedly makes the last approach better than the others. In all of them, the already existing resources in HeidelTime for Portuguese were kept and, if possible, appended with TE from the clinical texts. Figure 3 illustrates the process and also shows the main contributions of this paper, which are represented by parts (b) and (d) of the same figure. Bold words in (a) and (c) present different ways to add TEs in HeidelTime. Item (a) is a pattern resource (file) called "reDateWord" (italic). This resource already existed in the original HeidelTime (code taken from github (see footnote 1)). The pattern resource (file) "reFutureRefDate" (c) did not exist in the original HeidelTime, but was created by us. All bold parts of TEs in (c) came from the training set.



```
(e) RULENAME="date_clinical1",
    EXTRACTION="%reFutureRefDate(\s)(em|sob)(\s)%reMedicalEvents",
    NORM_VALUE="FUTURE_REF"
(f) RULENAME="date_r4c",
    EXTRACTION="%reDateWord(,)? %reWeekday",
    NORM_VALUE="%normDateWord(group(1))"
```

Fig. 3. Method processing step

Thus, the first approach called "Correct TEs" is the addition of TEs correctly written from the training set (in bold) in resources that already existed in HeidelTime (a) or were added to it (c). In the second approach called "Noisy TEs", the previous approach was kept, but noisy TEs found in the training set were also added to HeidelTime. It is represented by item (b), which was added to item (a), however, it could also have been added to a new file created by us.

In the third approach called "N-gram Noisy", the second approach is appended with the n-gram process. It is done by dividing in the middle each correctly written word of a TE with length greater than five characters (we used a changeable "n" which is equal to or greater than three characters), and each half (n-gram) was searched for in 870 clinical texts not used in the training and test sets. The aim of this approach is, to find misspelled words of TEs, in order to add the found patterns in the HeidelTime rules. The "N-gram Noisy" approach is represented in items (c) and (d). Item (d) illustrates the n-gram strategy, thus, the TE part "*persistir*" has a characters length equal to nine. The word was splitted in two parts "*pers*" and "*istir*". Afterwards, each part of the word (n-gram) was searched for in all 870 clinical texts. In item (d), the misspelled TE "*pwerdsistir*" represents a finding result of this process. Finally, the noisy TE is added to the item (c). This process was done for each correctly written word of TEs present in the pattern resource files of the "Correct TEs" approach. The "N-gram Noisy" approach is a type of string similarity strategy used to recognize misspelled words. It was adopted in this paper as an alternative way to cope with the noisy TEs in clinical texts written in Brazilian Portuguese, once a similar approach was already done based on the Edit Distance algorithm [32], a well-known algorithm for this purpose.

Items (e) and (f) are examples of rules. Item (e) is one example of rules created by us, its pattern resource "reFutureRefDate" is part of the extraction section of the rule, which also has another pattern resource called "reMedicalEvents" and is normalized with the imprecise and relative reference "FUTURE_REF". Item (f) already existed in the original HeidelTime, but had its pattern resource "reDateWord" appended with TEs from clinical texts.

The next section presents the experiments and results.

5 Experiments and Results

In this research we used 1,000 hospital discharge summaries from a Brazilian hospital produced between the years 2002 and 2007. Its use was approved by the Research Ethics Committee of PUCPR. From the total, 100 texts were used for training and 30 texts were used for testing. The hold-out method was applied. The other 870 texts were used with the n-gram approach already mentioned.

The training set and the test set were annotated by two annotators. A nurse assistant and a computer science master degree student. They marked TEs with TimeML TIMEX3 tags and normalized them with the TIMEX3 type and value attributes [18,24]. The annotators based their work on TimeML guidelines [8,18,24] and marked the TEs as noisy or correctly written. Their Kappa inter-annotator agreement coefficient was 75.2, which is a significant one considering the clinical domain challenges [31]. To evaluate the method, four experiments were done. They were evaluated with the precision, recall and F1 score metrics. They were tested with the same 30 clinical texts. All experiments used HeidelTime with a setup to the news domain. It was done to take advantage of the rules for Portuguese that already existed in the tagger and, because the discharge day date of all clinical texts was available (a prerequisite to use the news domain) [6,28].

The first experiment used HeidelTime with no changed rules or resources over the test set (code taken from github (see footnote 1)). In the second experiment we selected TEs correctly written from the training set and added them to HeidelTime, the "Correct TEs" approach. These TEs were added to already existing resources and rules, or added to new resources or rules designed specifically for the clinical texts written in Brazilian Portuguese.

In the third experiment, the previous setup was kept (experiments one and two). However, we selected also the noisy TEs of the training set and added them to resources or rules, the "Noisy TEs" approach described in Sect. 4. In the fourth experiment (N-gram Noisy approach), HeidelTime was kept with the third experiment setup, but rules and resources received also noisy TEs found in the process done with the n-gram strategy, described in Sect. 4. Table 1 presents the results of each experiment of the temporal tagging task.

Table 1. Experiments results

Step	Metric	Experiments			
		Original (1)	Correct (2)	Noisy (3)	N-Gram (4)
Extraction	Precision	79.78	**95.58**	93.18	94.52
	Recall	17.27	68.37	76.40	**83.94**
	F1 score	28.40	79.72	83.96	**88.92**
Normalization	Precision	77.53	91.50	92.58	**93.42**
	Recall	16.79	65.45	75.91	**82.97**
	F1 score	27.60	76.31	83.42	**87.89**

The fourth experiment had the best results, except by the precision of the second experiment. It incremented resources from the three previous experiments, plus the n-gram strategy. An interesting behavior was observed in the experiments, it is the increase of the recall between experiments 2 and 3. It happened because the patterns added to HeidelTime based on the noisy TEs, improved the ability of the tagger to make its work, which is identify, extract and normalize TEs. However, as results between the experiments two, three and four are similar, statistical relevance tests were done. Nevertheless, the results of the experiment one (original HeidelTime) were kept out of this statistical tests, because its results were far smaller than the others.

All experiments were done under the same conditions (paired samples), however, the Shapiro-Wilk test revealed that the distribution of the test set samples was not normal. Thus, a non-parametric test was applied. As there were more than two experiments to test, the Friedman test was chosen to check if there was a relevant statistical difference between the results of the three experiments. The pair comparison is done in a Dunn-Bonferroni post-hoc test showed that there was a relevant difference only between the second experiment and the fourth experiment. In order to test each of the three experiments by pairs, the Wilcoxon test was also performed and, the result was the same. In both cases, a confidence of 0.95 and a significance of 0.05 was used. The same statistical tests were done to the extraction and normalization steps and, the results were the same for both.

Referring to the normalization type of all TEs of the corpus (training and test), 43.98% were "date", 21.58% were "duration", 20.49% were "time" and 13.93% were "set" (frequency). Yet, 28.14% were normalized as imprecise [32] TEs. Among the imprecise ones, 91.15% were PAST_REF and only 8.84% were

FUTURE_REF. It happened because discharge summaries refer more about what happened to the patient up to the discharge day (past and present) and rarely refer to things that will happen after the discharge day (future).

The amount of 28 noisy TEs were found by the n-gram strategy. Examples of them are presented in their noisy version and correct version respectively: "*aanterior*" (*anterior*), "*seginte*" (*seguinte*), "*rcebe*" (*recebe*), "*acompanham,ento*" (*acompanhamento*), "*históri*" (*história*) and "*pó-infarto*" (*pós-infarto*). Only correctly written words of TEs served as input of the n-gram strategy, thus, there are probably more noisy TEs in the 870 clinical texts not identified. Examples of noisy TEs missed by HeidelTime in the test set are: "*2m*" (*2 meses*), "*1a5m*" (*1 ano e 5 meses*) e "*pós-op*" (*pós-operatório*). They were missed because they did not exist in the training set and thus, were not added to HeidelTime.

The TE "02/07" is an example of a problematic one. According to the gold standard it was the second day of July, however, HeidelTime understood it as February/2007, it highlights a regular expressions limitation, because they do not consider the surrounding words of a TE, what could be decisive in this case. Finally, about the annotation process, from a total of 1586 TEs annotated in the training and test sets, 7.69% were considered noisy by the annotators.

6 Conclusions

The focus of this work is to extract and normalize TEs from correct and incorrect clinical texts written in Brazilian Portuguese. The tagger we used showed that the n-gram strategy reached a statistical relevant improvement when compared to the tagging of only correctly written TEs in the clinical domain. The experiment results showed that our best result is similar to other works done in other languages, especially within the clinical domain [9,10,17], considering that none of them cope with noisy TEs in their extraction and normalization steps.

As future work, a combination of HeidelTime with a machine learning approach might improve the quality of the temporal tagging of correctly written and noisy TEs. For this, we intent to combine HeidelTime with classifiers such as SVM, CRF, Random Forest or Deep Learning.

References

1. Alonso, O., Strötgen, J., Baeza-Yates, R.A., Gertz, M.: Temporal information retrieval: challenges and opportunities. In: TWAW, vol. 11, pp. 1–8 (2011)
2. Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J., Verhagen, M.: SemEval-2015 task 6: clinical TempEval. In: SemEval@NAACL-HLT, pp. 806–814 (2015)
3. Campos, R., Dias, G., Jorge, A.M., Jatowt, A.: Survey of temporal information retrieval and related applications. ACM Comput. Surv. (CSUR) **47**(2), 15 (2015)
4. Chang, A.X., Manning, C.D.: SUTime: a library for recognizing and normalizing time expressions. In: LREC, vol. 2012, pp. 3735–3740 (2012)

5. Costa, F., Branco, A.: Extracting temporal information from portuguese texts. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS, vol. 7243, pp. 99–105. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28885-2_11
6. Costa, F., Branco, A.: TimeBankPT: a TimeML annotated corpus of Portuguese. In: LREC, pp. 3727–3734 (2012)
7. Gertz, M., Strötgen, J., Zell, J.: HeidelTime: tuning English and developing Spanish resources for TempEval-3, Atlanta, Georgia, USA, p. 15 (2013)
8. TimeML Working Group, et al.: Guidelines for temporal expression annotation for English for TempEval 2010 (2009)
9. Gupta, N., Joshi, A., Bhattacharyya, P.: A temporal expression recognition system for medical documents by taking help of news domain corpora. In: 12th International Conference on Natural Language Processing, ICON (2015)
10. Hamon, T., Grabar, N.: Tuning HeidelTime for identifying time expressions in clinical texts in English and French. In: EACL 2014, pp. 101–105 (2014)
11. Kreimeyer, K., et al.: Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J. Biomed. Inform. **73**, 14–29 (2017)
12. Lee, K., Artzi, Y., Dodge, J., Zettlemoyer, L.: Context-dependent semantic parsing for time expressions. In: ACL, vol. 1, pp. 1437–1447 (2014)
13. Li, H., Strötgen, J., Zell, J., Gertz, M.: Chinese temporal tagging with HeidelTime. In: EACL, vol. 2014, pp. 133–137 (2014)
14. Madkour, M., Benhaddou, D., Tao, C.: Temporal data representation, normalization, extraction, and reasoning: a review from clinical domain. Comput. Methods Progr. Biomed. **128**, 52–68 (2016)
15. Manfredi, G., Strötgen, J., Zell, J., Gertz, M.: HeidelTime at EVENTI: tuning Italian resources and addressing TimeML's empty tags. In: Proceedings of the Forth International Workshop EVALITA, pp. 39–43 (2014)
16. Meng, Y., Rumshisky, A., Romanov, A.: Temporal information extraction for question answering using syntactic dependencies in an LSTM-based architecture. arXiv preprint arXiv:1703.05851 (2017)
17. Moharasar, G., Ho, T.B.: A semi-supervised approach for temporal information extraction from clinical text. In: 2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future, RIVF, pp. 7–12. IEEE (2016)
18. Pustejovsky, J., et al.: TimeML: robust specification of event and temporal expressions in text. New Dir. Quest. Answ. **3**, 28–34 (2003)
19. Pustejovsky, J., Knippen, R., Littman, J., Saurí, R.: Temporal and event information in natural language text. Lang. Resour. Eval. **39**(2), 123–164 (2005)
20. Quaresma, P., Mendes, A., Hendrickx, I., Gonçalves, T.: Tagging and labelling Portuguese modal verbs. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.G. (eds.) PROPOR 2014. LNCS, vol. 8775, pp. 70–81. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09761-9_7
21. Roberts, K., Rink, B., Harabagiu, S.M.: A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. J. Am. Med. Inform. Assoc. **20**(5), 867–875 (2013)
22. Rodrigues, R., Gomes, P.: Improving question-answering for portuguese using triples extracted from corpora. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS, vol. 9727, pp. 25–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_3

23. Sarath, P., Manikandan, R., Niwa, Y.: Hitachi at SemEval-2017 Task 12: system for temporal information extraction from clinical notes. In: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval 2017, pp. 1005–1009 (2017)
24. Saurı, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., Pustejovsky, J.: TimeML annotation guidelines version 1.2. 1 (2006)
25. Schilder, F.: Extracting meaning from temporal nouns and temporal prepositions. ACM Trans. Asian Lang. Inf. Process. (TALIP) **3**(1), 33–50 (2004)
26. Skukan, L., Glavaš, G., Šnajder, J.: HEIDELTIME.HR: extracting and normalizing temporal expressions in Croatian. In: Proceedings of the 9th Slovenian Language Technologies Conferences, IS-LT 2014, pp. 99–103 (2014)
27. Strötgen, J., Gertz, M.: Multilingual and cross-domain temporal tagging. Lang. Resour. Eval. **47**(2), 269–298 (2013)
28. Strötgen, J., Gertz, M.: A baseline temporal tagger for all languages. In: EMNLP, vol. 15, pp. 541–547 (2015)
29. Strötgen, J., Gertz, M.: Domain-sensitive temporal tagging. Synth. Lect. Hum. Lang. Technol. **9**(3), 1–151 (2016)
30. Sun, W., Rumshisky, A., Uzuner, O.: Annotating temporal information in clinical narratives. J. Biomed. Inform. **46**, S5–S12 (2013)
31. Tissot, H., Roberts, A., Derczynski, L., Gorrell, G., Del Fabro, M.D.: Analysis of temporal expressions annotated in clinical notes. In: Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, ISA 2011 (2015)
32. Tissot, H.C.: Normalisation of imprecise temporal expressions extracted from text (2016)
33. UzZaman, N., Allen, J.F.: Event and temporal expression extraction from raw text: first step towards a temporally aware system. Int. J. Semant. Comput. **4**(04), 487–508 (2010)
34. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J.: SemEval-2007 Task 15: TempEval temporal relation identification. In: Proceedings of the 4th International Workshop on Semantic Evaluations, pp. 75–80. Association for Computational Linguistics (2007)
35. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., Pustejovsky, J.: The TempEval challenge: identifying temporal relations in text. Lang. Resour. Eval. **43**(2), 161–179 (2009)
36. Verhagen, M., Sauri, R., Caselli, T., Pustejovsky, J.: SemEval-2010 Task 13: TempEval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 57–62. Association for Computational Linguistics (2010)

Syntactic Knowledge for Natural Language Inference in Portuguese

Erick Fonseca(✉) and Sandra M. Aluísio

Institute of Mathematics and Computer Science, University of São Paulo,
Av. Trabalhador São-carlense, 400, São Carlos, SP, Brazil
erickrfonseca@gmail.com, sandra@icmc.usp.br

Abstract. Natural Language Inference (NLI) is the task of detecting relations such as entailment, contradiction and paraphrase in pairs of sentences. With the recent release of the ASSIN corpus, NLI in Portuguese is now getting more attention. However, published results on ASSIN have not explored syntactic structure, neither combined word embedding metrics with other types of features. In this work, we sought to remedy this gap, proposing a new model for NLI that achieves 0.72 F_1 score on ASSIN, setting a new state of the art. Our feature analysis shows that word embeddings and syntactic knowledge are both important to achieve such results.

Keywords: Natural Language Inference
Recognizing Textual Entailment · Feature engineering · Syntax

1 Introduction

Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), is the NLP task of determining whether a hypothesis H can be inferred from a premise P [7] (usually, P and H are sentences). Other semantic relations are also possible, such as contradiction [4,13] and paraphrase [10].

Datasets for NLI on English exist since 2005, with the RTE Challenges [6], and later with the SICK [13] and SNLI [4] corpora. For Portuguese, only recently the ASSIN [10] corpus was released, containing 10,000 sentence pairs annotated for NLI (entailment, paraphrase and neutral) and for semantic similarity [1].

Still, published results on the ASSIN dataset are worse than a word overlap baseline or only slightly better than it [2,3,8,9]. We hypothesize this is because these models focused on lexical overlap and similarity, without any attention to syntactic structure. On top of that, lexical similarity methods could be improved with the use of word embeddings, which have already been shown to be very effective in the semantic similarity task [11].

In this work, we sought to remedy this limitation by exploring richer representations for the input pairs. We extracted features including syntactic knowledge and embedding-based similarity, besides well established ones dealing with word alignments. Our model, named Infernal (*INFERence in NAural Language*),

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 242–252, 2018.
https://doi.org/10.1007/978-3-319-99722-3_25

achieved new state-of-the-art results, with 0.72 macro F_1 score on the complete ASSIN dataset (i.e., both Brazilian and European variants).

Moreover, we analyzed a sample of misclassified pairs in order to understand the difficulties of the task. We found that most of them pose significant difficulties, and that richer NLP resources (such as repositories of equivalent phrases) are necessary to improve performance on NLI and related tasks.

This paper is organized as follows. We first summarize the ASSIN dataset in Sect. 2, and briefly discuss previous related work on it in Sect. 3. We present our model in Sect. 4, and our experiments and findings in Sect. 5. We bring our conclusions in Sect. 6.

2 ASSIN Dataset

ASSIN [10] has 10,000 sentence pairs annotated for NLI (with entailment, paraphrase and neutral labels), half in Brazilian Portuguese (PT-BR) and half in European Portuguese (PT-PT)[1]. It is an unbalanced dataset: the neutral relation has 7,316 pairs, entailment has 2,080 while the paraphrases are a small portion of the set (604 pairs). Either language variant has 2,500 pairs for training, 500 for validation and 2,000 for testing, the three of them with the same proportions among classes.

Its sentences come from news articles, making the corpus more complex and with more varied topics than SNLI or SICK, which were produced from image captions. Thus, ASSIN presents challenges such as world knowledge, idiomatic expressions and named entities. Its difficulty is reflected in the fact that three out of four participants in the ASSIN shared task had performance below a word overlap baseline (a logistic regression classifier trained with the ratio of words in P that appear in H and the ration of words in H that appear in P).

3 Related Work

While works for NLI in English currently take advantage of large scale datasets to train deep neural networks [5,18], the lack of such a corpus for Portuguese has limited NLI strategies to shallow models dependent on feature engineering. Thus, we restrict our revision to published work with the ASSIN dataset.

The current state-of-the-art in ASSIN for PT-BR was achieved by Fialho et al. [9]. They consider different views of the input sentences: the original words, a lowercase version, stemmed words, among others. From each view, they extract metrics like string edit distance, word overlap, BLEU, ROUGE, and others. In total, 96 features are fed to an SVM, achieving 0.71 F_1 for PT-BR and 0.66 for PT-PT. Feitosa and Pinheiro [8] tried to improve on these results, adding eight new wordnet-based features to capture lexical similarity. However, they did not gain any significant improvement.

[1] Available at nilc.icmc.usp.br/assin/.

Rocha and Cardoso [17] reached the state of the art for PT-PT. They used a relatively small set of features, including counts of overlapping or semantically related tokens (such as synonyms and meronyms), named entities, word embedding similarity and whether both sentences have the same verb tense and voice. While the last one employs some syntactic knowledge, it is rather limited, and it is not clear how they deal with sentences with more than one verb, which are common in ASSIN. They only present results for PT-PT, with their best setup having 0.73 F_1. Curiously, while Fialho et al. [9] reports better results when combining PT-BR and PT-PT training data, Rocha and Cardoso [17] had a slight performance decrease when they did the same.

Reciclagem [2] did not use any machine learning technique; instead, it relied solely on lexical similarity metrics extracted from various resources. ASAPP, from the same authors, improved on this base by using an automatic classifier and included features such as counts of tokens, nominal clauses and named entities.

The Blue Man Group [3] extracted many word embedding-based similarity features from the pairs. They compare words of one sentence with the other, grouping similarity values in histograms, which are then fed to an SVM classifier. They also report negative results with deep neural networks, although they do not mention any performance value.

4 Data Modeling

4.1 Pre-processing

We perform some steps in an NLP pipeline in order to extract features. We run a syntactic parser, a named entity detector (NER), lemmatize words and find lexical alignments.

We used the Stanford CoreNLP dependency parser [12] trained on the Brazilian Portuguese corpus of the Universal Dependencies (UD) project[2], version 1.3. We used the spaCy pre-trained NER model for Portuguese[3], version 2.0. SpaCy also has a pre-trained syntactic parser for Portuguese, but we found that its performance is worse than the CoreNLP model.

For lemmatization, we checked the Unitex DELAF-PB dictionary[4] with POS tags produced by CoreNLP. The DELAF-PB is a Brazilian Portuguese resource; however, word forms in European Portuguese orthography can easily be checked against it after replacing some characters. If a word is not found in the dictionary, it is searched again after replacing some consonant clusters (ct and pt for t, cç and pç for ç, and mn for n).

Once we have word lemmas, we align words in the two sentences if they have the same lemma or share a synset in OpenWordNet-PT [14]. Using the same

[2] More information at http://universaldependencies.org.

[3] More information at http://spacy.io.

[4] The DELAF-PB dictionary maps inflected word forms to lemmas, according to their part-of-speech tag. More information at http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html.

resource, we also align verbs with nominalizations (such as *correr* and *corrida*). Named entities are aligned if they are exactly the same or if one is an acronym composed of the initial letters of the words of the other one.

4.2 Feature Extraction

Once we have preprocessed sentence pairs, we can extract features from them. We also depended on two other resources to compute features: a stopword list and a word embedding model. The former is the one available in NLTK[5], expanded with the punctuation signs and the words *é*, *ser* and *estar*, which were lacking from it. The embedding model was trained with Glove [16] over a collection of news texts, literary books and Wikipedia, with over 500 million tokens.

We also used the concept of Tree Edit Distance (TED) in some features, which measures how different two trees are from each other and has already been successfully used for NLI [19]. The idea of TED is to apply a sequence of edit operations to a tree that transforms it into another one. The possible operations are the insertion of a node, removal of a node, or replacement of a node for another. The cost for each operation must be defined individually *a priori* (possibly, costs may depend on the involved words or dependency labels). Given two trees and the cost of each possible operation, the minimal TED can be computed in polynomial time. We used the Zhang-Shasha algorithm [20] in our implementation.

The complete list of features is described as follows. Note that, while we list 14 features, some of them can be computed when aligning P to H or vice-versa, and others can be normalized by the length of either sentence. In total, we extract 28 feature values.

1. **BLEU.** (BiLingual Evaluation Understudy) is a common metric in Machine Translation. It computes how many n-grams with $1 \leq n \leq 4$ from one sentence appears in the other. It has two values: using P as reference (denoted here as $P \rightarrow H$) and using H ($H \rightarrow P$).
2. **Dependency overlap.** The proportion of overlapping dependency tuples. A dependency tuple is composed by the dependency label, parent node and child node; two tuples are considered as overlapping when they have the same label and aligned parent and child nodes. Additionally, the passive subject label (*nsubjpass*) is considered equivalent to the direct object label (*dobj*). This feature has two values: the ratio of overlapping tuples with respect to the length of P and to the length of H.
3. **Nominalization.** This feature checks whether one sentence has a verb aligned with a nominalized form used as direct object. It has two values, depending on which sentence has the verb and which has the nominalization.
4. **Length ratio.** Length ratio between the number of tokens in P and H, excluding *stopwords*.

[5] More information at http://www.nltk.org.

5. **Verb arguments.** This feature has two values that check whether verbs in the two sentences also have the aligned subject and direct object. If the objects differ, it has value $(0, 0)$; if they are aligned, it is $(1, 1)$. If only H has an object, it has value $(0, 1)$, if only in P, $(1, 0)$. The values are denoted, respectively, by *verb arguments* $P \rightarrow H$ and *verb arguments* $H \rightarrow P$.
6. **Negation.** This feature checks if an aligned verb is negated in one of the sentences. This happens when the verb has a modifier with the label *neg*.
7. **Quantities.** This feature has two values that check for quantities describing aligned words (indicated by the dependency label *num*). The value of the modifier is computed both for digits and fully written forms. The first feature value is 1 if any two words have the same quantity, 0 otherwise; the second one is 1 if there is a quantity mismatch. In case of no aligned words with quantity modifier, it is $(0, 0)$.
8. **Sentence cosine.** The cosine similarity between the two sentence vectors. Vectors are obtained as the elementwise average of all token vectors.
9. **Simple TED.** The TED between the two sentences, considering insert, removal and update costs as 1. Two nodes are matched when they have the same lemma and dependency label. This feature has three values: the TED value itself, TED divided by the length of P and by the length of H.
10. **TED with cosine distance.** The TED like the one above, except that update costs are equal to the cosine distance between embeddings.
11. **Word overlap.** The ratio of words in each sentence for which there is another word with the same lemma in the other sentence, excluding stopwords. The ratio to the length of P is denoted *overlap* P, while the ratio to H is *overlap* H.
12. **Synonym overlap.** Like the one above, but considering any aligned word, not only with the same lemma.
13. **Soft overlap.** This feature measures word embedding similarity instead of a binary match. For each word in a sentence, except stopwords, we take its highest cosine similarity with words from the other sentence, then we average all similarities. It has one value for each sentence.
14. **Named entities.** This feature checks for the presence of named entities, and has three binary values. The first indicates whether there is named entity in P without an equivalent in H; the second one indicates the opposite; and the third one indicates the presence of an aligned pair. All combinations of values are possible, depending on the number of named entities in the pair.

Our features contemplate different levels of knowledge: simple count statistics (1, 4, 11), resource-based lexical semantics (3, 12, 14), syntax (2, 3, 5, 6, 7, 9, 10) and embedding-based semantics (8, 10, 13). To the best of our knowledge, features 9, 10 and 13 have not been used before for NLI. Our implementation is available at https://github.com/erickrf/infernal.

5 Experiments

We trained different classifiers in our experiments, using the scikit-learn library [15]: Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF) and Gradient Boosting (GB).

We combined PT-BR and PT-PT training data, like in the best results reported by [9]. Before training classifiers, we normalize feature values: given a training data matrix $X \in \mathbb{R}^{n \times d}$, with n training examples and d features (28 with our full set), we normalize each column to have mean 0 and variance 1.

We did a 10-fold cross-validation in the training set in order to select the most relevant hyperparameters for some algorithms. For SVM, we used a penalty c of value 10, and an RBF kernel with γ coefficient 0.01. For RF, we used 500 trees which could use up to 6 features each, and expandable to the maximum. For GB, we used 500 trees with a maximum depth of 3 and learning rate η of 0.01. All other hyperparameters had default values of scikit-learn version 0.18.

Additionally, for LR and SVM, it is also possible to weight training examples to the inverse proportion of their class (in order to give more importance to paraphrase and entailment examples), and we also experimented that. Table 1 shows the results of our classifiers, as well as the previous state-of-the-art and the word overlap baseline.

Table 1. Infernal performance on ASSIN. The F1 values are the macro F1 (mean for all classes). The bottom part of the table shows previous state-of-the-art results and the word overlap baseline. RC refers to Rocha and Cardoso [17]

Model	Validation		PT-BR		PT-PT		All	
	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1
RL	85.50%	0.72	**87.30%**	**0.71**	85.75%	0.72	**86.52%**	**0.72**
RL, weighted	85.20%	**0.74**	85.00%	0.69	84.60%	**0.74**	84.80%	**0.72**
Random Forest	85.20%	0.72	86.20%	0.67	**86.20%**	**0.74**	86.20%	0.71
GB	**85.80%**	0.73	86.35%	0.67	86.10%	**0.74**	86.22%	0.71
SVM	85.60%	0.73	86.90%	0.70	85.75%	0.73	86.33%	**0.72**
SVM, weighted	80.20%	0.69	79.20%	0.64	80.95%	0.71	80.08%	0.68
L2F/INESC-ID [9]	—	—	85.85%	0.66	84.90%	0.71	—	—
RC [17]	—	—	—	—	83.5%	0.73	—	—
Baseline	81.40%	0.69	82.80%	0.64	81.75%	0.7	82.27%	0.67

Almost our models achieved higher accuracy and F_1 than the previous state of the art, showing that our features provide a good representation of the data for this problem. This is more evident when we consider that we used 28 features, while [9] used 96. No single algorithm stood out as best, but Logistic Regression seems interesting for coupling good performance with low computational cost and low sensibility to hyperparameters.

5.1 Feature Analysis

We also analyzed the relative importance of our features. Determining the exact importance of each one in a multidimensional setting where there may be some interdependence is impossible, but we can get reasonable estimates from methods like Random Forest and Gradient Boosting. These methods, which are ensembles of decision trees, can score feature importance based on how well they split the data in different classes.

Thus, we trained 10 instances of RF and GB, varying the random seed, and averaged the relative importance of each one, in order to get a more stable estimate. The importance of the features can be seen in Table 2, ordered according to the average importance for the two algorithms.

As expected, features related to word overlap have bigger weight, evidenced by the good performance of the word overlap baseline. Among them, our newly

Table 2. Features importance. The first and fifth column show the relative ordering of each feature; **%GB** and **%RF** indicate the percentual importance of each feature for each algorithm.

#	Feature	%GB	%RF	#	Feature	%GB	%RF
1	Soft overlap H	12.68%	14.06%	15	TED/length H	1.94%	3.23%
2	Overlap H	11.30%	13.74%	16	BLEU P → H	2.08%	3.07%
3	Synonym overlap H	4.85%	8.73%	17	TED cosine	1.97%	2.95%
4	Soft overlap P	7.56%	4.93%	18	Quantity mismatch	3.97%	0.84%
5	Cosine	6.64%	5.36%	19	TED	2.25%	1.93%
6	Overlap P	6.85%	4.89%	20	Quantity match	2.07%	0.71%
7	Length ratio	6.01%	5.13%	21	Non-aligned NE H	1.96%	0.38%
8	TED cosine/length H	6.63%	3.92%	22	Non-aligned NE P	1.16%	0.44%
9	TED/length P	3.50%	4.90%	23	Verb arguments P → H	0.25%	0.52%
10	Dependency overlap H	2.52%	5.51%	24	Verb arguments H → P	0.25%	0.52%
11	TED cosine/length P	3.06%	4.13%	25	Nominalization in P	0.63%	0.11%
12	Synonym overlap P	3.75%	3.14%	26	Negated verb	0.51%	0.14%
13	Dependency overlap P	2.81%	3.06%	27	Aligned NE	0.09%	0.49%
14	BLEU H → P	2.32%	3.02%	28	Nominalization in H	0.39%	0.13%

proposed soft overlap is one of the most important ones as well as the sentence cosine, showing that the flexible nature of word embeddings can be very useful for NLI.

In the middle positions, we see features related to syntactic structure: dependency overlap, matching quantities and TED. While less informative than lexical overlap, they still bring substantial information, which suggests they were responsible for the good performance of our models beyond lexical similarity.

As the least useful features, we have nominalizations, named entities, negation and verb structure features. While somewhat informative, we found that negated verbs and nominalizations are relatively rare in ASSIN, limiting their impact. The verb structure feature is too specific to be discriminative as well. We conjecture that named entity features had lower usefulness because the same entity may often be described in different ways—such as an omitted first or last name. Retraining our models without the least informative features resulted in a slight performance drop, indicating that they are still good to have.

5.2 Error Analysis

We manually analyzed 65 wrongly classified pairs by our LR model and listed the linguistic phenomena that led to the mistakes. The listing is shown in Table 3 and described as follows.

Table 3. Main sources of errors

Phenomena	Occurrences
Too much overlap	23
Rewrite	21
Contextual synonyms	19
Quantity specifier	5
Qualified named entity	4

Too much overlap is the main cause of neutral pairs misclassified as entailment or paraphrase. Example: *A presidente Dilma Rousseff empossa, nesta segunda-feira (5), os novos ministros, em cerimônia no Palácio do Planalto/Dez ministros tomaram posse nesta segunda-feira (5) numa cerimônia no Palácio do Planalto.*

Rewrite is the opposite, when the same content is described with different words or implicitly. This causes entailment and paraphrases to be classified as neutral. Example: *Os trabalhadores protestam contra a regulamentação da terceirização, a retirada de direitos trabalhistas e o ajuste fiscal/Os trabalhadores protestam contra o projeto de lei que regulamenta a terceirização no país.*

Contextual synonyms are words which have the same meaning only in very specific contexts, and thus not expected to be found in wordnets. Example: *Os demais agentes públicos serão* ***alocados*** *na classe econômica/Todo o resto dos funcionários públicos terá que* ***embarcar*** *na classe econômica.*

Quantity specifiers are expressions that specify that two quantities may differ and still keep an entailment relation, such as *at least*, *approximately*, etc. Example: *De acordo com a polícia, 56 agentes e 12 manifestantes ficaram feridos/Pelo menos 46 policiais e sete manifestantes ficaram feridos.*

Qualified named entity are named entities appearing in one sentence with a more detailed description, such as a title or profession (*actor*, *president*, etc.). Since this description is subsumed by the entity itself, it should not affect an entailment decision. Example: *Tite, no segundo tempo, trocou Ralf por Mendoza/O atacante Mendoza entrou no lugar do volante Ralf.*

These issues are hard to solve. For quantification, a list of expressions indicating approximate quantities can solve some cases. Concerning rewritten passages, resources containing equivalent expressions and phrases are also useful, although limited in the generalization capacity.

At any rate, a larger NLI corpus would be useful, allowing models to learn more subtleties from language and depend less on word overlap. Also, more data would make more feasible the efficient training of neural models, which have been successful in larger English corpora.

6 Conclusion

We have presented a new feature set for the NLI task on the ASSIN corpus, shown that it sets a new state-of-the-art with different classifiers, and performed a careful analysis of feature importance and sources of error.

The features we proposed encode syntactic knowledge about the pairs, something that, to the best of our knowledge, was missing in all published results on ASSIN to date. Also, we proposed a more flexible lexical similarity measure, the soft overlap, which is a strong indicator for NLI. Our feature set has been shown to be very useful for this task, and might be useful as well for other related tasks involving the semantics of two sentences.

Moreover, we pointed out the current challenges that ASSIN poses to NLI systems. Once we have efficient means to overcome them, even better performances can be expected.

Acknowledgments. This work was supported by FAPESP grant 2013/22973-0.

References

1. Agirre, E., et al.: SemEval-2015 Task 2: semantic textual similarity, English, Spanish and pilot on interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval 2015, pp. 252–263. Association for Computational Linguistics (2015)
2. Alves, A.O., Oliveira, H.G., Rodrigues, R.: ASAPP e Reciclagem no ASSIN: Alinhamento Semântico Automático de Palavras aplicado ao Português. Linguamática **8**(2), 43–58 (2016)
3. Barbosa, L., Cavalin, P., Martins, B., Guimarães, V., Kormaksson, M.: Blue Man Group no ASSIN: Usando Representações Distribuídas para Similaridade Semântica e Inferência Textual. Linguamática **8**(2), 15–22 (2016)
4. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 632–642. The Association for Computational Linguistics (2015)
5. Chen, Q., Zhu, X., Ling, Z.H., Wei, S., Jiang, H., Inkpen, D.: Enhanced LSTM for natural language inference. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Long Papers, vol. 1, pp. 1657–1668. Association for Computational Linguistics (2017)
6. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognising textual entailment challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS, vol. 3944, pp. 177–190. Springer, Heidelberg (2006). https://doi.org/10.1007/11736790_9
7. Dagan, I., Roth, D., Sammons, M., Zanzotto, F.M.: Recognizing Textual Entailment: Models and Applications. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, San Rafael (2013)
8. Feitosa, D.B., Pinheiro, V.C.: Análise de medidas de similaridade semântica na tarefa de reconhecimento de implicação textual. In: Proceedings of Symposium in Information and Human Language Technology (2017)
9. Fialho, P., Marques, R., Martins, B., Coheur, L., Quaresma, P.: INESC-ID no ASSIN: measuring semantic similarity and recognizing textual entailment. Linguamática **8**(2), 33–42 (2016)
10. Fonseca, E.R., dos Santos, L.B., Criscuolo, M., Aluísio, S.M.: Visão Geral da Avaliação de Similaridade Semântica e Inferência Textual. Linguamática **8**(2), 3–13 (2016)
11. Hartmann, N.S.: Solo queue no ASSIN: mix of a traditional and an emerging approaches. Linguamática **8**(2), 59–64 (2016)
12. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60 (2014)
13. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A SICK cure for the evaluation of compositional distributional semantic models. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation, pp. 216–223. European Language Resources Association (ELRA) (2014)
14. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: an open Brazilian WordNet for reasoning. In: Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012, pp. 353–360 (2012)

15. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
16. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Empirical Methods in Natural Language Processing, EMNLP, pp. 1532–1543 (2014). http://www.aclweb.org/anthology/D14-1162
17. Rocha, G., Cardoso, L.H.: Recognizing textual entailment: challenges in the Portuguese language. Information **9**(4), 76 (2018)
18. Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C.: DiSAN: directional self-attention network for RNN/CNN-free language understanding. ArXiv e-prints (2017)
19. Zanoli, R., Colombo, S.: A transformation-driven approach for recognizing textual entailment. Nat. Lang. Eng. **23**(4), 507–534 (2016)
20. Zhang, K., Shasha, D.: Simple fast algorithms for the editing distance between trees and related problems. SIAM J. Comput. **18**, 1245–1262 (1989)

Language Resources

RulingBR: A Summarization Dataset for Legal Texts

Diego de Vargas Feijó(✉) and Viviane Pereira Moreira

Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
{dvfeijo,viviane}@inf.ufrgs.br
http://www.inf.ufrgs.br/

Abstract. Text summarization consists in generating a shorter version of an input document, which captures its main ideas. Despite the recent developments in this area, most of the existing techniques have been tested mostly in English and Chinese, due in part to the low availability of datasets in other languages. In addition, experiments have been run mostly on collections of news articles, which could lead to some bias in the research. In this paper, we address both these limitations by creating a dataset for the summarization of legal texts in Portuguese. The dataset, called RulingBR, contains about 10K rulings from the Brazilian Federal Supreme Court. We describe how the dataset was assembled and we also report on the results of standard summarization methods which may serve as a baseline for future works.

Keywords: Summarization · Dataset · Legal · Law

1 Introduction

Text summarization is an important task in Natural Language Processing. It consists in generating a shorter version of the text given as input, capturing its main ideas. In the last few years, summarization has undergone significant developments. Notably, many of the new techniques being applied rely on deep learning strategies to go beyond the previously established state-of-the-art results [19,22,25,26]. Despite the recent boom in this area, the majority of works have been using English and Chinese datasets due in part to the low availability of resources in other languages.

Another limitation of the current research is that it focuses on news articles, for which the task consists in generating the headline or a very short summary. For example, models trained on the DUC-2004 task can only generate summaries of up to 75 characters [14,19], and the input consists of only one or two sentences.

News articles usually begin with a *teaser* sentence used as a catch for the reader, which sums up the contents of the full article. So, the task of guessing the title can generally obtain good results by simply extracting the first few words of the article. The excessive focus on this type of text introduces bias in the techniques being developed. For example, Google's Textsum model [24] for

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 255–264, 2018.
https://doi.org/10.1007/978-3-319-99722-3_26

summarization uses just the first two paragraphs of the article. Another possible approach is to weight the sentences in descending order from the start, in favor of the first few sentences [23].

We believe there is a need for datasets with different text styles and longer summaries with contents taken from several parts of the input. This would allow a more realistic setting and potential for employing summarization in a wider set of applications.

In this paper, we report on the creation of RulingBR− a dataset for the summarization of legal texts in Portuguese containing over 10K decisions from the Brazilian Federal Supreme Court. Our contribution aims at addressing two limitations of the research in summarization: (*i*) the low availability of resources for languages other than English and Chinese, and (*ii*) the excessive focus on summarizing news articles. We have assembled a language resource in Portuguese to enable the development of methods for this language. The second contribution is to do with the style of the texts, which will contribute to a greater variety in the research on summarization.

2 Related Work

There are a few datasets that have been used for evaluating summarization techniques on generic domains. The following are available in Brazilian Portuguese.

TeMário [17] is composed of 100 news articles. Each text contains a pair of reference summaries: one was made by a human an the other was automatically generated.

Summ-it [6] is an annotated corpus that contains anaphoric coreferences. These are newspaper articles annotated from the Brazilian *Folha de São Paulo* newspaper.

CSTNews [1] is another annotated corpus. It is composed of 50 text collections and each collection has about four documents. It uses texts from the following Brazilian news sources *Folha de São Paulo*, *Estadão*, *O Globo*, *Jornal do Brasil*, and *Gazeta do Povo*.

The most widely used datasets are available in English and are described below.

The Annotated Gigaword [18] is the largest static corpus of English news documents available [15]. It contains over 10 million documents from seven news sources, annotated with syntactic and discourse structure. It was not specifically built to be a summarization dataset, but it has been used for this purpose by simulating that the headline would be a summary of the article.

CNN/Daily Mail was purposely designed for summarization as each article comes paired with a short set of summarized bullet points that represent the highlights of the text. It is frequently used for question answering [5] and is composed of about 300 thousand articles.

Opinosis [8] contains customer reviews about a product they bought. Each product description has five reviews. This is a small dataset containing only 51 articles.

DUC[1] stands for Document Understanding Conference. It has run a specific summarization track since 2001. In 2008, DUC became a summarization track inside the Text Analysis Conferences (TAC). These datasets contain human-produced per-document and multiple document summaries.

RulingBR differs from these related datasets because in the legal domain, documents are generally lengthier and their structure is very different from the structure of news articles. As a consequence, the assumption that the most important ideas will be in the first few sentences is not valid.

3 A Summarization Dataset with Legal Documents

For the purpose of text summarization in the legal domain, we searched for a source with a large number of publicly available documents. Thus, we chose to use the *Supremo Tribunal Federal* (STF) as our source. The STF is the highest court in Brazil and has the final word interpreting the country's Federal Constitution. All of its decisions must be published online and are available in its internet portal[2].

3.1 Structure of the Documents

The full decision document, called (*inteiro teor*), is composed of four parts, namely: "Ementa", "Acórdão", "Relatório", and "Voto", which we now describe.

- The *Ementa* is a brief summary of the main topics discussed in each case and how the judges decided. We will be using the *Ementa* as the *reference summary* that automatic methods should aim to produce. In our corpus, the size of the *Ementa* was typically around 7% of the size of the full content.
- The *Acórdão* is a brief description of how each judge has decided and what the final decision was. This section represents around 2% of the full content.
- The *Relatório*, meaning report, is a compilation of the main arguments and events that happened during the trial. In general, this section accounts for about 22% of the full content.
- The last section, called *Voto*, may contain one vote, in case that the other judges agree with the first judge, or individual votes for each judge, otherwise. Because the votes need to address all the points raised by the petitioners, this tends to be the largest section covering around 69% of the full content.

The *Ementa* is useful for lawyers and other legal professionals when they are searching for decisions about a given topic. A good text should not be long, generally less than one page, making it a good summary of the full decision.

[1] https://duc.nist.gov/.
[2] http://www.stf.jus.br/.

{**"ementa"**: *"Embargos de declaração em recurso extraordinário com agravo. 2. Decisão monocrática. (...) 5. Agravo regimental a que se nega provimento."*,

"acordao": *"Vistos, relatados e discutidos estes autos, acordam os ministros do Supremo Tribunal Federal, em Segunda Turma, (...), por unanimidade, converter os embargos de declaração em agravo regimental e, a este, negar provimento, nos termos do voto do Relator."*,

"relatorio": *"(...) Trata-se de embargos de declaração opostos contra decisão que negou provimento a recurso, ao fundamento de que a natureza da matéria versada nos autos reveste-se de índole infraconstitucional. Aponta-se violação direta à Constituição Federal, em especial, aos artigos (...)."*,

"voto": *"(...) Tendo em vista o princípio da economia processual, recebo os embargos de declaração como agravo regimental e, desde logo, passo a apreciá-lo. (...)"*}

Fig. 1. Example of a document already divided into sections in JSON format.

3.2 Data Collection

In order to obtain the documents, the Scrapy [21] library was used to browse the search pages and to download the documents. Only a few documents from the years 2010 and 2011 could be successfully parsed. Thus most documents are dated from 2012 to 2018.

The raw text we obtained contains some undesired pieces of texts such as headings, footers, page numbers, *etc.* We used regular expressions to identify the starting and ending points of each section of interest and remove unwanted text. Finally, the text of the sections was dumped as a JSON object, one object per line.

In Fig. 1, we show an extract from a short document in the final JSON format. The ellipsis indicates the omission of content to save space.

The final file has about 173 MB and contains 10,623 decisions and can be downloaded from https://github.com/diego-feijo/rulingbr. There are around 26 million tokens in the entire dataset.

We investigated whether there is a correlation between the length (in tokens) of the *Ementa* section and all other sections combined (full document). This correlation would be important for us to determine the desired summary size when using the automatic summarizers. The calculated correlation coefficient was 0.39, which is considered weak and is reflected by a large dispersion.

4 Evaluating Summarization Systems on RulingBR

In this Section, we present results of out-of-the-box extractive summarization strategies on RulingBR dataset. The goal is to provide baseline results for future summarization techniques.

4.1 Experimental Setup

In order to establish some baselines using this corpus, we have run a few automatic summarization experiments using two common libraries.

The first library used was Gensim [20]. This is a software framework for Natural Language Processing that implements some popular algorithms for topical inference and has a TextRank implementation for summarization. This library implements a variation of the TextRank [13] algorithm.

The second library used was the Sumy package [3]. It has a large variety of algorithms implemented using a common interface which makes it easier to run and compare the results.

The choice of summarization algorithms was motivated by the fact that they could be applied directly to the text without requiring additional information such as part-of-speech tags or headlines.

The algorithms used in this experiment were the following.

TextRank uses a graph-based ranking model for text processing. This algorithm applies unsupervised methods for keyword and sentence extraction and is based on ideas borrowed from HITS [10] and PageRank [16]. Both Gensim and Sumy implement variations of the TextRank algorithm. The Gensim implementation was improved [2] replacing the cosine by the Okapi-BM25 similarity function.

Luhn [12], which uses statistical information derived from word frequency and distribution to compute a relative measure of significance, first for words and then for sentences. The set of sentences with the highest scores are extracted to make up the summary.

LexRank [7] is a graph-based method to compute a relative measure of importance which is based on the concept of eigenvector centrality in a graph representation of the sentences. In this model, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix for the graph representation of sentences.

4.2 Evaluation Metrics

The most commonly used metric for evaluating summarization algorithms is Rouge [11], which stands for Recall-Oriented Understudy for Gisting Evaluation. Its goal is to provide a measure of quality of an automatically generated summaries in comparison against a reference summary produced by humans.

The Rouge metric checks for overlapping text segments between the automatically generated summary and the reference summary. Different levels of granularity can be used. Rouge-1 counts the occurrences unigrams that appear in the automatically generated and the reference summaries. Rouge-2, counts how many bigrams were found (in the same order). Rouge-L stands for the longest common sub-sequence between the automatically generated and the reference summaries.

4.3 Experimental Procedure

Although both libraries used in the experiments support stopword removal, stemming, and tokenization, we opted to apply it beforehand as preprocessing steps to make sure that the same operations were applied in all settings. For most of the stages, we have used the Natural Language Toolkit (NLTK) [4], which is a widely used library for processing of natural language documents. It contains functions and trained models in many languages. We used this library for filtering, stemming, and tokenization.

Stopword Removal – In order to try to make a fair analysis of the content produced in the summaries, stopwords should be removed since their presence could artificially inflate the quality metrics (since the reference summaries would certainly contain many such words). We have used the Portuguese stop-list provided with NLTK. Also, we have filtered any token with fewer than two characters. This was done because these tokens have low discrimination power, and, as we are generating a summary, we expect that the words should contain relevant semantic meaning.

Stemming – This technique conflates the variant forms of words into a single stem. We used the NLTK implementation of the RSLP-Stemmer [9].

Tokenization – This is the task of separating the text into chunks. It is used for dividing the text into sentences and them into words. Recognizing the start and end of sentences is crucial for the extractive summarization algorithms because they will compute the score of each sentence and output the highest scored sentences. The tokenizer must identify situations such as when sentences were not being finished by a period (*e.g.* Hurry up!) or when a period was being used for an abbreviation (*e.g.* Mr. John) rather than to indicate the end of a sentence. Again, we used the NLTK implementation of the Punkt tokenizer trained for the Portuguese language.

Standardization – The documents in the corpus significantly vary in length due to the several subjects that are covered by the decisions. In order to try to generalize a pattern, some outliers needed to be dropped. Using a token (word) as measuring unit, we calculated the mean and the standard deviation for the summaries (99.53 ± 91.17) and for the full contents (1397.44 ± 2101.73). In order to reduce the dispersion, we removed outliers. Input documents with fewer than 300 words or more than the mean plus 3 times the standard deviation were treated as outliers. In a similar fashion, summaries with fewer than 19 words or more than the mean plus 3 times the standard deviation were also removed. With this standardization, we removed 616 decisions, which represent 5.80% of the total. Full contents mean became 1200.65, with a standard deviation of 893.86; Summary mean became 91.79, with a standard deviation of 62.92. The frequency distribution after the cleaning can be seen in the histograms of Fig. 2.

4.4 Model Parameters

It is important to notice that the evaluation scores could be affected by the size of the generated summaries. That happens because a longer summary would probably have a greater recall and, as consequence, a higher Rouge score.

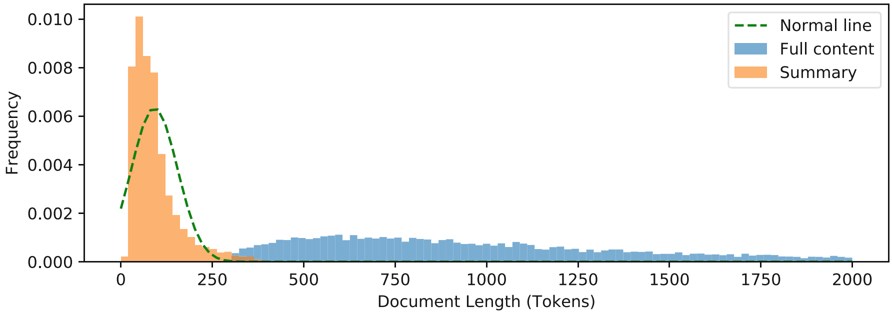


Fig. 2. Frequency distribution of the length of the summaries and the full content of the documents.

The libraries that generate automatic summaries receive as parameter the size of the desired output. As we discussed earlier, there is no strong correlation between the length of the document and the length of the summary. As shown in the histograms of Fig. 2, the size of the reference summaries can vary roughly between 30 and 150 tokens. So, setting the desired summary size to a fixed value will introduce an error as the size will be different from the size of reference summary. Nevertheless, we had to stick to a fixed size.

In these libraries, the output is entire sentences, so the total of words can be much smaller or larger than the desired output size. For example, the Gensim library receives the number of desired words, it computes the best sentences and will append them to the output until the difference between the desired output and the generated output is minimized. The Sumy library receives only the number of desired sentences, so the output may have a size completely different from the size of the reference summary (either much larger or much smaller).

In our dataset, sentence length can vary a lot. It is possible to find one-word sentences and sentences with a few hundred words. So, it is fairer to run our experiments with different size parameters. This way the results are not negatively impacted by an arbitrary choice of size.

4.5 Results

A higher Rouge score reflects a higher similarity between the automatically generated summary and the reference summary. Our goal when running this evaluation is to establish how standard extractive algorithms perform on this dataset. We have no intent in comparing those algorithms, as this would require evaluations under many different contexts and parameters.

As Table 1 shows, Rouge F-Score and Recall increase when the summary is longer. So, for a fair comparison, we used the scores of the runs in which the absolute differences between the length of the generated summary and its reference is minimized. Figure 3 shows the scores for Gensim using the desired output of 80 words, Luhn's algorithm with a fixed output of one sentence, LexRank and TextRank algorithms with a fixed output of two sentences.

In our experiments, the Gensim library generally performed slightly better. But, the highest Rouge-1 F-Score was obtained using the LexRank asking for four sentences in the Sumy library.

Table 1. Results of the summarization using different lengths of outputs. Best results per metric are shown in bold.

Algorithm	Length	Abs Dif	Rouge-1			Rouge-2			Rouge-L		
			F	P	R	F	P	R	F	P	R
Gensim	60	467,138	0.27	0.30	0.11	0.11	0.14	0.27	0.14	0.20	0.16
Gensim	80	**448,162**	0.29	0.29	0.14	0.12	0.14	0.32	0.15	0.19	0.19
Gensim	100	497,309	0.30	0.28	0.16	0.13	0.13	0.37	0.16	0.17	0.22
Gensim	120	587,463	0.30	0.27	0.18	0.13	0.12	0.40	0.15	0.16	0.25
Luhn	1	475,039	0.23	0.29	0.09	0.10	0.14	0.23	0.13	0.21	0.14
LexRank	1	616,140	0.21	**0.35**	0.07	0.09	0.16	0.18	0.11	**0.27**	0.11
TextRank	1	544,613	0.22	0.33	0.08	0.10	**0.17**	0.20	0.12	0.26	0.13
Luhn	2	503,009	0.27	0.27	0.14	0.12	0.13	0.32	0.15	0.17	0.20
LexRank	2	498,650	0.27	0.32	0.11	0.11	0.14	0.28	0.15	0.21	0.17
TextRank	2	543,740	0.24	0.31	0.12	0.11	0.15	0.27	0.14	0.21	0.18
Luhn	3	729,632	0.29	0.26	0.17	0.13	0.12	0.40	0.15	0.15	0.25
LexRank	3	513,033	0.29	0.30	0.15	0.12	0.13	0.35	**0.16**[a]	0.19	0.21
TextRank	3	673,067	0.26	0.29	0.15	0.12	0.14	0.33	0.14	0.18	0.22
Luhn	4	1,025,284	0.30	0.25	**0.20**	**0.13**	0.11	**0.45**	0.14	0.13	**0.28**
LexRank	4	609,894	**0.31**	0.29	0.17	0.13	0.12	0.40	**0.16**[a]	0.17	0.25
TextRank	4	859,861	0.27	0.27	0.17	0.12	0.13	0.38	0.14	0.16	0.25

[a] Both had exactly the same score.

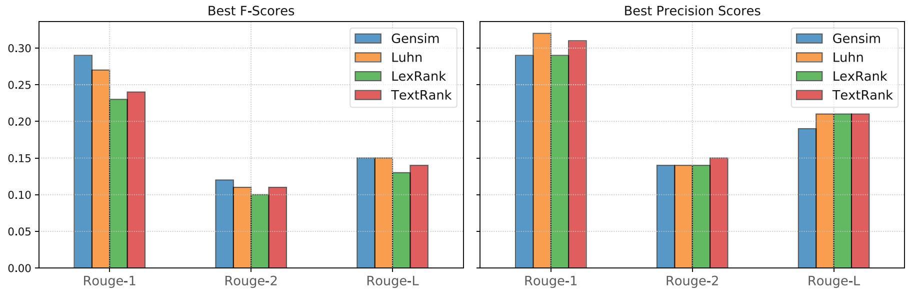


Fig. 3. F-Score and precision for the different summarization algorithms.

5 Conclusion

In this paper, we presented the RulingBR dataset, a corpus that can be used for natural language summarization. It differs from the existing corpora because it covers the legal domain and it is in Portuguese. We have analyzed different aspects of the dataset such as its organization, the size of each section, and how it can be used for the summarization task. We ran an experiment using different algorithms and libraries to establish baseline summarization results.

Despite the fact that the *Ementa* is a useful summary for legal professionals, it is not clear that the traditional general approaches for summarization could be directly applied to the legal domain producing texts that cover the same topics that a human would select.

The desired summary should contain the main topics discussed in the text. Perhaps, the desired output summary could be improved by appending these main topics, named entities, and compound terms. Also, we observed that the summary is composed of the final part of the *Acórdão*, the topics taken from the *Relatório*, and several ideas discussed in the *Voto*.

As a future work, we intend to test Neural models for summarization in order to identify the relevant aspects of the document and generate a summary in the style produced by a human.

References

1. Aleixo, P., Pardo, T.A.S.: CSTNews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory) (2008)
2. Barrios, F., López, F., Argerich, L., Wachenchauzer, R.: Variations of the similarity function of TextRank for automated summarization. arXiv preprint arXiv:1602.03606 (2016)
3. Belica, M.: Sumy: module for automatic summarization of text documents and HTML pages, April 2018. https://github.com/miso-belica/sumy
4. Bird, S.: NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive Presentation Sessions, COLING-ACL 2006, pp. 69–72. Association for Computational Linguistics, Stroudsburg (2006)
5. Chen, D., Bolton, J., Manning, C.D.: A thorough examination of the CNN/Daily mail reading comprehension task. CoRR abs/1606.02858 (2016). http://arxiv.org/abs/1606.02858
6. Collovini, S., Carbonel, T.I., Fuchs, J., Coelho, J.C., Rino, L., Vieira, R.: Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In: V Workshop em Tecnologia da Informação e da Linguagem Humana, Congresso da SBC, pp. 1605–1614 (2007)
7. Erkan, G., Radev, D.R.: LexRank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004)
8. Ganesan, K., Zhai, C., Han, J.: Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 340–348. Association for Computational Linguistics (2010)

9. Huyck, C., Orengo, V.: A stemming algorithm for the Portuguese language. In: International Symposium on String Processing and Information Retrieval, SPIRE, p. 0186, November 2001
10. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
11. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
12. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958)
13. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Proceedings of EMNLP 2004 and the 2004 Conference on Empirical Methods in Natural Language Processing, July 2004
14. Nallapati, R., Xiang, B., Zhou, B.: Sequence-to-sequence RNNs for text summarization. CoRR abs/1602.06023 (2016). http://arxiv.org/abs/1602.06023
15. Napoles, C., Gormley, M., Van Durme, B.: Annotated gigaword. In: Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction, AKBC-WEKEX 2012, pp. 95–100. Association for Computational Linguistics, Stroudsburg (2012)
16. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, pp. 161–172 (1998). citeseer.nj.nec.com/page98pagerank.html
17. Pardo, T.A.S., Rino, L.H.M.: Temário: Um corpus para sumarização automática de textos. Universidade de São Carlos, Relatório Técnico, São Carlos (2003)
18. Parker, R., Graff, D., Kong, J., Chen, K., Maeda, K.: English gigaword fifth edition, linguistic data consortium. Google Scholar (2011)
19. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304 (2017)
20. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA, Valletta, May 2010
21. ScrapingHub: Scrapy - a fast and powerful scraping and web crawling framework (2018). https://scrapy.org
22. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368 (2017)
23. Xiao: PyTeaser: Summarizes news articles, April 2018. https://github.com/xiaoxu193/PyTeaser
24. Xin Pan, P.L.: Models: models and examples built with TensorFlow, April 2018. https://github.com/tensorflow/models
25. Yin, W., Pei, Y.: Optimizing sentence modeling and selection for document summarization. In: Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI 2015, pp. 1383–1389. AAAI Press (2015)
26. Zhang, X., Lapata, M.: Sentence simplification with deep reinforcement learning. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, 9–11 September 2017, pp. 584–594 (2017)

Learning Word Embeddings from Portuguese Lexical-Semantic Knowledge Bases

Hugo Gonçalo Oliveira(✉)

CISUC, Department of Informatics Engineering, University of Coimbra,
Coimbra, Portugal
hroliv@dei.uc.pt

Abstract. This paper describes the creation of PT-LKB, new Portuguese word embeddings learned from a large lexical-semantic knowledge base (LKB), using the node2vec method. Resulting embeddings combine the strengths of word vector representations and, even with lower dimensions, achieve high scores in genuine similarity, which so far were obtained by exploiting the graph structure of LKBs.

Keywords: Lexical-semantic knowledge bases · Word embeddings
Lexical semantics · Semantic similarity

1 Introduction

Distributional word representations of language became a trend in natural language processing, mainly after the development of efficient models that learn low dimension word vectors (word embeddings) from very large corpora, with a neural network. Those include word2vec [8] and GloVe [10], extensively used in semantic similarity and analogy tasks. Distributional models reflect how language is used and enable to compute the similarity of words from the cosine of their vectors. Lexical-semantic networks (LKBs), such as WordNet [2], are alternative representations of word meaning, more theoretical, and generally handcrafted or created semi-automatically. A relevant difference is that, in LKBs, semantic relations are explicitly labelled, which does not happen in word embeddings, despite existing work on inducing specific relations from the word vectors [15]. Moreover, although there are several algorithms for computing word similarity [9] and relatedness [1] in LKBs, none is as straightforward as computing a cosine.

For Portuguese, several LKBs have been developed in the last 15 years and different word embeddings became available in the last two. Some of the previous were recently compared in word similarity tests [4], where LKBs seemed more suitable for computing genuine similarity, while word embeddings achieved better results for relatedness.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 265–271, 2018.
https://doi.org/10.1007/978-3-319-99722-3_27

This work describes how word embeddings can be learned from the structure of a large Portuguese LKB, using node2vec [6], a representation learning method for networks, which represents node neighbourhoods in a d-dimensional feature space, by applying a biased random walk procedure. Running node2vec in a Portuguese LKB results in PT-LKB, a collection of new word embeddings, where word similarity can be computed by the vector cosine, but where high scores can also be obtained for genuine similarity. This is confirmed by the results reported in four word similarity tests and compared to those obtained with other embeddings. It is also worth noticing that, due to the structure of the LKBs, new embeddings can use vectors with lower dimensions than the common 300.

2 Setup

Instead of picking one of the available LKBs for Portuguese, we built on previous work [4], where a large Portuguese LKB was created with relation instances in ten open lexical resources. This LKB is available online[1] in a single text file, with a relation instance in each line, followed by the number of lexical resources where it was found (e.g., `fruto HIPERONIMO_DE tomate 3`). This number can be seen as a sign of consensus on the validity/utility of the relation.

To learn the PT-LKB embeddings, a C++ implementation of node2vec, available as an example of the Stanford Network Analysis Platform[2], was used. This required the conversion of the LKB into a graph format accepted by node2vec: a file with a pair of related words in each line, followed by their weight, set to the number provided in the original LKB (e.g., `fruto tomate 3`). The relation name, not used by the embeddings, was lost in this process.

To take further conclusions, different embeddings were learned by changing the parameters of the algorithm, starting with the network used, either with all the relation instances (*All*) or only those in at least two, three, four, or five resources (*In2-5*), and either considering the weights or not. Other parameters include the dimension of the vectors (Dim), the walk length (Len(Walk)) and the number of walks (#Walks). Embeddings were first learned with default node2vec parameters (128, 80, 10), but other parameters were also tested, namely more walks (200) with a lower length (30), as well as a lower (64) and higher vector dimension (300, the most common in corpus-based embeddings).

3 Results in Word Similarity Tests

Resulting word embeddings were used to answer four Portuguese word similarity tests: PT-65 [5] and the Portuguese versions [12] of SimLex-999 (SL-999), WordSim-353 (WS-535) and RareWords (RW). These tests contain pairs

[1] http://ontopt.dei.uc.pt/index.php?sec=download_outros.
[2] http://snap.stanford.edu/node2vec/.

of Portuguese words and a gold similarity/relatedness score based on human judgements (e.g., `pássaro grua 0.24` or `menino rapaz 3.58`).

Validation consisted of measuring the correlation (Pearson ρ) between the gold scores with those obtained by the cosine of the word vectors in the embeddings. The obtained results were later compared to those of other embeddings available for Portuguese, namely a selection of the LX-DSemVectors [13] and of the NILC embeddings [7], all learned from large corpora, and also the Numberbatch embeddings [14] for Portuguese words, which combine data from the ConceptNet semantic network and other sources, including word2vec and GloVe.

Before computing similarity, the coverage of the words in the tests by the embeddings was analysed. This is relevant for further experiments where, when a word in a pair was not covered, similarity was set to 0. Tables 1 and 2 show the proportion of pairs covered by the PT-LKB embeddings and by the other embeddings tested, respectively. As expected, coverage by the PT-LKB embeddings is high when all instances are used, and decreases when minimum weight increases. Therefore, in the next experiment, we decided to use only the embeddings with instances in all, 2 or 3 resources (*All*, *In2*, *In3*). The coverage of the corpus-based embeddings is comparable to those of the *All* embeddings, 90–100%, except for RW, where every pair has a rare word, and coverage is 80–89%, against 71% of *All*. Numberbatch embeddings have a lower coverage than the corpus-based, possibly because ConceptNet is not as developed for Portuguese than for other languages.

Table 1. Word pairs covered by the PT-LKB embeddings.

Inst.	PT-65	SL-999	WS-353	RW
All	100%	98%	95%	71%
In2	95%	95%	88%	55%
In3	78%	85%	72%	39%
In4	45%	66%	44%	27%
In5	25%	44%	18%	17%

Table 2. Word pairs covered by other Portuguese embeddings.

Embedding	PT-65	SL-999	WS-353	RW
LX vanilla	100%	97%	98%	80%
LX p_17	100%	97%	98%	87%
NILC (all models)	100%	98%	93%	89%
Numberbatch	98%	96%	90%	47%

Table 3. Results in four word similarity tests, with the default parameters of node2vec in the All-network, Redun2 and Redun3.

Instances	Weights	ρ			
		PT-65	SL-999	WS-353	RW
All	Yes	**0.89**	**0.57**	**0.43**	**0.29**
All	No	0.87	0.56	**0.43**	0.28
In2	Yes	0.75	0.49	0.20	0.27
In2	No	0.73	0.47	0.19	0.26
In3	Yes	0.59	0.29	0.02	0.24
In3	No	0.56	0.27	0.00	0.24

Table 3 shows the similarity results for the PT-LKB embeddings, with the default node2vec parameters, considering and not considering the weights. In every test, ρ is higher for the *All* embeddings and, though with small differences, this happens when the weights are considered, which suggests that the number of resources where each relation instance was found is a relevant aspect.

Table 4 has the results obtained for PT-LKB embeddings learned with different parameters in node2vec, and shows that, with the exception of PT-65, the same results can be obtained with a higher number of walks, though with smaller 64-sized vectors. To have an idea, while the text file of the NILC embeddings with 300-sized vectors uses about 2.5G, the *All* embeddings use about 125 MB, 250 MB and 600 MB, respectively with 64, 128 and 300-sized vectors. This also benefits from the fact that the latter were learned from a LKB, which only contains lemmas. This does not have a noticeable impact in the similarity tests but, when it comes to other tasks where words are inflected, lemmatisation has to be made before retrieving the word vector. A similar issue happens for the Numberbatch embeddings. However, in this case, even though it is only available with 300-sized vectors, the Portuguese part only uses 108 MB.

Table 4. Results in four word similarity tests, with different parameters of node2vec in the All-network.

Instances	Weights	Dim	Len(Walk)	#Walks	ρ			
					PT-65	SL-999	WS-353	RW
All	Yes	64	80	10	0.82	0.57	0.43	0.28
All	Yes	64	30	200	0.85	**0.58**	**0.45**	**0.30**
All	Yes	128	80	10	**0.89**	0.57	0.43	0.29
All	Yes	128	30	200	0.86	**0.58**	**0.45**	**0.30**
All	Yes	300	80	10	0.88	0.56	0.41	0.29
All	Yes	300	30	200	0.88	0.55	0.43	**0.30**

Table 5. Results of other word embeddings in four word similarity tests.

Source	Model	Dim	ρ			
			PT-65	SL-999	WS-353	RW
ConceptNet	Numberbatch	300	0.81	**0.63**	**0.50**	0.31
NILC	fastText skip-gram	300	0.78	0.33	0.41	**0.42**
NILC	word2vec c-bow	600	0.60	0.25	0.33	0.36
NILC	GloVe	300	0.74	0.30	0.32	0.38
NILC	GloVe	600	0.75	0.30	0.30	0.37
LX	word2vec skip-gram (p_17)	300	0.66	0.33	0.48	0.35
LX	word2vec skip-gram (vanilla)	300	0.57	0.23	0.36	0.27
NILC	fastText c-bow	100	0.69	0.22	0.20	0.29
NILC	fastText skip-gram	100	0.71	0.28	0.28	0.40
NILC	word2vec skip-gram	100	0.52	0.22	0.34	0.34
NILC	word2vec c-bow	100	0.38	0.15	0.24	0.31
NILC	GloVe	100	0.72	0.27	0.31	0.37
NILC	Wang2vec skip-gram	100	0.69	0.30	0.36	0.38
NILC	Wang2vec c-bow	100	0.65	0.34	0.36	0.39
All-LKB	Adj-Cos	N/A	0.86	0.58	0.44	0.38
All-LKB	PR-Cos	N/A	0.87	0.61	0.46	0.23
CONTO.PT	μ	N/A	0.74	0.47	0.30	0.41

Table 5 shows a selection of results for the same similarity tests. The first part includes the corpus embeddings used in previous work [4]. Here, we highlight the good performance of Numberbatch, especially in SL-999 and WS-353. This confirms that combining knowledge in a semantic network with corpus-based embeddings provides a good balance between similarity and relatedness, as it happened for word similarity tests in other languages [14]. The main drawback is the lower coverage of Numberbatch for Portuguese, which explains its low storage size, but leads to lower scores in RW.

The second part of the table has results obtained with 100-sized vector NILC embeddings, which, despite having a dimension between the LKB embeddings with lower dimensions (64 and 128), perform poorly in all tests but RW. The results in the last part of the table were not obtained with embeddings, but by exploiting the same LKB with other algorithms, also described in previous work [4]: (i) similarity of the adjacencies of each word, i.e. directly-connected words, computed with the cosine similarity (Adj-Cos); (ii) PageRank vectors [11], obtained after running PageRank once for each word, creating a word vector with the resulting word ranks, and finally computing the cosine similarity of the target word vectors (PR-Cos); (iii) Memberships (μ) of words in the synsets of CONTO.PT [3], a fuzzy wordnet extracted from the same LKB.

Overall, with the default parameters, the PT-LKB embeddings achieve the best reported results in PT-65 (0.89 vs 0.87), previously obtained with the PR-Cos algorithm. Yet, even if the PageRank vectors are pre-computed, PR-Cos requires larger vectors, with size equals to the number of words in the LKB. In SL-999, the PT-LKB embeddings perform better than any corpus embedding, but are outperformed by Numberbatch and by PR-Cos. When it comes to WS-353 and RW, the PT-LKB embeddings perform below the corpus embeddings with 300-sized vectors. Yet, on WS-353, even the PT-LKB embeddings with 64-sized vectors achieve higher results than those of corpus embeddings with 100-sized vectors.

4 Conclusion

This paper described how node2vec was used to learn new Portuguese word embeddings from LKBs, dubbed PT-LKB, then validated with word similarity tests. New embeddings performed well, in some cases better than corpus-based embeddings, even when using lower-dimensions.

Yet, this should be seen as a preliminary validation. The PT-LKB embeddings should further be tested in other tasks, such as semantic textual similarity or analogies. In fact, we suspect that they are not suitable for solving analogies in available datasets, as the latter typically contain named entities (e.g. counties and their capitals), generally not present in a LKB. Other experiments that may also be performed include analysing the impact of learning this kind of embeddings from individual LKBs or using different weights for different relation types. But this might also depend on the target task. Even though synonymy and hypernymy are more relevant for genuine similarity, other relations play an important role for computing relatedness.

The PT-LKB embeddings with best results in the similarity tests are freely available from http://ontopt.dei.uc.pt/index.php?sec=download_outros.

References

1. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–47 (2006)
2. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press, Cambridge (1998)
3. Gonçalo Oliveira, H.: CONTO.PT: groundwork for the automatic creation of a fuzzy Portuguese WordNet. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS, vol. 9727, pp. 283–295. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_29
4. Gonçalo Oliveira, H.: Distributional and knowledge-based approaches for computing Portuguese word similarity. Information **9**(2), 35 (2018)
5. Granada, R., Trojahn, C., Vieira, R.: Comparing semantic relatedness between word pairs in Portuguese using Wikipedia. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.G. (eds.) PROPOR 2014. LNCS, vol. 8775, pp. 170–175. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09761-9_17

6. Grover, A., Leskovec, J.: Node2vec: scalable feature learning for networks. In: Proceedings of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 855–864. ACM (2016)
7. Hartmann, N.S., Fonseca, E.R., Shulby, C.D., Treviso, M.V., Rodrigues, J.S., Aluísio, S.M.: Portuguese word embeddings: evaluating on word analogies and natural language tasks. In: Proceedings the 11th Brazilian Symposium in Information and Human Language Technology, STIL 2017 (2017)
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the Workshop Track of the International Conference on Learning Representations, ICLR, Scottsdale, Arizona (2013)
9. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity: measuring the relatedness of concepts. In: Demonstration Papers at HLT-NAACL 2004, pp. 38–41. ACL Press, Stroudsburg (2004)
10. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Empirical Methods in Natural Language Processing, EMNLP, pp. 1532–1543 (2014)
11. Pilehvar, M.T., Jurgens, D., Navigli, R.: Align, disambiguate and walk: a unified approach for measuring semantic similarity. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Sofia, Bulgaria, Long Papers, vol. 1, pp. 1341–1351. ACL Press (2013)
12. Querido, A., et al.: LX-LR4DistSemEval: a collection of language resources for the evaluation of distributional semantic models of Portuguese. Revista da Associaçáo Portuguesa de Linguística **3**, 265–283 (2017)
13. Rodrigues, J., Branco, A., Neale, S., Silva, J.: LX-DSemVectors: distributional semantics models for Portuguese. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS, vol. 9727, pp. 259–270. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_27
14. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: an open multilingual graph of general knowledge. In: Proceedings of 31st AAAI Conference on Artificial Intelligence, San Francisco, California, USA, pp. 4444–4451 (2017)
15. Vylomova, E., Rimell, L., Cohn, T., Baldwin, T.: Take and took, gaggle and goose, book and read: evaluating the utility of vector differences for lexical relation learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Long Papers, vol. 1, pp. 1671–1682. ACL Press (2016)

SIMPLEX-PB: A Lexical Simplification Database and Benchmark for Portuguese

Nathan S. Hartmann[1](✉), Gustavo H. Paetzold[2], and Sandra M. Aluísio[1]

[1] Institute of Mathematics and Computer Science, University of São Paulo,
São Paulo, Brazil
{nathansh,sandra}@icmc.usp.br

[2] Department of Computer Science, University of Sheffield, Sheffield, UK
g.h.paetzold@sheffield.ac.uk

Abstract. Lexical Simplification has the function of changing words or expressions for synonyms that can be understood by a larger number of people. It is very common to have in mind a target audience which will benefit from the task, such as children, low-literacy audiences, and others. In recent years there has been great activity in this field of research, especially for English, but also for other languages such as Japanese and multilingual and cross-lingual scenarios. Few works have children as target audience. Currently, in Brazil, the *Programa Nacional do Livro Didático* (PNLD) is an initiative with a broad impact on education, as it aims to choose, acquire, and distribute free textbooks to students in public elementary schools. In this scenario, adapting the level of complexity of a text to the reading ability of a student is a determinant of his/her improvement and whether he/she reaches the level of reading comprehension expected for that school year. On the other hand, there have not been publicly available resources on lexical simplification for Portuguese as yet. Therefore, the development of this material is urgent and welcome. This work compiled the SIMPLEX-PB, the first available corpus of lexical simplification for Brazilian Portuguese. We also make available a benchmark for evaluating the most well-known methods of LS in our dataset.

Keywords: Lexical simplification · Dataset · Benchmark · Evaluation

1 Introduction

Lexical Simplification (LS) has the function of changing words or expressions for synonyms that can be understood by a larger number of people. It is very common to have in mind a target audience which will benefit from the task, such as children, low-literacy audiences, people with cognitive disabilities and second language learners [24]. An automatic system for LS performs the following steps in a *pipeline* (see Fig. 1): (i) lexical complexity analysis, which selects words or expressions that are considered complex for a reader and/or task; (ii) search

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 272–283, 2018.
https://doi.org/10.1007/978-3-319-99722-3_28

for substitutes, in general, synonyms with the same meaning used in context; and (iii) ranking of the pitch synonyms according to how simple they are to the reader and/or task [29]. After choosing the appropriate synonym, the word being focused on is replaced by a synonym, which can ask for adjustments in the writing of the words of the sentence, such as the adequacy of gender and/or number.

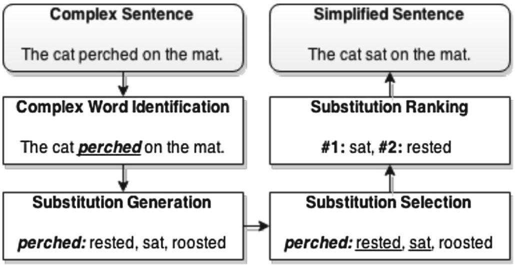


Fig. 1. Lexical simplification pipeline.

In recent years there has been great activity in this field of research, especially for English, [4,17,29–31] but also for other languages such as Japanese and multilingual and cross-lingual scenarios [12,13,30,32]. Only two studies focus on children [12,13].

Currently, in Brazil, the PNLD is an initiative with a broad impact on education, as it aims to choose, acquire, and distribute free textbooks to students in public elementary schools. Since 2001, the PNLD has been focusing on lexicography [6], selecting and acquiring specific dictionaries for each school year. In this scenario, adapting the level of complexity of a text to the reading ability of a student is a determinant of his/her improvement and whether he/she reaches the level of reading comprehension expected for that school year. On the other hand, there have not been publicly available resources on lexical simplification for Portuguese as yet. Therefore, the development of this material is urgent and welcome.

This paper presents the SIMPLEX-PB, the first publicly available corpus of lexical simplification for Brazilian Portuguese, targeting children from the 3rd to 9th years (Elementary School). We also make available a benchmark for evaluating the most well-known methods of LS in our dataset[1].

2 Related Works

The benchmark compiled by [29] for the SemEval 2012 Text Simplification shared-task was based on the Semeval 2007 Lexical Substitution *gold-standard* (LEXSUB) [17]. The 2007 joint task asked participants to generate substitutes

[1] Our corpus and benchmark are available at github.com/nathanshartmann/simplex.

for a target word. The LEXSUB dataset consists of 2,010 sentences, 201 target words each with 10 sentences as contexts which were annotated by 5 native English speakers. It covers mostly polysemous target words, including nouns, verbs, adjectives, and adverbs. For the joint task of 2012 (SemEval 2012 dataset, the authors reused this dataset and asked annotators to rank substitutes for each individual context in ascending order of complexity, thus enabling the joint task in Lexical Simplification. The selected annotators (graduated students) had high proficiency levels in English as second language learners.

CW Corpus [27] is composed of 731 sentences from the Simple English Wikipedia in which exactly one word had been simplified by Wikipedia editors from the standard English Wikipedia. These simplifications were mined from Simple Wikipedia edit histories and each entry gives an example of a sentence requiring simplification by means of a single lexical edit. This dataset has been used in the Complex Word Identification (CWI) task. In the CWI task of SemEval 2016 [20], 400 non-native English speakers annotated the dataset, mostly university students or staff. Using the total of 9,200 sentences, 200 of them were split into 20 subsets of 10 sentences, and each subset was annotated by a total of 20 volunteers. The remaining 9,000 sentences were split into 300 subsets of 30 sentences, each of which was annotated by a single volunteer.

For Japanese, there are two LS datasets available. The SNOW R4 dataset [12], with 2,330 sentences, contains simplifications created and ranked by 5 annotators. These simplifications were rated as appropriate or not based on the following two criteria: if the sentence became unnatural as a result of the substitution and if the meaning of the sentence changed as a result of the substitution. The rank of each target word was decided based on the average of the rank from each annotator, following the previous research [29]. The BCCWJ dataset [13] was built to overcome several deficiencies of SNOW R4 dataset. It is the first controlled and balanced dataset for Japanese lexical simplification with high correlation with human judgment. A crowdsourcing application was used to annotate 1,800 sentences. Five annotators wrote substitutes, five annotators selected a substitution word that did not change the meaning of the sentence and five annotators performed the simplification ranking.

3 Corpus

3.1 Complex Words for Elementary School Years

To identify which words were difficult for children, we compiled lists of difficult words from dictionaries for the 3rd to 7th years (Elementary School 1) and 8th to 9th years (Elementary School 2). These lists delimit the lexical knowledge addressed in those school years. The choice of dictionaries was based on the list of these resources provided by the PNLD. The target audience of these dictionaries are school year cycles categorized by the *Ministério da Educação* (MEC) of the Brazilian government. The 1st to 3rd years are categorized as the 1st cycle, 4th and 5th year as the 2nd cycle, 6th and 7th years as the 3rd cycle, and 8th and 9th years as the 4th cycle. We focused on the end of the first cycle and cycles 2 to 4.

The PNLD considers that the level 1 dictionaries (1st cycle) contain a small number of entries (between 1,000 and 3,000 lexical units) belonging to thematic fields related to the daily routine of children. In general, these dictionaries limit word classes to nouns, adjectives, and verbs. The dictionaries of level 2 (second cycle) are characterized by an intermediate information density. They present a greater number of entries (3,000 to 15,000 lexical units), taking into account the growth of children's world knowledge. All grammatical classes are present and there are advances in the density of grammatical information. Finally, the dictionaries of levels 3 and 4 (3rd and 4th cycles or Elementary School 2) differ from the other two in several aspects, such as the much larger number of entries (19,000 to 35,000 lexical units) and the inclusion of regionalisms and technical-scientific terms. In general, they are close to the major dictionaries of the Portuguese language.

It is known that dictionaries compiled for a given school year are composed of words to be learned in this cycle - dictionaries are compiled for queries of unknown words. Therefore, the words contained in the level 1 dictionaries can be simple words compared to the next level, and so on.

3.2 Benchmark for Lexical Simplification in Portuguese

A Lexical Simplification benchmark should contain sentences that have at least one complex word identified and a list of suitable gold substitutes so that once a simplifier generates a candidate substitution for a given word, we can check whether it is suitable or not. We rely on the benchmark developed by [29]. We chose the corpus compiled by [7], which contains texts written for children. We searched for sentences with at least 5 tokens and containing exactly 1 complex word. Here, complex words are all those not contained in the word list for the 8th and 9th years. A total of 16,170 sentences with exactly one complex word were identified.

Due to the time required to annotate all of these sentences, we chose to sample them. In a different way from the random sampling performed for English [29], we selected 1,719 instances following the proportion of content words found in our corpus: 56% nouns, 18% adjectives, 18% verbs, and 6% adverbs. From this distribution, we also made the subdivision equally distributed to privilege: more frequent words, words with a greater number of synonyms and words with more senses. Words with more synonyms are problematic for an automatic system that must identify which synonym is appropriate, and words with more meanings are ambiguous. Altogether, 757 distinct words were a target of simplification in our corpus. This corpus can grow up in a future annotation.

Given a complex word, we lemmatized them using DELAF [19] and we retrieved all their synonyms which contain the same POS tag from TeP 2.0 database [16]. TeP (*Thesaurus eletrônico para o Português do Brasil*) contains 19,888 synonyms for 44,678 words in Portuguese. In order to list synonyms for human annotation, we sorted them by their frequencies in a large corpus and the appropriateness of the words to the given context. We used the corpus compiled in [9] to query the frequency of words. To fit the context, we used the embedding model also trained in [9] and calculated the mean of cosine similarity between a synonym and the context of the difficult word (3 words to the left and 3 to the right). More adequate synonyms have greater cosine similarity between their embedding and the embeddings of the context words. The synonyms were sorted by the average of the two established criteria. We limited the number of synonyms for the 15 best ranked, because we identified that some words contained more than 300 synonyms (e.g., the verb to be (verb *ser* in Portuguese), making human annotation impossible.

The dataset annotation was performed by three linguistic specialists for children. Two of them have a MSc and the third one has a Ph.D. All of them have years of expertise in teaching. The annotator filtered which words were appropriate to replace the original complex word. They could also suggest replacements that were not listed. The Ph.D. linguist annotated all sentences and each of the MSc linguists annotated half of them in a double-blind procedure. The Cohen Kappa [3] was 0.74 for the first pair of annotators and 0.72 for the second pair. Following Cohen Kappa authors, these are substantial annotation agreements. For the final dataset, we considered only words which both annotators agreed upon (for both filtered and suggested words). The replacements were also ranked by simplicity and not adequacy to context.

4 Methods and Evaluation

Our LS methods for Portuguese are based on the pipeline of Fig. 1 and are performed in two steps: Substitution Generation and Substitution Ranking. We chose not to address Complex Word Identification in order to more easily draw comparisons with LS contributions for English, almost all of which do the same. Much like [5], we chose to address Substitution Selection jointly with Substitution Ranking by grammaticality and meaning preservation features as inputs.

In what follows, we describe the approaches we used to tackle each of these steps.

Substitution Generation. We used two approaches to generate candidates:

- **Lexicon-Based:** Given a target complex word, synonyms are extracted from a lexicon to be used as candidate substitutions. This approach was first introduced by [2], who used it to simplify texts for the aphasic.
- **Embeddings-Based:** Extracts as candidate substitutions the 10 non- morphological variants with the highest cosine similarity with the target complex word given a word embeddings model. This approach was made popular by [5,23], who achieved impressive results for the English language.

Our lexicon-based approach extracts synonyms from the TeP 2.0 database. We also created multiple variants of embeddings-based approaches by using four different approaches: **word2vec** which employs the traditional continuous bag-of-words model of [18]; **wang2vec** that is a variant of word2vec that incorporates word order information [14]; **fasttext** which is known for its good performance modeling characters [11]; and **glove** that incorporates matrix factorization principles during training [26]. All four models use 300 dimensions and are trained over the [8] corpus, containing almost 1.4 billion words from assorted sources.

Substitution Ranking. We evaluate three ranking approaches:

- **Frequency-based (Frequency):** Ranks candidates according to their frequency in a large corpus [2].
- **Rank averaging (Rank Avg.):** Calculates features for each candidate, ranks them according to each one, then averages the ranks across all features in order to produce a final ranking [5].
- **Pairwise regression (Regression):** A supervised ridge regression model that learns how to quantify the simplicity difference between words. During training, it receives as input features for a pair of candidate substitutions, along with the difference in simplicity between them. During testing, it receives as input a set of candidate substitutions, calculates the simplicity difference between each and every pair, then averages the simplicity difference for each candidate in order to produce a final ranking [21].

We extracted word frequencies from the Brazilian Portuguese portion of the OpenSubtitles2016 corpus [15]. The corpus contains nearly 1.4 billion words extracted from subtitles. We chose this corpus because, as demonstrated in the literature [1,28], word frequencies from spoken text corpora tend to correlate much more closely to word familiarity than frequencies from other domains.

Our rank averaging and pairwise regression approaches use the same 17 features as input: the frequency of n-grams surrounding the target complex word in the OpenSubtitles2016 corpus considering all combinations of n-grams formed by independent windows of $0 \leq n \leq 2$ tokens to its left and right (9 features); the cosine similarity between the candidate and the target complex word based on the four embedding models (4 features); and the average cosine similarity between the candidate and the six words surrounding the target (three tokens to the left, and three to the right), also based on our four embedding models (4 features).

We chose these features for their effectiveness in English LS [5,10,21] and because they require only raw text to be produced.

4.1 Candidate Generation Proficiency

In our first experiment, we compared the performance of our four Substitution Generation approaches alone. For evaluation, we used the benchmark dataset compiled in this work, where each instance is composed of a target complex word in a sentence and a set of manually ranked gold substitutions. Four evaluation metrics were used: **Potential**, which is the proportion of instances for which at least one gold candidate was generated; **Precision**, the proportion of generated candidates that are among the gold candidates; **Recall**, the proportion of gold candidates generated; and **F1**, the harmonic mean between Precision and Recall.

While Precision, Recall, and F1 are very well-known metrics, Potential is unique, and measures how well the generator would do when paired with a 100% accurate Substitution Selection approach in an idealized scenario. The results are featured in Table 1.

Table 1. Candidate generation evaluation results. TeP represents our lexicon-based approach, and the remainder our embeddings-based approaches.

	Potential	Precision	Recall	F1
TeP	**0.506**	**0.068**	**0.506**	**0.121**
wang2vec	0.368	0.044	0.336	0.078
glove	0.378	0.043	0.328	0.076
word2vec	0.335	0.038	0.291	0.068
fasttext	0.259	0.028	0.208	0.050

Our lexicon-based approach (TeP) achieved the highest scores for all four metrics. Among embeddings-based methods, the glove model performed best, with fasttext featuring the least impressive results. At first glance, the results may seem surprising, since embedding models have been shown to be much more effective than lexicon-based approaches for English [21]. Inspecting our benchmarking dataset, we found that this is caused mainly by the fact that it greatly differs from benchmarking datasets for English, such as LexMTurk [10] and BenchLS [25]; while LexMTurk and BenchLS feature an average of 12.64 ± 6.4 and 7.36 ± 5.3 gold candidate substitutions, respectively, ours has much lower coverage, featuring only 1.43 ± 0.7 in average. The unusually lower coverage of our dataset puts embeddings-based approaches at a disadvantage, since, conversely to how lexicon-based approaches work, embeddings-based approaches are usually better at maximizing the Recall of generated candidates instead of their Precision. It must also be pointed out that the synonyms in our benchmarking dataset were extracted from the TeP itself, which gives it an inherent advantage over our embeddings.

Nevertheless, embeddings-based approaches usually offer much higher Recall and Potential because the vocabulary they are built upon is much larger than that of human-made lexicons. However, because of the low coverage of our dataset, not many gold candidates are featured within the 10 words with the highest cosine similarity with the target, which greatly compromises Recall and Potential. By increasing the number of candidates generated by our embeddings from 10 to 60, for example, we actually surpass TEP in both Potential and Recall, but because the Precision is so severely compromised, the F1 ends up being much lower.

4.2 Target Replacement Proficiency

In our second experiment, we paired our four Substitution Generation with our three Substitution Rankers to produce 15 full simplifiers, then compared their performance in replacing complex words.

The evaluation metric we use is Accuracy: the proportion of instances for which the highest ranking candidate is within the gold simplifications. We train our supervised pairwise regressor over the ranks included in our benchmark dataset. Notice that this does not necessarily bias the supervised ranker, since the rankings present in the benchmark dataset are gold candidates only, and the candidates that are ranked by the regressor during testing are produced by a generator, and hence are potentially spurious.

The results in Table 2 revealed that rank averaging outperforms both our frequency-based and supervised rankers for all generators. Although frequency-based approaches tend to be outperformed by more elaborate strategies, rank averaging is often outperformed by pairwise regression models in LS benchmarks for English [25]. We hypothesize that our supervised ranker performed poorly because the rankings present in our dataset were produced in context-unaware fashion i.e., the gold candidates were ranked independently from the context of the target complex word. Because of that, the rankings do not capture grammaticality or meaning preservation, which causes our supervised ranker to neglect these properties and hence prioritize spurious candidates more often.

Table 2. Candidate replacement evaluation results. Each line represents a Substitution Generation approach, each column a Substitution Ranking approach, and each cell the Accuracy resulting from their combination.

	Frequency	Rank avg.	Regression
TeP	0.227	0.292	0.211
wang2vec	0.167	0.200	0.121
fasttext	0.172	0.175	0.118
glove	0.086	0.174	0.095
word2vec	0.112	0.158	0.094

4.3 Error Analysis

In our final experiment, we conduct an error analysis of some of our simplifiers. For that purpose we use PLUMBErr [22], an error analysis platform for LS. As input it takes a benchmarking dataset, a list of words deemed complex by a certain target audience, the ranked candidates produced by a lexical simplifier, and an optional list of binary judgments made by a Complex Word Identification system for each target word in the benchmarking dataset. As output it identifies 7 types of scenarios that can happen throughout the simplification process: **2A:** The target word is in the complex word list but was identified as simple by the system; **2B:** The target word is not in the complex word list but was identified as complex by the system; **3A:** None of the candidate substitutions generated are among the gold candidates in the dataset; **3B:** All of the valid candidate substitutions generated are in the complex word list; **4:** The highest ranking candidate is not within the gold candidates of the dataset; **5:** The highest ranking candidate is in the list of complex words; and **1:** The target word is replaced by a gold candidate that is not in the complex word list (no error).

As the list of complex words we use the list of words in our dictionaries for the 8th and 9th school years. We chose this list because it has a substantial overlap with the list of target words in our dataset. Since we do not address Complex Word Identification in this contribution, we assume all target words are complex. We used PLUMBErr to analyze five simplifiers: each and every one of our generators paired with our rank averaging ranker, which performed best in our previous experiment. Table 3 features the cumulative proportion of instances in which each error was made. In other words, if, for instance, the simplifier made an error of type 2A in 30% of the instances, then it will only have the opportunity to make an error of type 2B for the remaining 70% of instances.

Table 3. Error analyses. Each line represents a Substitution Generation approach paired with rank averaging.

	2A	2B	3A	3B	4	5	1
TeP	0.00%	14.04%	25.86%	48.89%	7.76%	2.22%	1.23%
fasttext	0.00%	14.04%	63.30%	20.69%	0.37%	0.12%	1.48%
wang2vec	0.00%	14.04%	52.22%	29.68%	3.20%	0.74%	0.12%
word2vec	0.00%	14.04%	57.39%	25.37%	2.59%	0.49%	0.12%
glove	0.00%	14.04%	49.88%	33.13%	2.96%	0.00%	0.00%

The results further highlight the differences between lexicons and embeddings. Our lexicon-based generator makes a much smaller number of 3A errors than the others, meaning they find valid candidates more easily. However, this is mostly due to the fact that it exploits the same resource used to create the benchmark dataset. On the other hand, the embeddings tend to produce simpler candidates, since the proportion of 3B errors they make is much smaller.

Finally, it can also be noticed that the proportion of successful simplifications is very low for all simplifiers. As discussed in our previous experiments, our benchmark dataset has a very low gold candidate coverage, which makes the task inherently challenging. However, it becomes even more challenging when you incorporate the restriction that the replacements produced cannot have been judged complex by an external target audience, which is why our simplifiers managed to correctly simplify at best 1.48% of the instances.

5 Conclusion

We presented SIMPLEX-PB, the first lexical simplification corpus for Brazilian Portuguese. We conducted a benchmark comparing several well-known methods for LS using SIMPLEX-PB. We found that combining lexicons with an unsupervised ranking approach yields the best results. In the future, we aim to complement the gold candidates in our dataset and conceive new LS approaches.

Acknowledgments. We thank Larissa Pícoli, Livia Cucatto and Magali Duran for annotating our corpus. This work was supported by FAPESP proc. 2016/00500-1.

References

1. Brysbaert, M., New, B.: Moving beyond kucera and francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american English. Behav. Res. Methods **41**, 977–990 (2009)
2. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of English newspaper text to assist aphasic readers. In: Proceedings of the AAAI 1998 Workshop on Integrating Artificial Intelligence and Assistive Technology, pp. 7–10 (1998)
3. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 37–46 (1960)
4. De Belder, J., Moens, M.-F.: A dataset for the evaluation of lexical simplification. In: Gelbukh, A. (ed.) CICLing 2012. LNCS, vol. 7182, pp. 426–437. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28601-8_36
5. Glavaš, G., Štajner, S.: Simplifying lexical simplification: do we need simplified corpora? In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015), vol. 2, pp. 63–68 (2015)
6. da Graça Krieger, M.: Dicionários escolares e ensino de língua materna. Estudos Linguísticos (São Paulo 1978) **41**(1), 169–180 (2016)
7. Hartmann, N., Cucatto, L., Brants, D., Aluísio, S.: Automatic classification of the complexity of nonfiction texts in portuguese for early school years. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 12–24. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_2
8. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: evaluating on word analogies and natural language tasks. arXiv preprint arXiv:1708.06025 (2017)
9. Hartmann, N.S.: ASSIN shared task - solo queue group: mix of a traditional and an emerging approaches. In: Avaliação de Similaridade Semântica e Inferência Textual (ASSIN), Propor Workshop (2016)
10. Horn, C., Manduca, C., Kauchak, D.: Learning a lexical simplifier using wikipedia. In: Proceedings of the 52nd Annual Meeting of the ACL (ACL 2014), pp. 458–463 (2014)
11. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)

12. Kajiwara, T., Yamamoto, K.: Evaluation dataset and system for Japanese lexical simplification. In: ACL (Student Research Workshop), pp. 35–40. The Association for Computer Linguistics (2015)
13. Kodaira, T., Kajiwara, T., Komachi, M.: Controlled and balanced dataset for Japanese lexical simplification. In: Proceedings of the ACL 2016 Student Research Workshop, pp. 1–7. Association for Computational Linguistics (2016)
14. Ling, W., Dyer, C., Black, A., Trancoso, I.: Two/too simple adaptations of Word2Vec for syntax problems. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2015)
15. Lison, P., Tiedemann, J.: OpenSubtitles 2016: extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the 10th LREC (2016)
16. Maziero, E.G., Pardo, T.A., Di Felippo, A., Dias-da Silva, B.C.: A base de dados lexical e a interface web do tep 2.0: thesaurus eletrônico para o português do brasil. In: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, pp. 390–392. ACM (2008)
17. McCarthy, D., Navigli, R.: Semeval-2007 task 10: English lexical substitution task. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), pp. 48–53. Association for Computational Linguistics (2007)
18. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013), pp. 746–751 (2013)
19. Muniz, M.C.M.: A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. Master's thesis, Universidade de São Paulo, Brasil (2004)
20. Paetzold, G., Specia, L.: SemEval 2016 task 11: complex word identification. In: Proceedings of the 10th International Workshop on Semantic Evaluation, pp. 560–569. Association for Computational Linguistics, San Diego, California, June 2016
21. Paetzold, G., Specia, L.: Lexical simplification with neural ranking. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, pp. 34–40. ACL (2017)
22. Paetzold, G.H., Specia, L.: PLUMBErr: an automatic error identification framework for lexical simplification. In: Proceedings of 1st Quality Assessment for Text Simplification (LREC-QATS 2016), pp. 1–9 (2016)
23. Paetzold, G.H., Specia, L.: Unsupervised lexical simplification for non-native speakers. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016), pp. 3761–3767 (2016)
24. Paetzold, G.H., Specia, L.: A survey on lexical simplication. J. Artif. Intell. Res. **60**, 549–593 (2017)
25. Paetzold, G.H., Specia, L.: Benchmarking lexical simplification systems. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pp. 3074–3080 (2016)
26. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), vol. 12, pp. 1532–1543 (2014)
27. Shardlow, M.: The CW corpus: a new resource for evaluating the identification of complex words. In: Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pp. 69–77. Association for Computational Linguistics (2013)

28. Shardlow, M.: Out in the open: finding and categorising errors in the lexical simplification pipeline. In: Proceedings of The International Conference on Language Resources and Evaluation (LREC 2014), pp. 1583–1590 (2014)
29. Specia, L., Jauhar, S.K., Mihalcea, R.: SemEval-2012 task 1: English lexical simplification. In: Proceedings of the 1st SEM, pp. 347–355. ACL (2012)
30. Yimam, S.M., et al.: A report on the complex word identification shared task 2018. In: Proceedings of the 13th BEA. Association for Computational Linguistics (2018)
31. Yimam, S.M., Štajner, S., Riedl, M., Biemann, C.: CWIG3G2 - complex word identification task across three text genres and two user groups. In: Proceedings of the 8th IJCNLP, pp. 401–407. Asian Federation of Natural Language Processing (2017)
32. Yimam, S.M., Štajner, S., Riedl, M., Biemann, C.: Multilingual and cross-lingual complex word identification. In: Proceedings of RANLP, pp. 813–822 (2017)

Analysing Semantic Resources for Coreference Resolution

Thiago Lima[1(✉)], Sandra Collovini[1(✉)], Ana Leal[2(✉)], Evandro Fonseca[1(✉)], Xiaoxuan Han[2(✉)], Siyu Huang[2(✉)], and Renata Vieira[1(✉)]

[1] Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil
{thiago.lima.001,sandra.abreu,evandro.fonseca}@acad.pucrs.br,
renata.vieira@pucrs.br

[2] University of Macau, Macau, China
analuleal@gmail.com, hanxiaoxuan1113@gmail.com, oliviahsy316@gmail.com

Abstract. This paper compares the use of two different semantic bases, Onto.PT and ConceptNet, for the purpose of coreference resolution in the Portuguese language. These semantic bases have relations such as hyponymy and synonymy that may indicate a coreference relationship between two or more mentions. Quantitative and qualitative analyses of the impact of semantic information in coreference resolution are discussed.

Keywords: Coreference resolution · Semantic knowledge
Corpus analysis

1 Introduction

An important task in computational linguistics is coreference resolution, which is the task of finding all expressions that refer to the same entity in a text, as in the example: "France is resisting. The country is one of the first in the ranking (...)". The noun phrases [the country] and [France] are considered coreferent.

In the review of literature, most of the approaches for coreference resolution use lexical and syntactic knowledge, which are important for the task, but semantic knowledge is also necessary. In "O garoto é descoberto por uma raposa e foge. O rapaz sai a caminho das montanhas..." [The boy is discovered by a fox and runs away. The youngster heads toward the mountains...], for example, identifying a referential relation between [o garoto]$_1$ and [o rapaz]$_2$ requires semantic knowledge.

In this work we compare two different semantic bases for the task of Portuguese coreference resolution. For this, we used a coreference resolution tool for Portuguese (CORP) [6] along with OntoPT [16] and ConceptNet [23].

This work is organized as follows. In Sect. 2, we present related work. The semantic resources are detailed in Sect. 3. The coreference tool is presented in Sect. 4. The corpus used in this study is described in Sect. 5. In Sect. 6, we describe the experiments and results obtained. Finally, Sect. 7 presents some concluding remarks.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 284–293, 2018.
https://doi.org/10.1007/978-3-319-99722-3_29

2 Related Work

In the literature we find research that explore the use of semantic knowledge to improve coreference resolution systems. For English, Soon et al. [22] tested a model with a Semantic Class feature, using WordNet[1]. Strube and Ponzetto [17,18], based on [22], used Wikipedia and WordNet to disambiguate mentions and compared the results obtained with each base. Rahman and Ng [20] evaluated world knowledge using Yago [24] and FrameNet [1]. Lee et al. [12] used semantic information to identify mentions that refer to the Person category in order to resolve pronominal coreference. Hou et al. [10] proposed a rule based model to resolve unrestricted bridging using WordNet.

Specifically for Portuguese, Coreixas [14] proposed and evaluated coreference resolution methods, focusing on named entities and using semantic categories. Silva [21] proposed a coreference resolution model using semantic tags from the HAREM corpus [7]. As knowledge base they used TeP 2.0 – Electronic Thesaurus for Brazilian Portuguese [13], which stores synonymy and antonym word forms. Garcia and Gamallo in [8] proposed a model based on rules for multiple languages (Portuguese, Spanish and Galician). The authors focused solely on the Person category. Fonseca et al. [4] evaluated the impact of semantic knowledge on coreference resolution for Portuguese.

In this work, we compare the results obtained applying two different semantic bases in coreference resolution for Portuguese. Onto.PT and ConceptNet are used in this comparison. We also provide a qualitative analysis of the results found for one text in the corpus rich in semantic relations. This work was implemented in the tool CORP, further explained on Sect. 4.

Although there are studies exploring the use of semantic bases for Portuguese coreference resolution, we are not aware of any study comparing the use of different semantic bases for the task. As bases are built using different methods and sources, the comparison between the results obtained using one or another base might tell us about the adequacy of each for the task.

3 Semantic Resources

In this section we present the semantic bases for Portuguese used in this work: Onto.PT and ConceptNet.

Onto.PT is a lexical ontology for Portuguese [16]. It was built automatically using Portuguese textual resources such as OpenThesaurus.PT[2], Wikicionário.PT[3], Dicionário Aberto[4], TeP[5], OpenWN-PT[6] and the PAPEL[7]

[1] https://wordnet.princeton.edu/.

[2] https://paginas.fe.up.pt/~arocha/AED1/0607/trabalhos/thesaurus.txt.

[3] https://pt.wiktionary.org.

[4] http://www.dicionario-aberto.net.

[5] http://www.nilc.icmc.usp.br/tep2/index.htm.

[6] https://github.com/own-pt/openWordnet-PT.

[7] https://www.linguateca.pt/PAPEL/.

Table 1. Onto.PT relations

Relation	Description	Noun examples
HypernymOf	X is a superclass of Y	insect, bee
HyponymOf	X is a subclass of Y	bee, insect
MeronymOf	X is part of Y	window, house

project. Its structure is based on synsets[8], similar to WordNet, although it is not aligned with it. Synonyms can be found in the Synset itself, using it's lexical forms. For instance, Synset 21967 contains many lexical forms, including "casa", "morada", "habitação" e "domicílio", which can be seen as synonyms. There are also other relations such as hyperonymy, hyponymy, meronymy. Examples of relations between a pair of words are presented in Table 1.

ConceptNet is a freely available semantic base [23]. It originated from Open Mind Common Sense, a MIT Media Lab crowdsourcing project from 1999. Currently it is maintained by Luminoso Techonologies[9]. It has support for hundreds of languages from different families. This base uses various data sources, ranging from DBPedia[10], Wikitionary[11], Open Multilingual WordNet[12], among others. The sources are combined with word embedding models such as Word2Vec. Information is presented as a graph, wherein edges connect two terms. Each object contains an URI, for instance: `/c/en/house` is the word "house" in English. Each edge contains a relation, which can be of many types. Table 2 shows some relation types.

Table 2. ConceptNet relations

Relation URI	Description	Examples
/r/RelatedTo	The most general relation. There is some relation between X and Y	Learn ↔ erudition
/r/IsA	X is a subclass of Y (hyponym)	Car → the vehicle of Chicago → city
/r/PartOf	X is a part of Y. (meronym)	engine → car
/r/HasA	X belongs to Y	wing → bird
/r/Synonym	X and Y have very similar meanings	Sunlight → sunshine
/r/Antonym	X and Y has inverse meaning	Hot ↔ cold

[8] Groups of synonyms that have the meaning of words that contain the possible meanings for a given word in its lexical form.

[9] https://luminoso.com/.

[10] http://wiki.dbpedia.org.

[11] http://www.wiktionary.org.

[12] http://compling.hss.ntu.edu.sg/omw.

In this work, the relations of interest are hyponymy and synonymy. Onto.Pt has a larger number of these relations, as shown in Table 3. For ConceptNet only Portuguese relations were considered.

Table 3. Onto.PT and ConceptNet hyponym and synonym relations

Relation	Onto.PT	ConceptNet
Hyponym	79425	3757
Synonym	156566	13727

4 Coreference Tool and Semantic Rules

CORP is a coreference resolution tool for Portuguese. A set of rules, further explained in [2], are used to decide if two noun phrases are coreferent. CORP only deals with identity nominal coreference (proper and common nouns), independently of entity category.

Semantic relationships are extracted from each base and stored in files, using a internal representation, which CORP uses to run the semantic rules. This allows not only the use of different bases, but the option to select the desired base during runtime. Each relationship type, for each base, is stored in a separate file. When using more than one base, the tool unifies the relations.

Currently CORP has two rules regarding semantic knowledge: Synonymy and Hyponymy. Hyperonymy relations are not considered since it is more common to introduce an entity in a more specific form and, and then, in the following mentions, to use more generic terms to refer back to it. These two semantic rules are detailed below.

Hyponymy: Considering two mentions m_i and m_j, we may group them if the following conditions are satisfied:

- Head mention lemma from m_i and m_j must have a hyponymy relation;
- m_i and m_j must agree in number;
- Different modifiers in the mentions cannot occur;
- If m_i has a definite article, m_j can't have an indefinite one.

This last condition refers to a restriction on the order of articles, since it is not common, after a definite article, to use an indefinite one. For instance, it is not common to use "the car" and, after it, "a car".

Synonymy: To group two mentions based on synonymy these conditions must be satisfied:

- Head mention lemma from m_i and m_j must have a synonymy relation;
- If m_i has a definite article, m_j can't have an indefinite one;
- m_i and m_j must agree in number;
- Different modifiers in the mentions cannot occur;
- For each grouped mention in a coreference chain, it must have a synonymy relation with all other mentions in the chain.

This last condition is due to likely problems of sense ambiguity.

5 Corpus

In this work we used Corref-PT corpus for the experiments. Corref-PT[13] is a Portuguese corpus annotated with coreference information, which was created as a collective effort during IBEREVAL-2017 (Evaluation of Human Language Technologies for Iberian Languages)[14] - task "Collective Elaboration of a Coreference Annotated Corpus for Portuguese Texts" [5]. The objective of this task was to elaborate collectively a corpus with coreference annotation for Portuguese. To do so, each participant team presented texts which they considered interesting to annotate. The task involved 21 annotators from seven teams, all Portuguese native speakers.

The corpus is composed by journalistic texts, books, magazines, Wikipedia articles, among others. The corpus constitution is described in Table 4.

Table 4. Corref-PT

Corpus	Texts	Tokens	Mentions	Coreferent mentions	Coreference chains	Biggest chain	Avg. chain size
CST-News	137	54445	14680	6797	1906	25	3.6
Le-Parole	12	21607	5773	2202	573	38	3.8
Wikipedia	30	44153	12049	4973	1308	53	3.8
Fapesp magazine	3	3535	1012	496	111	33	4.5
Total	182	123740	33514	14468	3898	53	3.7

6 Experiments

We implemented a mechanism to add new semantic bases on CORP, giving the possibility to select, during runtime, which databases to use. We also implemented a mechanism to unify the relations from different databases.

Our objective was to evaluate the performance of Onto.PT and ConceptNet for the task.

In order to add a new semantic base we need to generate the relations file. In the database file available on ConceptNet site there were 791502 records, from various languages. All records that were not Portuguese-Portuguese relations were deleted, remaining 280915 records. From the remaining records, 3757 were hyponym relations and 13727 were synonym relations.

For the evaluation, we used Corref-PT corpus, described in Sect. 5, and CoNLL-Scorer, which was developed for the CoNLL 2011 Shared Task [19]. We tested three alternatives of semantic knowledge input: Onto.PT, ConceptNet and Onto.PT unified with ConceptNet.

[13] Available in http://www.inf.pucrs.br/linatural/wordpress/index.php/recursos-e-ferramentas/corref-pt/.

[14] http://sepln2017.um.es/ibereval.html.

6.1 Mention Clustering

Experiments were performed with two clustering methods, described in more detail in [3]. The first method, which we call clustering A, links one mention with its closest antecedent when at least one rule is satisfied.

The second method, which we call clustering B, is inspired on Heim's work [9] and it consists on exploring discourse representation. We assume that any mention is new in discourse if it does not have a link with the available antecedents. Thus, whenever a mention does not fire a linking rule, a new chain is generated. When there are valid rules with more than one previously introduced chain, a clustering criteria is adopted for the decision to which unique chain it will be linked.

6.2 Results Analysis

We used MUC metric to evaluate the use of different semantic bases. Tables 5 and 6 show results for MUC metric with clustering A and clustering B, respectively. We can notice that despite of a lower Recall, ConceptNet obtained a greater Precision. We can notice that, with clustering B, precision and F-measure were greater. This is due to the fact the clustering B avoids the generation of too large chains.

Table 5. Results - MUC - Clustering A

Base	Recall	Precision	F1
Onto.PT	51.9%	43.96%	47.6%
ConceptNet	49.58%	54.15%	51.77%
Onto.PT+ConceptNet	51.86%	43.85%	47.52%

Table 6. Results - MUC - Clustering B

Base	Recall	Precision	F1
Onto.PT	49.93%	54.6%	52.16%
ConceptNet	47.83%	62.89%	54.33%
Onto.PT+ConceptNet	49.96%	54.52%	52.14%

Tables 7 and 8 show CoNLL Average results with clustering A and clustering B, respectively. The CoNLL Average is obtained by the average of MUC, B^3 and $CEAF_e$ F1 values. We can notice that ConceptNet obtained the highest average.

Results with both bases are close to those with OntoPT. That can be explained by the fact that Onto.PT contains more relations than ConceptNet.

Table 7. Results - CoNLL average - Clustering A

Base	MUC	B_3	CEAF_e	CoNLL Avg
Onto.PT	47.6%	40.05%	44.93%	44.19%
ConceptNet	51.77%	45.27%	47.88%	48.31%
Onto.PT+ConceptNet	47.52%	39.94%	44.88%	44.11%

Table 8. Results - CoNLL average - Clustering B

Base	MUC	B_3	CEAF_e	CoNLL Avg
Onto.PT	52.16%	47.19%	49.29%	49.55%
ConceptNet	54.33%	48.54%	50.57%	51.15%
Onto.PT+ConceptNet	52.14%	47.18%	49.26%	49.19%

ConceptNet is multilingual, and we are using only the Portuguese relations contained in it. Moreover, many relations in ConceptNet are already contained in OntoPT. Again clustering method B had a positive impact on the results.

6.3 Qualitative Analysis

We selected one text from Corref-PT corpus with a high number of the semantic relations for a detailed analysis. This text has 34 coreference chains, from which 14 are based on semantic relations. An example of chains where lexical and syntactic rules were sufficient to identify relations between the mentions is 1. Examples 2, 3 and 4 show cases where semantic rules are required.

1. [Um estudante de artes], [O estudante][15]
2. [um pacato vilarejo], [o vilarejo], [a vila], [o seu vilarejo], [o vilarejo], [a vila], [o vilarejo], [o lugar][16]
3. [as pessegueiras], [o jardim de pessegueiras], [o jardim], [o antigo jardim], [as preciosas árvores], [as pessegueiras][17]
4. [energia nuclear], [a radiação], [as fumaças radiativas][18]

In example 2, although ConceptNet contains "vilarejo", it has no relationships with other Portuguese terms. A relationship between "vilarejo" e "lugar" was found on Onto.PT. Using clustering A, that chain was generated. However when using clustering B, the mention "lugar" was placed in the chain 5:

[15] [An arts student], [The student].

[16] [a peaceful hamlet], [the hamlet], [the village], [his hamlet], [the hamlet], [the village], [the hamlet], [the place].

[17] [the peach trees], [the peach tree garden], [the garden], [the old garden], [the precious trees], [the peach trees].

[18] [nuclear energy], [radiation], [the radioactive smoke].

5. [lugar], [vida], [arte][19]

Those chain elements are lexical forms contained in Synset 21265 in Onto.PT, therefore synonyms. Since clustering B uses a weight for each rule, and the weight for the synonymy rule is greater than for the hiponymy rule, "lugar" was included in that chain. Those mentions being synonyms, although possible, are unlikely in many cases.

Semantic relationships for the elements contained in chains 3 and 4 were not found. For "pessegueiras", a chain was formed using only lexical rules. In the case of 4, although both terms "nuclear" and "radiação" are found in Onto.PT and ConceptNet, they are not related in any of them.

When using clustering A, a large coreference chain with 39 mentions was generated grouping together many different elements, from which a fragment is shown in example 6:

6. [um garoto], [uma mulher], [casa], [uma árvore], [a casa], [uma espada curta], [uma menina][20]

This chain was generated based on the relations contained on Onto.PT between garoto and árvore, and also espada and mulher. Some Onto.PT synsets contain many relations, as an example, [árvore] synset contains 1354 hyponyms. This may be due to the automatic information extraction techniques used for its generation.

7 Conclusions

In this paper we presented experiments using two different semantic bases, aiming to compare the differences obtained with each one. We found that with ConceptNet it was possible to achieve higher precision and F-measure than Onto.PT. Through our qualitative analysis we can observe that there is noise generated by the straightforward verification of semantic relations holding between mentions. More sophisticated semantic treatment may improve the results.

Our first analysis considered two semantic bases, as future work we aim to extend our experiments to other bases. We plan to consider also BabelNet [15], Conto.PT[21] and DBPedia[22]. We also plan to explore the use of relationships weight measure, available in Conto.PT, ConceptNet and BabelNet. Another option is to explore word embeddings and information extraction [11,25], in order to build new semantic resources for Portuguese.

Acknowledgements. This paper was achieved in cooperation with Hewlett Packard Brasil LTDA. Using incentives of Brazilian Informatics Law (Law n° 8.2.48 of 1991). We thank also CNPQ, CAPES and FAPERGS for their financial support.

[19] [place], [life], [art].

[20] [a boy], [a woman], [house], [a tree], [the house], [a short sword], [a girl].

[21] http://ontopt.dei.uc.pt/index.php?sec=contopt.

[22] http://wiki.dbpedia.org/.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, ACL 1998, pp. 86–90. Association for Computational Linguistics, Montreal (1998)
2. Fonseca, E.: Resolução de Correferência Nominal Usando Semântica em Língua Portuguesa. Ph.D. thesis. PUCRS: Programa de Pós-Graduação em Ciência da Computação, PUCRS (2018)
3. Fonseca, E., Vanin, A., Vieira, R.: Mention clustering to improve portuguese semantic coreference resolution. In: Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., Meziane, F. (eds.) NLDB 2018. LNCS, vol. 10859, pp. 256–263. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91947-8_25
4. Fonseca, E., Vieira, R., Vanin, A.: Improving coreference resolution with semantic knowledge. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 213–224. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_21
5. Fonseca, E., et al.: Collective elaboration of a coreference annotated corpus for portuguese texts. In: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages, pp. 68–82 (2017)
6. Fonseca, E., et al.: CORP: Uma abordagem baseada em regras e conhecimento semântico para a resoluçao de correferências. Linguamática **9**(1), 3–18 (2017)
7. Freitas, C., et al.: Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, pp. 3630–3637 (2010)
8. Garcia, M., Gamallo, P.: An entity-centric coreference resolution system for person entities with rich linguistic information. In: COLING, pp. 741–752 (2014)
9. Heim, I.: File change semantics and the familiarity theory of definiteness. In: Semantics Critical Concepts in Linguistics, pp. 108–135 (1983)
10. Hou, Y., Markert, K., Strube, M.: A rule-based system for unrestricted bridging resolution: recognizing bridging anaphora and finding links to antecedents. In: Conference on Empirical Methods in Natural Language Processing, pp. 2082–2093 (2014)
11. Kamel, M., et al.: Extracting hypernym relations from Wikipedia disambiguation pages: comparing symbolic and machine learning approaches. In: 12th International Conference on Computational Semantics - Long papers, IWCS 2017 (2017)
12. Lee, H.: Deterministic coreference resolution based on entitycentric, precision-ranked rules. Comput. Linguist. **39**(4), 885–916 (2013)
13. Maziero, E.G., et al.: A base de dados lexical e a interface web do TeP 2.0: thesaurus eletrônico para o Português do Brasil. In: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, pp. 390–392. ACM (2008)
14. Moraes, T.C.: Resolução de correferência e Categorias de Entidades Nomeadas. M.A. thesis. Programa de Pós-Graduação em Ciência da Computação, PUCRS (2010)
15. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
16. Oliveira, H.G., Gomes, P.: ECO and Onto.PT: a flexible approach for creating a Portuguese WordNet automatically. Lang. Resour. Eval. **48**(2), 373–393 (2014)

17. Ponzetto, S.P., Strube, M.: Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 192–199. Association for Computational Linguistics (2006)
18. Ponzetto, S.P., Strube, M.: Knowledge derived from Wikipedia for computing semantic relatedness. J. Artif. Intell. Res. **30**(1), 181–212 (2007)
19. Pradhan, S., et al.: Scoring coreference partitions of predicted mentions: a reference implementation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 30–35. Association for Computational Linguistics, Baltimore (2014)
20. Rahman, A., Ng, V.: Coreference resolution with world knowledge. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT 2011, pp. 814–824. Association for Computational Linguistics, Portland (2011)
21. da Silva, J.F.: Resolução de correferência em múltiplos documentos utilizando aprendizado não supervisionado. M.A. thesis. Universidade de São Paulo (2011)
22. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Comput. Linguist. - Spec. Issue Comput. Anaphora Resolut. **27**(4), 521–544 (2001)
23. Speer, R., Havasi, C.: Representing general relational knowledge in ConceptNet 5. In: LREC 2012, pp. 3679–3686 (2012)
24. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 697–706. ACM, Banff (2007)
25. Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: Proceedings of Conference on Language Resources and Evaluation, vol. 8, pp. 1646–1652 (2008)

Annotation of a Corpus of Tweets for Sentiment Analysis

Allisfrank dos Santos(✉), Jorge Daniel Barros Júnior,
and Heloisa de Arruda Camargo

Department of Computer Science, Federal University of São Carlos (UFSCar),
Rodovia Washington Luís, km 235, 310 - SP, 13565-905 São Carlos, Brazil
allisknarf@gmail.com, jorge.barros@ufscar.br,
heloisa@dc.ufscar.br

Abstract. This article describes the process of creation and annotation of a tweets corpus for Sentiment Analysis at sentence level. The tweets were captured using the #masterchefbr hashtag, in a tool to acquire the public stream of tweets in real time and then annotated based on the six basic emotions (joy, surprise, fear, sadness, disgust, anger) commonly used in the literature. The neutral tag was adopted to annotate sentences where there was no expressed emotion. At the end of the process, the measure of disagreement between annotators reached a Kappa value of 0.42. Some experiments with the SVM algorithm (Support Vector Machine) have been performed with the objective of submitting the annotated corpus to a classification process, to better understand the Kappa value of the corpus. An accuracy of 52.9% has been obtained in the classification process when using both discordant and concordant text within the corpus.

Keywords: Annotation · Emotion · Tweets · Corpus

1 Introduction

The identification of emotions has been one of the areas of great interest in natural language processing and machine learning, in an attempt to automatically identify opinions expressed by users in texts, mainly the ones posted on websites and social networks. Sentiment analysis is the denomination for the research area of identification of emotions in textual data.

Sentiment analysis requires resources which allow the computational systems to process textual data so that their outputs represent the actual user's sentiment at the time of posting. One of the main features is the annotated textual corpus as a form of representation of the user's discourse.

This article aims to describe the annotation process of a textual corpus for Brazilian Portuguese, based on the emotions expressed in the text. The motivation to the construction of the corpus was the lack of corpus for the Portuguese language to serve as a basis for the sentiment analysis research. The specific goal in the work described here was to build a textual database for multiclass classification of textual data using the six

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 294–302, 2018.
https://doi.org/10.1007/978-3-319-99722-3_30

basic emotions (joy, surprise, fear, sadness, disgust and anger) proposed by [1] with texts extracted from social media.

In order to achieve the proposed goal, a corpus was built containing texts extracted from the Twitter platform regarding the specific domain of "MasterChef Professionals Brazil". The capture of texts occurred through the Twitter API for capturing public tweets, using the #masterchefbr hashtag as the filter for obtaining tweets from the desired domain.

This paper is organized as follows: Sect. 2 discusses papers that deal with corpus in sentiment analysis; in Sect. 3, the details of the process of construction of the corpus used in this work are presented; Sect. 4 describes with the analysis of agreement between the annotators. Section 5 discusses the experiments and results; Sect. 6 presents with the conclusions and future work perspective.

2 Corpus of Sentiment Analysis

According to [2] the literature is not extensive for papers describing the construction of corpus in Brazilian Portuguese for sentiment analysis, including the annotation process, its methodologies, the results obtained in this process, the evaluation measures and the subjectivity degree between annotators. Although a few papers were identified in the literature describing the process of constructing corpus for Portuguese, they consist basically the analysis of sentiment with annotation according to polarity (positive and negative).

In [3] the construction of a corpus of tweets regarding news comments was carried out using 3 annotators relying on manual processes. An annotated corpus was built with 850 tweets of which 425 were annotated as positive and 425 as negatives.

The work of [4] shows the ReLi (Book Review), comprehending 14 different books, in a total of 12470 sentences annotated according to the polarity at sentence level, to identify entities within the texts.

In [2] a corpus constructed by news from globo.com is presented, using 1750 texts annotated with the six basic emotions proposed by Ekman [1], each of which has 250 texts plus 250 of the neutral class.

In the context of the sentiment analysis, the work described in [5] explores an approach of sentiment analysis of journalistic texts for Brazilian Portuguese. The objective is to identify the basic emotions, using a supervised machine learning approach and the multiclass SVM to perform the task of textual classification.

In [6], the author performs the sentiment analysis orientated to aspects of the Portuguese language, exploring methods based on frequency and in machine learning, comparing experiments with corpus in Portuguese and English, annotated according to the polarity.

In the construction of a corpus, parameters such as Kappa Coefficient, which is a statistical method to evaluate the level of agreement between two or more annotators, are used to measure the degree of relevance of this corpus to serve as the basis for a computational model.

In Computational Linguistics, according to [2], his limit can vary from researcher to researcher. In the work of [7], the author says that a corpus is considered acceptable when its agreement degree reaches a Kappa Coefficient above 0.67. On the other hand, according to [8], the agreement value must be higher than 0.8, for an annotation to be considered good enough to be used. In [9], the authors state that rather than relying exclusively on the Kappa Coefficient, we must analyze factors such as the description of the annotation process, its detailing, the number of annotations, which guidelines were applied for annotation and others, in order to evaluate and accept an annotated corpus with precision.

Obtaining a level of Kappa agreement is not a trivial task when using an approach with six categories of emotion, mainly because there is no clear distinction on the boundaries between some of the emotions. This situation is not exclusive to corpus in Portuguese, as can be seen in [10] with a Kappa of 0.52 for the six emotions, in [11] with a Kappa of 0.66 for annotation of polarity and in [12] with Kappa between 0.21 and 5.1 for the corpus annotated with the six emotions.

3 Corpus Building Process

The corpus described in this article refers to the domain of MasterChef Professionals Brazil, a culinary competition program, composed of 16 participants screened from September to December 2017 by Rede Bandeirantes.

The choice of MasterChef as the domain for the building of the corpus was due to its large audience and a great interaction between the public through Twitter, thus generating a large amount of textual content for analysis.

The main motivation for the construction of the corpus was the need for annotated texts to be used in an ongoing research that aims to perform multiclass sentiment analysis, using machine learning classifiers for data stream. Due to the shortage of corpus annotated with the six basic emotions for Portuguese, it was necessary to build an annotated corpus that met this need.

The capture of the tweets occurred through an API (Application Programming Interface) provided by Twitter to capture the public stream of tweets in real time, the Streaming API. Using Streaming API it is possible to capture tweets, as well as parameters such as language, geographical location, user and creation date. To access the tool, it is necessary to have an active Twitter account, and install the application. After that, access codes are generated to work with the API.

Once the application was created, a python script was developed to link the Twitter account to the application via the generated access codes. From then on, the capture was performed using a search restriction filter. For the corpus in question, the official hashtag of the program, “#masterchefbr”, was used.

The capture of the tweets was performed on all episodes of the MasterChef Professional edition, always from the beginning of the program until one hour after the end of the program. This period has been chosen because many emotional comments can be identified after the end of the program, when users mainly express their opinion about the elimination of the day.

At the end of the data capture process, we obtained 14 files with tweets corresponding to each episode, with an average of 50000 tweets each one. In order to reduce the dimensionality of the data to be annotated due to resource constraints, such as the limited number of annotators, some measures were adopted. The first measure was to define that the tweets to be annotated would refer to the finalists of the program, since it would be possible to follow the trajectory of the participant throughout the program, and to have the perception of the feeling expressed by the users regarding the participant from the beginning to the end of the program edition. The second measure was the use of a filter, using as the keyword the name of the finalist participant, to select candidate tweets for annotation.

The annotation process was performed by 3 volunteer annotators. Each text was annotated by 2 different annotators and, in case of disagreement, a third annotator analyzed the text and assigned a label. The first step in the annotation process was to annotate 3000 randomly chosen texts after the filter application. In this phase, two volunteer annotators participated in the process and each annotator had 3 months to complete this stage. At the end of this period only 2550 were annotated twice.

The annotation was basically consisted of reading the tweets to identify the presence of one or more of an emotion, if there were, and indicate the predominant emotion (joy, surprise, fear, sadness, disgust, anger), assigning an intensity to the predominant emotion (high, medium or low) and a polarity (positive, negative). In case of no emotion or polarity in the tweet, the annotators rated this tweet as neutral.

To assist annotators in the annotation process, a file containing 30 tweets previously labeled exclusively to serve as an annotation template was provided. Along with this file, instructions on how annotation should occur, and a list of emotional words distributed among the six basic emotions were also provided. These instructions aim to level the annotators' knowledge, in an attempt to reduce subjectivity in the annotation task. A round of collective annotation was made so that possible doubts were resolved before the official rounds began. The Frame 1 shows a few examples of tweets labeled in the test round. Because they are real tweets, some of them may not obey the standard rules of the language.

After the two participants performed the annotations, the discordant tweets between annotators were identified and analyzed by a third annotator, who resolved the conflict and defined the correct label for these tweets. The next step was to analyze the annotated corpus, measuring the agreement between annotators.

Frame 1. Annotation template example.

Text	Emotion by sentence	Predominant emotion	Intensity
My greatest sadness is not that Mirna left but Francisco and Clecio stayed #MasterChefBR	Sadness	Sadness	High
I laugh at everything Pablo says is very funny #MasterChefBR	Joy	Joy	High
What sadness irina to be in the group of assholes, does not even like to twist #MasterChefBR	Sadness/disgust	Sadness	Medium

4 Degree of Agreement Between Annotators

It is necessary to use some measure to evaluate the reliability of the agreement between the annotators of the corpus, before being submitted to some algorithmic processing. In order to prove that this corpus is adequate to test and evaluate the output of a computational process.

Several measures can be used to evaluate the reliability of agreement. The correct choice of method depends on the data, the resources used in the annotation process and the expected result. Methods such as Pearson's Correlation Coefficient and Kappa Coefficient are known to ensure the reliability of a corpus. In this work, the Kappa Coefficient was used as a measure of the reliability among annotators.

Some experiments were performed after the completion of the corpus annotation. The first experiment consisted of finding the level of agreement between annotators, calculating the Kappa Coefficient. The calculated value of 0.42 is below the ones indicated as acceptable in the literature of corpus linguistics. Table 1 shows the confusion matrix of agreement between annotators.

Table 1. Matrix of confusion of concordance between annotators.

		Annotator 2						
	Emotion	**Joy**	**Sadness**	**Anger**	**Fear**	**Disgust**	**Surprise**	**Neutral**
Annotator 1	**Joy**	656	8	7	7	20	20	135
	Sadness	14	48	3	12	12	7	15
	Anger	17	6	159	3	216	9	81
	Fear	8	7	0	38	7	2	45
	Disgust	18	6	34	5	114	7	65
	Surprise	46	9	10	6	28	67	75
	Neutral	111	7	10	10	30	25	305

Another experiment aimed to verify differences between two annotations when the same annotator labels a text in different occasions, showing the amount of annotator subjectivity in the labeling process. There were 200 tweets annotated twice, without the annotator knowing which of the tweets were duplicated. Tables 2 and 3 show the confusion matrices of texts annotated twice by annotators 1 and 2, respectively.

Table 2. Matrix of confusion of texts annotated twice by annotator 1.

		2nd Time						
	Emotion	**Joy**	**Sadness**	**Anger**	**Fear**	**Disgust**	**Surprise**	**Neutral**
1st Time	**Joy**	45	0	3	1	2	3	10
	Sadness	0	10	0	1	0	0	0
	Anger	0	2	21	0	6	0	0
	Fear	0	1	0	6	1	1	1
	Disgust	1	0	9	0	10	1	1
	Surprise	5	1	4	1	1	16	2
	Neutral	8	0	2	1	4	1	18

Table 3. Matrix of confusion of texts annotated twice by annotator 2.

		2nd Time					
	Emotion	**Joy**	**Sadness**	**Anger**	**Fear**	**Disgust**	**Surprise**	**Neutral**
1st Time	**Joy**	52	0	2	0	3	0	5
	Sadness	0	5	1	0	1	0	0
	Anger	2	2	5	0	3	0	1
	Fear	2	0	0	8	0	0	0
	Disgust	1	4	2	2	26	3	3
	Surprise	1	1	1	0	2	7	3
	Neutral	6	2	1	1	3	5	34

The Kappa Coefficient found for annotator 1 was 0.54 and for annotator 2 was 0.59. These results show a high level of subjectivity in the annotation process, which contributes to the low Kappa value of a corpus, since there is no standardization in the process of annotation of emotional corpus with many classes.

After the process of extracting the necessary evaluation measures, the next step was the normalization and preprocessing of the texts that made up the corpus.

Normalization is the transformation of words that are out of the standard into words called canonical, that is, in the cultured form of a certain language. The "out of standard" words can be generated by incorrect punctuation, misspellings, acronyms, Internet slangs, and so on. Depending on the source of extraction of the texts (Web, Social Networks, evaluation sites), the author tends to have little concern about using the cultured standard of the language, requiring a standardization process to correct these problems.

In order to perform normalization of the corpus, the Enelvo tool proposed by [13] was used to normalize noisy words in user generated content written in Portuguese, identifying and normalizing spelling errors, Internet slangs, among others.

For the preprocessing of texts, a script developed in Python was used to eliminate stop-words, remove duplication of spaces and characters, as well as remove special characters, links and text accents.

5 Experiment and Results

Experiments were performed with the corpus properly normalized and preprocessed using the SVM algorithm, for textual classification of the six basic emotions.

Support Vector Machine (SVM) is a binary classifier proposed by [14], which aims to build an optimal hyperplane, so that it can separate different classes of data with a greater precision margin. These hyperplanes are constructed by small subsets of training data, also called support vectors. The SVM proved to be quite promising in textual classification [15]. However, several classification problems have more than two classes, and to solve this problem, we can use the strategy of combining classifiers generated in binary subproblems as shown in [16].

The training and testing of SVM classifier was done using Weka, a tool that provides a collection of machine learning algorithms for data mining tasks [17], using the standard parameters of the tool and with cross-validation of 10 folds.

The evaluation measures considered in the experiment were accuracy, precision, recall and F-measure.

Table 4 shows the results of experiments for the classification of 1163 discordant texts, submitted to the method of identification of emotions.

Table 4. Emotion classification of discordant texts.

Emotion	Joy	Sadness	Anger	Disgust	Fear	Surprise	Neutral	Precision	Recall	F-Measure
Joy	65	4	4	45	2	5	74	0.38	0.32	0.35
Sadness	13	1	2	16	0	2	19	0.05	0.01	0.02
Anger	12	0	31	63	0	6	27	0.44	0.22	0.29
Disgust	28	5	22	158	0	11	75	0.36	0.52	0.42
Fear	6	1	1	18	0	2	24	0	0	0
Surprise	19	4	7	48	0	9	46	0.22	0.69	0.10
Neutral	29	2	3	90	2	5	160	0.37	0.55	0.44
Accuracy: 36.4%										

Table 5. Emotion classification of concordant texts.

Emotion	Joy	Sadness	Anger	Disgust	Fear	Surprise	Neutral	Precision	Recall	F-Measure
Joy	534	3	10	11	4	8	86	0.71	0.81	0.75
Sadness	11	12	2	4	1	1	17	0.70	0.25	0.36
Anger	39	1	67	15	1	0	36	0.62	0.42	0.5
Disgust	39	1	12	24	1	0	37	0.32	0.21	0.25
Fear	16	0	2	5	7	0	8	0.50	0.18	0.26
Surprise	29	0	6	7	0	6	19	0.31	0.09	0.14
Neutral	84	0	9	8	0	4	200	0.49	0.65	0.56
Accuracy: 61.2%										

On the experiment presented in Table 5, a set of 1387 text with full concordance among annotators were submitted to emotions classification method, reaching a 61.2% success rate, against the 36.4% success rate for discordant texts. Therefore, it is noticeable that the sentiment analysis method trained with a set of texts with full concordance between annotators shows a superior performance when compared to the same method trained with a set of discordant texts.

Another experiment was performed, mixing discordant and concordant texts, with the objective of understanding the method's behavior when there is data of both sorts within the corpus. The results are shown in Table 6.

Table 6. Evaluation measures using discordant and concordant texts within the corpus.

Emotion	Joy	Sadness	Anger	Disgust	Fear	Surprise	Neutral	Precision	Recall	F-Measure
Joy	630	6	11	47	0	11	150	0.61	0.73	0.67
Sadness	31	24	2	15	0	4	25	0.70	0.23	0.35
Anger	50	1	110	66	1	11	59	0.61	0.39	0.46
Disgust	99	1	35	157	0	13	108	0.39	0.38	0.38
Fear	28	0	2	18	13	3	26	0.92	0.14	0.25
Surprise	65	1	9	31	0	23	68	0.35	0.11	0.17
Neutral	123	1	9	61	0	8	394	0.47	0.66	0.55
Accuracy: 52.9%										

The accuracy, the measure that calculates the percentage of examples that were correctly classified by the classifier, was 52.9%. When analyzing the accuracy percentages obtained in Tables 4, 5 and 6, it can be seen that texts with total agreement have a better performance when compared to the complete set of texts, making it evident that a higher concordance rate between the annotators has a direct impact in classifier learning.

Analyzing the accuracy obtained by the classifier in Table 6, we notice a disproportion between emotions, especially in “disgust” and “surprise”, showing a low rate of accuracy when compared to the other emotions analyzed. One can attribute this result to the lack of a clear boundary definition for such emotions, causing an overlap between their limits and, therefore, leading to a misclassification in the cases that would fall under either disgust or surprise.

6 Conclusions and Future Work

The purpose of this article was to describe the process of building a corpus of annotated tweets following Ekman’s six basic emotions. The experiments performed show a low degree of agreement between the annotators.

However, the results help us understand the great mismatch between emotions, which in many cases is justified by the lack of clear boundaries between emotions, as observed for “disgust” and “surprise” described earlier. This fact leads us to believe that there may be words and/or expressions invoking both emotions at once. Such lack of clear boundaries reflected in the concordance between the annotators in the analysis of some tweets.

Even with Kappa Coefficient results below the limits of acceptance proposed by some authors, the values obtained in this work are reasonable when compared to articles that present similar works. Thus, the main contribution of this work is the construction of a corpus annotated at sentence level based on the six emotions of Ekman, which will be available for other studies of sentiment analysis in Brazilian Portuguese.

For future work, the objective is to expand the number of corpus annotators in an attempt to obtain a higher Kappa Coefficient, and insert of the irony class, since many of the texts can be considered ironic.

References

1. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(3–4), 169–200 (1992)
2. Dosciatti, M.M., Ferreira, L.P.C., Paraiso, E.C.: Anotando um Corpus de Notícias para a Análide de Sentimento: Um relato de experiência. In: Proceedings of Symposium in Information and Human Language Technology (STIL), pp. 121–130 (2015)
3. Nascimento, P., et al.: Análise de sentimento de tweets com foco em notícias. In: Brazilian Workshop on Social Network Analysis and Mining (2012)
4. Freitas, C., Motta, E., Milidiú, R.L., César, J.: Sparkling vampire… lol! annotating opinions in a book review corpus. In: Aluísio, S., Tagnin, S.E.O. (eds.) New Language Technologies and Linguistic Research: A Two-Way Road, pp. 128–146. Cambridge Scholars Publishing (2014)
5. Dosciatti, M.M.: Um Método para Identificação de Emoções Básicas em Textos em Português do Brasil Usando Máquinas de Vetores de Suporte em Solução Multiclasse. 2015. Tese(Doutorado em Informática) – Instituto de Informática, Pontifícia Universidade Católica do Paraná, Paraná (2015)
6. Balage Filho, P.P.: Aspect extraction in sentiment analysis for portuguese language. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos (2017)
7. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, Chap. 12. Sage, Beverly Hills (1980)
8. Artstein, R., Poesio, M.: Bias decreases in proportion to the number of annotators. In: Proceedings of FG-MoL 2005, Edinburgh, pp. 141–150 (2005)
9. Di Eugenio, B., Glass, M.: The kappa statistic: a second look. Comput. Linguist. **30**(1), 95–101 (2004)
10. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the 2008 ACM Symposium on Applied Computing, pp. 1556–1560. ACM (2008)
11. Habernal, I., Ptácek, T., Steinberger, J.: Supervised sentiment analysis in Czech social media. Inf. Process. Manag. **50**(5), 693–707 (2014)
12. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of Human Language Technology Conference on Empirical Methods in Natural Language Processing (2005)
13. Bertaglia, T.F.C., das Graças Volpe Nunes, M.: Exploring word embeddings for unsupervised textual user-generated content normalization. In: Proceedings of the 2nd Workshop on Noisy User-Generated Text (WNUT) (2016)
14. Panik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to the probabilities. Theory Probab. Appl. **16**(2), 283–305 (1971)
15. Alves, A.L.F., Baptista, C.S., Firmino, A.A., Oliveira, M.G., Paiva, A.C.: A comparison of SVM versus Naive-Bayes techniques for sentiment analysis in tweets: a case study with the 2013 FIFA confederations cup. In: Proceedings of the 20th Brazilian Symposium on Multimedia and the Web, WebMedia 2014, pp. 123–130. ACM, New York (2014)
16. Facelli, K., Lorena, A.C., Gama, J., Carcalho, A.C.P.L.F.: Inteligência Artificial – Uma abordagem de Aprendizado de Máquina. LTC. Rio de Janeiro (2011)
17. https://www.cs.waikato.ac.nz/ml/weka/. Accessed 09 Apr 2018

SICK-BR: A Portuguese Corpus for Inference

Livy Real[1(✉)], Ana Rodrigues[1], Andressa Vieira e Silva[1], Beatriz Albiero[1], Bruna Thalenberg[1], Bruno Guide[1], Cindy Silva[1], Guilherme de Oliveira Lima[1], Igor C. S. Câmara[2], Miloš Stanojević[3], Rodrigo Souza[1], and Valeria de Paiva[4]

[1] University of São Paulo, São Paulo, Brazil
livyreal@gmail.com, bruna@ime.usp.br
{ana2.rodrigues,andressa.vieira.silva,beatriz.albiero,bruno.guide, cindy.silva,guilherme.oliveira.lima,rodrigo.aparecido.souza}@usp.br
[2] University of Campinas, Campinas, Brazil
igor0csc@gmail.com
[3] University of Edinburgh, Edinburgh, UK
m.stanojevic@ed.ac.uk
[4] Nuance Communications, Burlington, USA
valeria.depaiva@gmail.com

Abstract. We describe SICK-BR, a Brazilian Portuguese corpus annotated with inference relations and semantic relatedness between pairs of sentences. SICK-BR is a translation and adaptation of the original SICK, a corpus of English sentences used in several semantic evaluations. SICK-BR consists of around 10k sentence pairs annotated for neutral/contradiction/entailment relations and for semantic relatedness, using a 5 point scale. Here we describe the strategies used for the adaptation of SICK, which preserve its original inference and relatedness relation labels in the SICK-BR Portuguese version. We also discuss some issues with the original corpus and how we might deal with them.

Keywords: Portuguese · Open corpus · Textual inference · NLI
Semantic relatedness

1 Introduction

Determining semantic relationships between sentences is essential for machines that understand and reason with natural language. While the task of detecting Natural Language Inference (NLI) may be considered implicit in all the work done in Natural Language Semantics, one could argue that the strategies currently used for NLI are superficial and unable to deal with the real problem: reasoning with semantic content. We are working towards reasoning with Portuguese texts and have been working for a while on different strategies to obtain open systems that can help with processing Portuguese [1–3]. Inference can be seen as one of the most basic tasks for semantic reasoning. Our long term

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 303–312, 2018.
https://doi.org/10.1007/978-3-319-99722-3_31

research goal is the development of symbolic and statistics based approaches that can deal with inference in Portuguese.

Much work has been done for inference in English, symbolic or otherwise, and not so much for Portuguese. The initial Recognizing Textual Entailment (RTE)[1] challenges (from 2005 to 2013) have drawn attention to the problems of detecting inference. Nowadays, with the success of deep learning techniques, the inference tasks are again in the spotlight for Natural Language Processing, now under the label of Natural Language Inference (NLI). Large datasets have been constructed to serve as supervised data for systems that want to learn to perform inference e.g. amongst others SICK [4], SNLI [5], MultiNLI [6]. However, it is not clear how trustworthy these datasets are, how much they codify the flexible human intuitions of inference. Also, it is still debatable how these datasets should be constructed, since it has been shown [7] that neural systems can be largely influenced by the way the training corpus is assembled. This makes it possible, for example, that a system assigns an inference relation between two sentences only by looking at the first sentence, thereby missing completely the point of the inference. For this and other reasons, it is particularly important to have different corpora available to train and test systems that detect inference in different languages.

1.1 Related Work

There is already one Portuguese corpus for textual inference publicly available, the ASSIN corpus [8]. This was released for a shared task[2], associated with the conference PROPOR2016. The tasks in the competition were both Semantic Similarity and Textual Entailment. Competitors could chose to participate in only one of the tasks. Only 4 of the 6 teams participating in the task decided to work on textual entailment, while all the 6 teams worked on semantic similarity.

ASSIN was the first attempt to gather the Portuguese NLP community to discuss textual entailment and semantic similarity. However, since the ASSIN corpus was built over international news, there is a large amount of temporal expressions, named entities and other complex phenomena that make the reasoning over the sentences very difficult. No system could do better than the offered baselines. This led the ASSIN creators to suggest that maybe a simpler corpus for NLI, organized in the style of SICK [4], with less subjectivity was needed [8,9]. Thus, in some ways, this work can be seen as a continuation of the ASSIN work.

However, we can also see problems with the ASSIN corpus. The ASSIN corpus is annotated for entailment, paraphrase and neutral relations. We agree with [10] and [11] that *contradictions* are an essential component of human reasoning, so we want to have a corpus which annotates contradictions too. The ASSIN corpus has also the label 'paraphrase' and paraphrases are also entailments, so one could say that these two ASSIN categories overlap with each other. This is something

[1] https://aclweb.org/aclwiki/Recognizing_Textual_Entailment.

[2] http://propor2016.di.fc.ul.pt/?page_id=381.

that we would like to avoid. A corpus with a restricted scope of the semantic and syntactic phenomena it intends to tackle seems a sensible idea and SICK is exactly that.

1.2 Our Goal

Given our long term goal of doing semantic reasoning in Portuguese and the sensible suggestion of the need for a simplified corpus, we decided to take on the task of building a NLI corpus in Portuguese.

The corpus SICK concentrates on compositional phenomena and it is annotated with the kind of inference we want to be able to distinguish. The SICK construction process, coming up from picture captions, restricts its scope to concrete actions and scenes. Moreover some care was taken to restrict the amount of world knowledge required to perform the intended inferences. Finally, because the inferences aimed at were fairly basic, they held the promise of common sense reasoning, which we believe is somewhat universal.

Finally there were practical considerations. Building a resource like SICK is very time and money consuming. Bootstrapping the creation of a Portuguese corpus via automatic translating and adapting SICK for Portuguese, giving rise to SICK-BR, seemed the easier route. This approach has also the added value of producing a parallel resource, since the pairs of SICK and SICK-BR are aligned and offer the same labels, both for relatedness and inference relations.

2 SICK

The corpus SICK[3] was conceived to provide a benchmark for compositional distributional semantic models [4]. The corpus SICK consists of English sentence pairs annotated to account for inference relations (entailment, contradiction and neutral) and relatedness (on a 5-point rating scale). The corpus is simplified in aspects of language processing not fundamentally related to composionality: there are no named entities, the tenses have been simplified to the progressive, there are few modifiers, few compounds, few pronouns, etc. The data set consists of 9840 English sentence pairs (composed from some 6k unique sentences), generated from existing sets of captions of pictures.

The authors of SICK randomly selected a subset from the caption sources and applied a 3-step generation process to obtain the pairs. After a normalization phase, when undesirable phenomena were excluded or re-written, desirable phenomena were added, as negation and active/passive alternation. After these generation steps, the sentences in pairs were sent to Amazon 'mechanical turkers' who annotated them for inference and relatedness.

Inference annotation led to 5595 neutral pairs, 1424 contradiction pairs, and 2821 entailment pairs, hence 4245 informative pairs in total. SICK was the resource used in the SemEval 2014 task 1.

[3] http://clic.cimec.unitn.it/composes/sick.html.

3 SICK-BR

We bootstrapped the creation of a simplified corpus for NLI in Portuguese by making use of the human annotations in the original SICK. We start from a basic machine translation of the corpus. We want to be sure that our translated pairs get exactly the same truth-conditional semantics as the original ones. We also want to have, as much as possible, the same kind of linguistic phenomena that SICK discusses. Another parallel goal is to keep the relatedness between the paired sentences, which imposes challenges on lexical choices. Here we explain some of our strategies to keep the translations of SICK-BR as close to the original SICK as possible and we describe the phases of the construction of SICK-BR. The process of building SICK-BR had the following phases: 1. pre-processing and machine translation; 2. guidelines creation, training and translation checking; 3. post-processing and reconstruction; 4. label checking. We discuss each of these steps in this section.

3.1 Pre-processing and Machine Translation

Firstly, we got all the (6076) unique sentences that are part of the 9480 SICK pairs and translated them to Portuguese using a state-of-the-art online tool. As expected the output of the automatic translation is full of mistakes. For example, in SICK, the most used verb, apart from the verb *to be*, is the verb *to play.* This needs to be translated by different verbs in Portuguese *tocar/play an instrument, brincar/play with other kids* and *jogar/play sports*, etc. So we expected this to be difficult for machine translation, and it was. We also found many spelling mistakes in the translation, such as *estao* or *estáo* for *estão*, which is easy to correct, but that can cause trouble when processing the corpus, since most of the systems would just not recognize these misspelled forms.

3.2 Guidelines, Training and Checking

As discussed in [12,13], many mistakes in SICK are due to the lack of clear guidelines for annotators and to the fact that they did not have linguistic training. To avoid introducing mistakes in the corpus and to try to ensure the quality of SICK-BR, concerted effort was put in this phase.

Since we worked on a translation to produce a new corpus, SICK-BR, which should keep the same labels (for relatedness and inference) from the previous one, SICK, we call here 'annotators' the linguists that worked in the translation, rather than the people who actually annotated the original labels of the pairs. Also, we call 'guidelines' the instructions to be used for translation checking, rather than instructions to actually label the relations within a given pair. In this translation phase, ten annotators took part in the work, all of them native Brazilian Portuguese speakers, proficient in English and all have linguistic training.

Once we had an automated machine translation of the corpus, two of us selected 55 sentences that showcased the intended linguistic phenomena in SICK and also other phenomena that may be difficult to translate from English. These sentences were given to the 10 annotators without the machine translation and, after a detailed discussion, an agreement was reached on how to translate these sentences. We then compared these 'golden' translations to the ones produced automatically and got some insights on where the machine translation system systematically goes wrong.

Considering our main goals—(i) keep the inference labels of SICK, (ii) the relatedness labels and (iii) having a naturally sounding corpus in Portuguese—and the results of this initial discussion, we reached our main guidelines:

- 1. Translated sentences should keep the same truth values as the original sentences;
- 2. We try to maintain, over the Portuguese corpus, the same lexical choices for the same English expressions within reason;
- 3. We keep, as much as possible, the same phenomena that we believe the original sentence was showcasing;
- 4. We keep naturally sounding Portuguese sentences, as much as possible;
- 5. We keep word alignment, whenever possible.

The guidelines are to be followed in this order, which tells us that keeping the same labels as the sentences have in SICK is more important than to keep the naturalness of the Portuguese sentences, for example. Although we tried to not have sentences that sound odd our main goal was to keep the labels aligned.

We also produced and updated during the project a glossary[4] for the most used terms, such as the many multiword expressions (MWEs) we found in SICK, despite the original SICK creators efforts to not have any MWEs. This might be useful to scholars interested in MWE and named entities recognition in Portuguese, since these choices are informative.

The 6076 unique sentences of the corpus were equally distributed among the 10 annotators. We used an online platform for the checking. This made it possible for annotators to look at each others' work when translating their sentences. We also kept an online forum for discussing issues, where more than 2k messages were exchanged during this work. The glossary was always updated when a solution was reached. Each annotator could also mark out complex sentences that they thought needed further review. Differently from other corpora creation processes we know about, our annotators could always say they were not able to annotate something, an easy strategy that helps to ensure the quality of the work. Finally, an experienced annotator double checked all sentences considered complex and proposed a final translation for these sentences.

Rethinking Our Steps. During this phase, we realized that one of our previous goals was not reachable. We would like to have SICK and SICK-BR aligned as parallel corpora at a **sentence level**. This would mean that each sentence of SICK would be translated in SICK-BR by one sentence. However, some translation issues showed us that it was an impossible goal.

[4] Available at https://github.com/livyreal/SICK-BR/tree/master/Glossary.

One third of the pairs in SICK differ by only one or two words, for example the pair $A =$ *Kids in red shirts are playing in the leaves.* $B =$ *Children in red shirts are playing in the leaves.* Since we could not 'perfectly' translate this pair of sentences into Portuguese (Portuguese has no two words for 'child' that have no (ontological) gender attached) keeping exactly the same referent, we would not keep the original NLI labels. We looked at these pairs[5], called one-word-apart pairs, and, for most of them, we could find words in Portuguese that kept the same truth value for the sentences.

However, the scenario changes when we consider pairs as *kid* and *child*. We have many words for child in Portuguese, but most of them have (ontological) gender attached to them, only *criança* can be used for boys and girls. Considering the many pairs in SICK based on *child/kid* difference, we could not translate both of them to *criança*, without ending up with sentences that were literally the same. Also in SICK, there are many pairs based on the difference between *kid/child* and gender specific words such as *boy/girl*, therefore translating *kid/child* for pairs as *garoto/menino* would not solve the problem, but would rather create a new one. Because of that, we decided to have a new step in our corpus building. We translated both *kid* and *child* by *criança* and had a new step to make sure there is no sentence pair in the corpus with exactly the same sentence repeated. In the sentence translation phase, so, both $A =$ *Kids in red shirts are playing in the leaves.* $B =$ *Children in red shirts are playing in the leaves.* were translated by *Crianças de camisas vermelhas estão brincando nas folhas.* After in the corpus construction, these exactly-the-same pairs were reanalyzed and re-translated by: $A =$ *Meninos de camisas vermelhas estão brincando nas folhas.* $B =$ *Garotos de camisas vermelhas estão brincando nas folhas.* With this solution, we still keep the inference label for the pair (A_entails_B, B_entails_A).

These choices had two main consequences: we needed a new phase of corpus construction; and we did not have a corpus aligned at a sentence level. However, we still have a corpus aligned to the original corpus SICK at the **level of paired sentences**. The new corpus keeps exactly the same labels for the intended tasks and these are possible to trace since the `id` pair in SICK-BR is the same as the one in SICK. The pairs in SICK and SICK-BR are aligned and have the same labels, but one sentence in SICK may be translated by more than one sentence in SICK-BR. Therefore, SICK-BR has the same amount of pairs as SICK, but SICK-BR has a slightly bigger number of sentences than SICK.

3.3 Post-processing and Reconstruction

This phase was concerned with making sure we were not introducing new mistakes to SICK-BR. We ran a state-of-the-art speller and grammar checkers on all the 6k unique sentences. We also made sure that we had no extra spaces and final periods in the sentences. Although SICK pairs in general do not have any punctuation, a few sentences still have it and for people interested in syntactic

[5] We thank Katerina Kalouli for the processing of original SICK, made public available in https://github.com/kkalouli/SICK-processing.

parsing having punctuation or not in a sentence can change the parsing. We then used the glossary we prepared for the annotators, checking to make sure no one missed an agreed lexical choice during the translation phase.

Finally, the corpus was reconstructed: the sentences were paired as the original ones and the original labels were assigned to the Portuguese pairs. We then reviewed all the 'same sentences' pairs.

Bellow, one example of an entry in SICK-BR:

```
580 | Um grupo de meninos está brincando com uma bola em frente
a uma porta grande feita de madeira | Um monte de meninos está
brincando com uma bola em frente a uma porta grande feita de
madeira | ENTAILMENT | 4.9 | A_entails_B | B_entails_A | A group
of boys are playing with a ball in front of a large door made of
wood | A bunch of boys are playing with a ball in front of a large
door made of wood | FLICKR | FLICKR | TEST
```

The first field (1) is the ID pair. The next two fields (2, 3) are the human proposed version of the original sentences. The four following fields (4, 5, 6, 7) are the original SICK labels we reused in our corpus. The next two fields (8, 9) are the original SICK pair. The following two fields (10, 11) indicate the dataset where the original sentences in SICK came from. Finally, the last field (12) indicates the set of SEMEVAL 2014 dataset split this pair was part of.

3.4 Checking Labels

We then verified how well the original SICK labels fitted our translated pairs. We checked 400 labels for relatedness and 800 labels for inference relations, chosen randomly but equally distributed between the different label types. This step showed that we do not always agree with the original SICK labels. The annotation for 'semantic relatedness' is especially problematic. This is a subtle classification, that was presented by the original SICK annotators only through examples, therefore the labels are not always consistent. Since our goal was not to re-annotate SICK, but rather to think of strategies that would keep the original human annotation, the 'mistakes' we found in SICK labels are also present in SICK-BR.

The lack of clear guidelines on what would be considered a related pair made impossible for us checking the relatedness scores without considering the original English pair. We compared the relatedness score of the Portuguese pairs checking the pairs in English in parallel. We found that when a certain score was given to an English pair, this score still holds in Portuguese. Since relatedness is a so subtle phenomenon, very difficult to annotate, only huge discrepancies would be considered mistakes and, over 400 checked labels, we didn't find any. Although SICK-BR is a translation of SICK, that could make the relatedness between the sentences different, the fact that SICK is a simplified corpus and that we kept as much as possible the same lexical choices over the whole corpus, made feasible the reuse of semantic relatedness scores from SICK to SICK-BR. Despite of the fact that 100% of the checked relatedness scores were reasonable when applied to the Portuguese pairs, we recall that this annotation is not (in both languages) as reliable as we would like it to be.

Checking the consistency of these labels made us realize some new issues with the original corpus. For example, the sentences *A woman is not riding a horse./A woman is riding a horse* are part of two pairs with different `ids`. So in SICK, we have both the pair *id = 4305 A = A woman is not riding a horse. B = A woman is riding a horse* and the pair *id = 4587 A = A woman is riding a horse. B = A woman is not riding a horse.* Since all the pairs were annotated for inference in both directions (whether A entails B and also B entails A), it does not make sense to have repeated pairs. The situation gets worse when we consider that these two pairs have different labels for relatedness in SICK: while the first pair has a 4.5 relatedness score, the second one is scored as only 3.8. This clearly shows how the relatedness score is subjective and debatable.

We also found some inconsistency on inference labels as [12,14] have already shown. From the 800 pairs checked for inference, 20 do not hold for Portuguese. All these 20 pairs are labeled as ENTAILMENT. From our analysis (double checked by two native English speakers), 14 of those 20 inference labels were already wrong in SICK. As pointed in [12] some sentences in SICK are non-sensical or ungrammatical. For example *A motorcycle is riding standing up on the seat of the vehicle* or *The players is maneuvering for the ball*[6]. The other six pairs are debatable labels even in English. It happens that one of the sentences is ambiguous (as *The kid is still in the snow.*) or that the entailment among the pairs is not obvious but possible (is a shore always by the beach? Is a lady a girl?).

In SICK-BR, we corrected all the ungrammatical sentences (18 sentences), but we do not correct (35) non-sensical sentences since correcting them would mean radically changing their interpretation. We listed 35 sentences as non-sensical, such as *A woman is bowling two eggs to a break dancer* and *A man is pouring a pot of cheese sauce into a shredded plate.* It seems that these sentences are the result of the expansion phase of the SICK creation process. They were created by scrambling the original words. Although this way of generating new sentences might have seemed a good idea, it created a lot of noise in SICK. Almost always these non-sensical and ungrammatical sentences are part of pairs that were labeled as neutral for inference, suggesting that when the annotators could not judge the sentences, they just annotated them as neutral.

4 Results

The Portuguese pairs of SICK-BR can be downloaded in https://github.com/livyreal/SICK-BR. SICK-BR has the sentences in Portuguese and keeps the same identifiers `id` and labels for inference and relatedness as the original corpus.

Our hypotheses that the logical phenomena in both languages would be similar and that entailment and contradiction relations between sentences would

[6] These analyses are available in https://github.com/livyreal/SICK-BR.

work the same way both in English and in Portuguese have been mostly confirmed. From 800 inference labels checked, we disagree on only 20 in SICK-BR: 14 of them were already wrong in SICK and the other 6 are somehow debatable in the original resources as well. Considering 400 relatedness score, we confirmed that, for all the checked pairs, if the English label was reasonable, ours was also reasonable. This makes SICK-BR as reliable as SICK for the relatedness task.

Of course, this translation, preserving relations, was possible because of the simplification of data that SICK aimed for.

Some of the issues found in SICK are also present in our corpus. We still have, for example, sentences that are not common sense such as *Um hamster está cantando* (translated from *A hamster is singing*). However, given the need to manually verify all the translations, we have managed to correct non-grammatical sentences, sentences with smaller typos and such like. We have decided to keep the sentences lacking commonsense, to keep the parallelism between the corpora. Many of the goals stated by the SICK creators were not really fully realized. For instance, not all sentences are in the progressive. We found around 90 sentences that were not in progressive, such as *A topless boy has a clean face.* To preserve as much as possible the original label assignments, we also kept some of the ambiguity and bias from the original corpus.

SICK-BR is more uniform than SICK as far as punctuation goes. We also corrected some spelling and processing mistakes. SICK has sentences, such as *A black dog is jumping from **n** hay ball to another hay ball* (should *n* be *an*?). It also has sentences such as *The man is not adding seasoning to **the/some** water in a bowl* and *A piece of bread, which is big, is having butter spread upon it by a man **OR A piece of bread, which is big, is being spread with butter by a man**.* In these cases, it seems that some steps of the construction phase were messy and left behind the choice markers used by the SICK creators. For all these cases, we have a single and grammatical sentence in SICK-BR.

5 Conclusions and Future Work

We described the construction of a Natural Language Inference (NLI) corpus for Portuguese, SICK-BR, which is based on and aligned to the English corpus SICK. We focused on linguistic strategies to guarantee (i) the reuse of the original NLI and relatedness labels of SICK into SICK-BR; (ii) a natural register of Portuguese and (iii) the existence and discussion of the same linguistic phenomena found in SICK. The issues found with the labels in SICK-BR were almost all already found in the original SICK. Due to some specificities of the languages involved, it was impossible to keep SICK-BR aligned to SICK at the sentence level, instead we have SICK and SICK-BR aligned at the pair level. We leave to future work the investigation of different approaches to automatically detecting inference relations in SICK-BR. Concentrating on SICK-BR, we would like to make sure that existing lexical resources for Portuguese, such as OpenWordNet-PT [1], are capable of dealing with the information in SICK-BR. Finally, we would like to investigate the phenomena of implicatives and factives in Portuguese, following up on the work of Kartunnen and others [15,16].

References

1. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: an open Brazilian WordNet for reasoning. In: COLING 2012: Demonstration Papers (2012)
2. de Paiva, V., Real, L., Rademaker, A., de Melo, G.: NomLex-PT: a lexicon of Portuguese nominalizations. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, May 2014
3. Real, L., Rademaker, A., Chalub, F., de Paiva, V.: Towards temporal reasoning in Portuguese. In: LREC2018 Workshop Linked Data in Linguistics (2018)
4. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A SICK cure for the evaluation of compositional distributional semantic models. In: Proceedings of LREC 2014 (2014)
5. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326 (2015)
6. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv (2017). http://arxiv.org/abs/1704.05426
7. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. CoRR abs/1803.02324 (2018). http://arxiv.org/abs/1803.02324
8. Fonseca, E., Borges dos Santos, L., Criscuolo, M., Aluisio, S.: Visao geral da avaliacao de similaridade semantica e inferencia textual. Linguamatica **8**(2) (2016)
9. Fonseca, E.R.: Reconhecimento de implicação textual em português. Ph.D. thesis, ICMC-USP (2018)
10. Condoravdi, C., Crouch, D., De Paiva, V., Stolle, R., Bobrow, D.: Entailment, intensionality and text understanding. In: HLT-NAACL 2003 Workshop on Text Meaning (2003)
11. de Marneffe, M.C., Rafferty, A.N., Manning, C.D.: Finding contradictions in text. In: Proceedings of ACL 2008 (2008)
12. Kalouli, A.L., Real, L., de Paiva, V.: Textual inference: getting logic from humans. In: Proceedings of the 12th International Conference on Computational Semantics (IWCS) (2017)
13. Kalouli, A.L., Real, L., De Paiva, V.: Annotating logic inference pitfalls. In: Workshop on Data Provenance and Annotation in Computational Linguistics (2018)
14. Kalouli, A.L., Real, L., de Paiva, V.: Correcting contradictions. In: Proceedings of Computing Natural Language Inference (CONLI) Workshop (2017)
15. de Melo, G., de Paiva, V.: Sense-specific implicative commitments. In: Gelbukh, A. (ed.) CICLing 2014. LNCS, vol. 8403, pp. 391–402. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54906-9_32
16. Nairn, R., Condoravdi, C., Karttunen, L.: Computing relative polarity for textual inference. In: Inference in Computational Semantics (ICoS-5), pp. 20–21 (2006)

LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text

Pedro Henrique Luz de Araujo[1](✉), Teófilo E. de Campos[2],
Renato R. R. de Oliveira[1], Matheus Stauffer[1], Samuel Couto[1],
and Paulo Bermejo[1]

[1] R&D Center for Excellence and Public Sector Transformation - NEXT,
Universidade de Brasília - UnB, Brasília, Brazil
pedrohluzaraujo@gmail.com,
{renatooliveiraz,matheusstauffer,samuelcouto,paulobermejo}@next.unb.br

[2] Department of Computer Science, University of Brasília, Brasília, DF, Brazil
t.decampos@st-annes.oxon.org
http://www.cic.unb.br/~teodecampos/LeNER-Br/

Abstract. Named entity recognition systems have the untapped potential to extract information from legal documents, which can improve information retrieval and decision-making processes. In this paper, a dataset for named entity recognition in Brazilian legal documents is presented. Unlike other Portuguese language datasets, this dataset is composed entirely of legal documents. In addition to tags for persons, locations, time entities and organizations, the dataset contains specific tags for law and legal cases entities. To establish a set of baseline results, we first performed experiments on another Portuguese dataset: Paramopama. This evaluation demonstrate that LSTM-CRF gives results that are significantly better than those previously reported. We then retrained LSTM-CRF, on our dataset and obtained F_1 scores of 97.04% and 88.82% for Legislation and Legal case entities, respectively. These results show the viability of the proposed dataset for legal applications.

Keywords: Named entity recognition · Natural language processing
Portuguese processing

1 Introduction

Named entity recognition (NER) is the process of finding, extracting and classifying named entities in natural language texts. Named entities are objects that can be designated by a proper noun and fit predefined classes such as persons, locations and organizations. In addition, the NER community has found useful to include temporal and numeric expressions (e.g. dates and monetary values) as named entities [18].

The state-of-the-art entity recognition systems [13,14] are based on Machine Learning techniques, employing statistical models that need to be trained on a

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 313–323, 2018.
https://doi.org/10.1007/978-3-319-99722-3_32

large amount of labeled data to achieve good performance and generalization capabilities [15]. The process of labeling data is expensive and time consuming since the best corpora are manually tagged by humans.

There are few manually annotated corpora in Portuguese. Some examples are the first and second HAREM [5,22] and Paramopama [17]. Another approach is to automatically tag a corpus, like the one proposed in [19] that originated the WikiNER corpus. Such datasets have lower quality than manually tagged ones, as they do not take into consideration sentence context, which can result in inconsistencies between named entity categories [17].

An area that can potentially leverage the information extraction capabilities of NER is the judiciary. The identification and classification of named entities in legal texts, with the inclusion of juridical categories, enable applications such as providing links to cited laws and legal cases and clustering of similar documents.

There are some issues that discourage the use of models trained on existing Portuguese corpora for legal text processing. Foremost, legal documents have some idiosyncrasies regarding capitalization, punctuation and structure. This particularity can be exemplified by the excerpts below:

> EMENTA: APELAÇÃO CÍVEL - AÇÃO DE INDENIZAÇÃO POR DANOS MORAIS - PRELIMINAR - ARGUIDA PELO MINISTÉRIO PÚBLICO EM GRAU RECURSAL - NULIDADE - AUSÊNCIA DE INTERVENÇÃO DO PARQUET NA INSTÂNCIA A QUO - PRESENÇA DE INCAPAZ - PREJUÍZO EXISTENTE - PRELIMINAR ACOLHIDA - NULIDADE RECONHECIDA.

> HABEAS CORPUS 110.260 SÃO PAULO RELATOR : MIN. LUIZ FUX PACTE.(S) :LAERCIO BRAZ PEREIRA SALES IMPTE.(S) :DEFENSORIA PÚBLICA DA UNIÃO PROC.(A/S)(ES) :DEFENSOR PÚBLICO-GERAL FEDERAL COATOR(A/S)(ES) :SUPERIOR TRIBUNAL DE JUSTIÇA

In these passages, not only are all letters capitalized, but also there is no ordinary phrase structure of subject and predicate. Intuitively, it follows that the distribution of such documents differs from the existing corpora in a way that models trained on them will perform poorly when processing legal documents. Also, as they do not have specific tags for juridical entities, the models would fail to extract such legal knowledge.

The present paper proposes a Portuguese language dataset for named entity recognition composed entirely of manually annotated legal documents. Furthermore two new categories (LEGISLACAO, for named entities referring to laws; and JURISPRUDENCIA, for named entities referring to legal cases) are added to better extract legal knowledge.

Some efforts have been made on NER in legal texts. For instance, Dozier et al. [4] propose a NER system for Title, Document Type, Jurisdiction, Court and Judge tagging. Nevertheless, only the first entity is identified using a statistical approach, while the others are classified with contextual rules and

Table 1. Sentence, token and document count for each set.

Set	Documents	Sentences	Tokens
Training set	50	7,827	229,277
Development set	10	1,176	41,166
Test set	10	1,389	47,630

lookup tables. Cardellino et al. [2] used the Wikipedia to generate an automatically annotated corpus, tagging persons, organizations, documents, abstraction (rights, legal doctrine) and act (statutes) entities. As far as we are aware, the present paper is the first to propose a benchmark dataset and a baseline method for NER in legal texts in Portuguese.

The rest of this paper is organized as follows. In Sect. 2, we discuss the dataset creation process. We present the model used to evaluate our dataset in Sect. 3, along with the training of the model and our choice of hyper-parameters in Sect. 4. In Sect. 5 we present the results achieved regarding the test sets and Sect. 6 presents the final considerations.

2 The LeNER-Br Dataset

To compose the dataset, 66 legal documents from several Brazilian Courts were collected. Courts of superior and state levels were considered, such as Supremo Tribunal Federal, Superior Tribunal de Justiça, Tribunal de Justiça de Minas Gerais and Tribunal de Contas da União. In addition, four legislation documents were collected, such as Lei Maria da Penha, giving a total of 70 documents.

For each document, the NLTK [1] was used to split the text into a list of sentences and tokenize them. The final output for each document is a file with one word per line and an empty line delimiting the end of a sentence.

After preprocessing the documents, WebAnno [3] was employed to manually annotate each one of the documents with the following tags: “ORGANIZACAO” for organizations, “PESSOA” for persons, “TEMPO” for time entities, “LOCAL” for locations, “LEGISLACAO” for laws and “JURISPRUDENCIA” for decisions regarding legal cases. The last two refer to entities that correspond to “Act of Law” and “Decision” classes from the Legal Knowledge Interchange Format ontology [10] respectively.

The IOB tagging scheme [21] was used, where “B-” indicates that a tag is the beginning of a named entity, “I-” indicates that a tag is inside a named entity and “O-” indicates that a token does not pertain to any named entity. Named entities are assumed to be non-overlapping and not spanning more than one sentence.

To create the dataset, 50 documents were randomly sampled for the training set and 10 documents for each of the development and test sets. The total number of tokens in LeNER-Br is comparable to other named entity recognition corpora such as Paramopama and CONLL-2003 English [24] datasets (318,073, 310,000

Table 2. Named entity word count for each set.

Category	Training set	Development set	Test set
Person	4,612	894	735
Legal cases	3,967	743	660
Time	2,343	543	260
Location	1,417	244	132
Legislation	13,039	2,609	2,669
Organization	6,671	1,608	1,367

and 301,418 tokens respectively). Table 1 presents the number of tokens and sentences of each set and Table 2 displays the number of words in named entities of each set per class. Table 3 presents an excerpt from the training set.

3 The Baseline Model: LSTM-CRF

To establish a methodological baseline on our dataset, we chose the LSTM-CRF model, proposed in [13]. This model is proven to be capable of achieving state-of-the-art performance on the English CoNLL-2003 test set [24] (a F1-score of 90.94%). It also has readily available open source implementations [6], which was adapted for the needs of the present work.

The architecture of the model consists of a Bidirectional [7] Long Short-Term Memory Layer (LSTM) [9] followed by a CRF [12] layer. The input of the model is a sequence of vector representations of individual words constructed from the concatenation of both word embeddings and character level embeddings.

For the word lookup table we used 300 dimensional GloVe [20] word embeddings pretrained on a multi-genre corpus formed by both Brazilian and European Portuguese texts [8]. These word embeddings are fine tuned during training.

The character level embeddings are obtained from a character lookup table initialized at random values with embeddings for every character in the dataset. The embeddings are fed to a separate bidirectional LSTM layer. The output is then concatenated with the pretrained word embeddings, resulting in the final vector representation of the word. Figure 1 presents an overview of this process.

To reduce overfitting and improve the generalization capabilities of the model a dropout mask [23] is applied to the outputs of both bidirectional LSTM layers, i.e., the one following the character embeddings and the one after the final word representation. Figure 2 shows the main architecture of the model.

4 Experiments and Hyper-parameters Setting

This section presents the methods employed to train the model and displays the hyper-parameters that achieved the best performance.

Table 3. Two excerpts from the training set. Each line has a word, a space delimiter and the tag corresponding to the word. Sentences are separated by an empty line.

A	O	TJMG	B-ORGANIZACAO
falta	O	-	O
de	O	Apelação	B-JURISPRUDENCIA
intervenção	O	Cível	I-JURISPRUDENCIA
do	O	1.0549.15.003028-2/003	I-JURISPRUDENCIA
Ministério	B-ORGANIZACAO	,	O
Público	I-ORGANIZACAO	Relator	O
nas	O	(	O
ações	O	a	O
em	O	)	O
que	O	:	O
deva	O	Des	O
figurar	O	.	O
como	O	(	O
fiscal	O	a	O
da	O	)	O
lei	O	Otávio	B-PESSOA
e	O	Portes	I-PESSOA
da	O	,	O
Constituição	B-LEGISLACAO	16ª	B-ORGANIZACAO
(	O	CÂMARA	I-ORGANIZACAO
custus	O	CÍVEL	I-ORGANIZACAO
legis	O	,	O
et	O	julgamento	O
constituitionis	O	em	O
’	O	28/09/2017	B-TEMPO
)	O	,	O
enseja	O	publicação	O
de	O	da	O
forma	O	súmula	O
inexorável	O	em	O
a	O	06/10/2017	B-TEMPO
nulidade	O	)	O
do	O	Assim	O
processo	O	sendo	O
,	O	,	O
segundo	O	entendo	O
prescreve	O	que	O
o	O	deve	O
artigo	B-LEGISLACAO	ser	O
279	I-LEGISLACAO	acolhida	O
...	...	...	...

Both Adam [11] and Stochastic Gradient Descent (SGD) with momentum were evaluated as optimizers. Although SGD had slower convergence, it achieved better scores than Adam. Gradient clipping was employed to prevent the gradients from exploding.

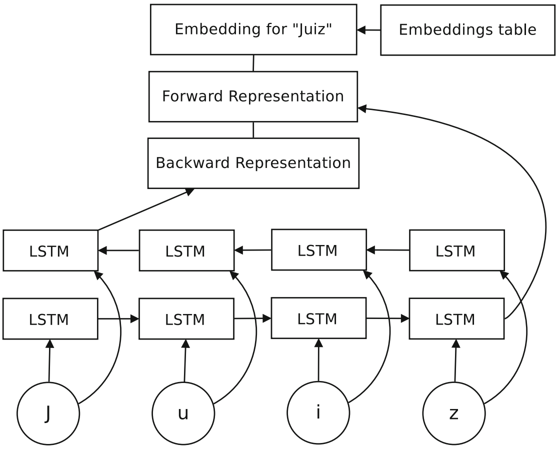


Fig. 1. Each word vector representation is a result of the concatenation of the outputs of a bidirectional LSTM and the word level representation from the word lookup table.

Table 4. Model hyper-parameter values.

Hyper-parameter	Value
Word embedding dimension	300
Character embedding dimension	50
Number of epochs	55
Dropout rate	0.5
Batch size	10
Optimizer	SGD
Learning rate	0.015
Learning rate decay	0.95
Gradient clipping threshold	5
First LSTM layer hidden units	25
Second LSTM layer hidden units	100

After experimenting with hyper-parameters, the best performance was achieved with the ones used in [13], presented in Table 4. It is worth noting that the number of LSTM units refers to one direction only. Since the LSTM are bidirectional, the final number of units doubles. Moreover, the learning rate decay is applied after every epoch. The net parameters were saved only when achieving better performance on the validation set than past epochs.

The model was first trained using the Paramopama Corpus [17] to evaluate if it could achieve state-of-the-art performance on a Portuguese dataset. This dataset contains four different named entities: persons, organizations, locations

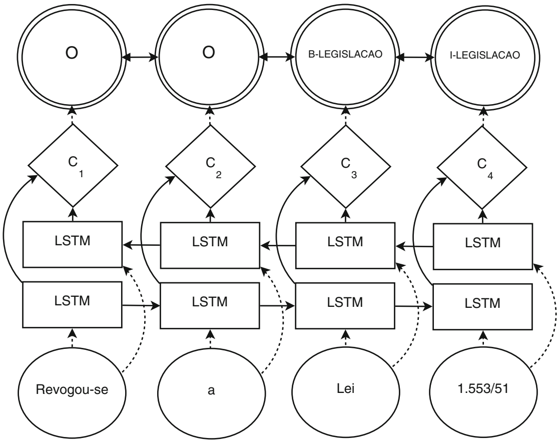


Fig. 2. The LSTM-CRF model. The word vector representations serve as input to a bidirectional LSTM layer. C_i represents the concatenation of left and right context of word i. Dotted lines represent connections after a dropout layer is applied.

and time entities. After confirming that the model performed better than the state-of-the-art model (ParamopamaWNN [16]), the LSTM-CRF network was trained with the proposed dataset.

The preprocessing steps applied were lowercasing the words and replacing every digit with a zero. Both steps are necessary to match the preprocessing of the pretrained word embeddings. Since the character-level representation preserves the capitalization, this information is not lost when the words are lowercased.

5 Results

The metric used to evaluate the performance of the model on both datasets was the F_1 Score. Tables 5 and 6 compare the performance of the LSTM-CRF [13] and ParamopamaWNN [16] models on different test sets. Test Set 1 and Test Set 2 are the last 10% of the WikiNER [19] and HAREM [22] corpora respectively. Table 7 shows the scores achieved by the LSTM-CRF model when training on the proposed dataset.

The obtained results show that the LSTM-CRF network outperforms the ParamopamaWNN on both test sets, achieving better precision, recall and F_1 scores in the majority of the entities. Furthermore, it improved the overall score by 2.48% and 4.58% on the first and second test sets respectively.

As far as we are aware, there is no published material about legal entities recognition in Portuguese, so it was not possible to establish a baseline for comparison on LeNER-Br. Despite that, the obtained results on LeNER-Br show that a model trained with it can achieve performance in legal cases and legislation recognition comparable to the ones seen in Paramopama entities, with F_1

Table 5. Results on paramopama test set 1 (10% of the WikiNER [19]).

Entity	ParamopamaWNN			LSTM-CRF		
	Precision	Recall	F_1	Precision	Recall	F_1
Person	83.76%	90.50%	87.00%	**91.80%**	**92.43%**	**92.11%**
Location	87.55%	**88.09%**	87.82%	**92.80%**	87.39%	**90.02%**
Organization	69.55%	82.35%	75.41%	**72.27%**	**83.94%**	**77.67%**
Time	86.96%	89.06%	88.00%	**92.54%**	**96.66%**	**94.56%**
Overall	86.45%	89.77%	88.08%	**90.01%**	**91.16%**	**90.50%**

Table 6. Results on paramopama test set 2 (HAREM [22]).

Entity	ParamopamaWNN			LSTM-CRF		
	Precision	Recall	F_1	Precision	Recall	F_1
Person	84.36%	88.67%	86.46%	**94.10%**	**95.78%**	**94.93%**
Location	84.08%	86.85	85.44%	**90.51%**	**92.26%**	**91.38%**
Organization	81.48%	54.15%	65.06%	**83.33%**	**78.46%**	**80.82%**
Time	**98.37%**	87.40%	92.56%	91.73%	**94.01%**	**92.86%**
Overall	83.83%	88.65%	86.17%	**90.44%**	**91.10%**	**90.75%**

Table 7. Results on LeNER-Br, our dataset for NER on Legal texts from Brazil.

Entity	Precision	Recall	F_1
Person	94.44%	92.52%	93.47%
Location	61.24%	59.85%	60.54%
Organization	91.27%	85.66%	88.38%
Time	91.15%	91.15%	91.15%
Legislation	97.08%	97.00%	97.04%
Legal cases	87.39%	90.30%	88.82%
Overall	93.21%	91.91%	92.53%

scores of 88.82% and 97.04% respectively. In addition, person, time entities and organization classification scores were compatible with the ones observed in the Paramopama scenarios, obtaining scores greater than 80%.

However, location entities have a noticeably lower score than the others on LeNER-Br. This drop could be due to many different reasons. The most important one is probably the fact that words belonging to location entities are rare in LeNER-Br, representing 0.61% and 0.28% of the words pertaining to entities in the train and test sets respectively. Furthermore, location entities are easily mislabeled, as there are words that, depending on the context, may refer to a person, a location or a organization. A good example is treating the name of an avenue as the name of a person. For instance, instead of identifying "avenida José Faria da Rocha" as a location, the model classified "José Faria da Rocha" as a person.

6 Conclusion

This paper presented LeNER-Br, a Portuguese language dataset for named entity recognition applied to legal documents. As far as we are aware, this is the first dataset of its kind. LeNER-Br consists entirely of manually annotated legislation and legal cases texts and contains tags for persons, locations, time entities, organizations, legislation and legal cases. A state-of-the-art machine learning model, the LSTM-CRF, trained on this dataset was able to achieve a good performance: average F_1 score of 92.53%. There is room for improvement, which means that this dataset will be relevant to benchmark methods that are sill to be proposed.

Future work would include the expansion of the dataset, adding legal documents from different courts and other kinds of legislation, e.g. Brazilian Constitution, State Constitutions, Civil and Criminal Codes, among others. In addition, the use of word embeddings trained on a large corpus of legislation and legal documents could potentially improve the performance of the model.

Acknowledgements. TEdC acknowledges the support of "Projeto de Pesquisa& Desenvolvimento de aprendizado de máquina (machine learning) sobre dados judiciais das repercussões gerais do Supremo Tribunal Federal - STF".

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media Inc., Newton (2009)
2. Cardellino, C., Teruel, M., Alonso Alemany, L., Villata, S.: A low-cost, high-coverage legal named entity recognizer, classifier and linker. In: Proceedings of the 16th International Conference on Artificial Intelligence and Law (ICAIL), London, United Kingdom, June 2017. https://hal.archives-ouvertes.fr/hal-01541446
3. de Castilho, R.E., et al.: A web-based tool for the integrated annotation of semantic and syntactic structures. In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pp. 76–84 (2016)
4. Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., Wudali, R.: Named entity recognition and resolution in legal text. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS (LNAI), vol. 6036, pp. 27–43. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12837-0_2
5. Freitas, C., Mota, C., Santos, D., Oliveira, H.G., Carvalho, P.: Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In: Language Resources and Evaluation Conference (LREC). European Language Resources Association (2010)
6. Genthial, G.: Sequence tagging - named entity recognition with Tensorflow (2017). GitHub repository https://github.com/guillaumegenthial/sequence_tagging/tree/0048d604f7a4e15037875593b331e1268ad6e887

7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
8. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: evaluating on word analogies and natural language tasks. In: Proceedings of Symposium in Information and Human Language Technology. Sociedade Brasileira de Computação, Uberlandia, MG, Brazil, 2–5 October 2017. https://arxiv.org/abs/1708.06025
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
10. Hoekstra, R., Breuker, J., Bello, M.D., Boer, A.: The LKIF core ontology of basic legal concepts. In: Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (2007)
11. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015). https://arxiv.org/abs/1412.6980
12. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning (ICML). ACM (2001). http://portal.acm.org/citation.cfm?id=655813
13. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of NAACL-HLT, pp. 260–270. Association for Computational Linguistics (ACL), San Diego, 12–17 June 2016. https://arxiv.org/abs/1603.01360
14. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1064–1074. ACL, Berlin, 7–12 August 2016. https://arxiv.org/abs/1603.01354
15. Mansouri, A., Affendey, L.S., Mamat, A.: Named entity recognition approaches. Int. J. Comput. Sci. Netw. Secur. **8**(2), 339–344 (2008)
16. Mendonça Jr., C.A.E., Barbosa, L.A., Macedo, H.T., São Cristóvão, S.: Uma arquitetura híbrida LSTM-CNN para reconhecimento de entidades nomeadas em textos naturais em língua portuguesa. In: XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). SBC (2016)
17. Mendonça Jr., C.A.E., Macedo, H., Bispo, T., Santos, F., Silva, N., Barbosa, L.: Paramopama: a Brazilian-Portuguese corpus for named entity recognition. In: XII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). SBC (2015)
18. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
19. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from Wikipedia. Artif. Intell. **194**, 151–175 (2013)
20. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
21. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., Yarowsky, D. (eds.) Natural Language Processing Using Very Large Corpora. TLTB, pp. 157–176. Springer, Dordrecht (1999). https://doi.org/10.1007/978-94-017-2390-9_10. Preprint available at: http://arxiv.org/abs/cmp-lg/9505040

22. Santos, D., Cardoso, N.: A golden resource for named entity recognition in Portuguese. In: Vieira, R., Quaresma, P., Nunes, M.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 69–79. Springer, Heidelberg (2006). https://doi.org/10.1007/11751984_8
23. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
24. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, vol. 4, pp. 142–147. Association for Computational Linguistics (2003)

Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results

Rafael A. Monteiro[1], Roney L. S. Santos[1], Thiago A. S. Pardo[1](✉), Tiago A. de Almeida[2], Evandro E. S. Ruiz[3], and Oto A. Vale[4]

[1] Interinstitutional Center for Computational Linguistics (NILC), University of São Paulo, São Carlos, Brazil
{rafael.augusto.monteiro,roneysantos}@usp.br, taspardo@icmc.usp.br

[2] Federal University of São Carlos, Sorocaba, Brazil
talmeida@ufscar.br

[3] University of São Paulo, Ribeirão Preto, Brazil
evandro@usp.br

[4] Interinstitutional Center for Computational Linguistics (NILC), Federal University of São Carlos, São Carlos, Brazil
otovale@ufscar.br

Abstract. Fake news are a problem of our time. They may influence a large number of people on a wide range of subjects, from politics to health. Although they have always existed, the volume of fake news has recently increased due to the soaring number of users of social networks and instant messengers. These news may cause direct losses to people and corporations, as fake news may include defamation of people, products and companies. Moreover, the scarcity of labeled datasets, mainly in Portuguese, prevents training classifiers to automatically filter such documents. In this paper, we investigate the issue for the Portuguese language. Inspired by previous initiatives for other languages, we introduce the first reference corpus in this area for Portuguese, composed of aligned true and fake news, which we analyze to uncover some of their linguistic characteristics. Then, using machine learning techniques, we run some automatic detection methods in this corpus, showing that good results may be achieved.

Keywords: Fake news · Reference corpus · Linguistic features
Machine learning

1 Introduction

Since the earliest times, long before the advent of computers and the web, fake news (also known as deceptive news) were transmitted through the oral tradition, in the form of rumors (face to face) or in the yellow/sensational press, either to "innocently" talk about other people lives, or to intentionally harm the

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 324–334, 2018.
https://doi.org/10.1007/978-3-319-99722-3_33

reputation of other people or rival companies. Nowadays, social networks and instant messenger apps have allowed such news to reach an audience that was never imagined before the web era. Due to their appealing nature, they spread rapidly [20], influencing people behavior on several subjects, from healthy issues (e.g., by revealing miraculous medicines) to politics and economy (as in the recent Cambridge Analytica/Facebook scandal[1] and in the Brexit situation[2]).

As the spread of fake news has reached a critical point, initiatives to fight back fake news have emerged. On the one hand, journalistic agencies have supported fact checking sites (e.g., Agência Lupa[3] and Boatos.org[4]) and big digital companies (as Facebook[5]) have attempted to block fake news and to educate users. On the other hand, academic efforts have been made by studying how such news spread, the behavior of the users that produce and read them, and language usage characteristics of fake news, in order to identify such news. This last research line - on language characteristics - has been mainly explored in the Natural Language Processing (NLP) area.

In NLP, the attempts to deal with fake news are relatively recent, both on the theoretical (e.g., [7,12,24]) and practical points of view [1,11,16,18]. Some previous work has showed that humans perform poorly on separating true from fake news [3,10] and that the domain may affect this [14], but others have produced promising automatic results. Despite the advances already made, the lack of available corpora may compromise the evaluation of different approaches.

To fill this important gap, in this paper we investigate the issue of fake news detection for the Portuguese language. Inspired by previous initiatives for other languages, to the best of our knowledge, we introduce the first reference corpus in this area for Portuguese. This corpus is composed of aligned true and fake news, which we analyze to uncover some of their linguistic characteristics. Then, using traditional machine learning techniques and following some of the ideas of [16,22], we perform tests on the automatic detection of fake news, achieving good results. One of our main goals is that our corpus and methods may support future researches in the area.

The remainder of this paper is organized as follows. In Sect. 2, we briefly review the essential related work. Section 3 offers details about the newly-created corpus. In Sect. 4, we report our machine learning approaches for fake news detection. Finally, Sect. 5 concludes this paper.

2 Related Work

According to [17], there are three main types of deception in texts: (i) the ones with humor, clearly for fun, using sarcasm to produce satires and parodies; (ii)

[1] http://fortune.com/2018/04/10/facebook-cambridge-analytica-what-happened/.

[2] https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy/.

[3] http://piaui.folha.uol.com.br/lupa/.

[4] http://www.boatos.org/.

[5] https://newsroom.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/.

fake content, which intends to deceive people and to cause confusion; and (iii) rumors, which are non-confirmed and usually publicly accepted information. Fake content, in particular, may appear in different contexts. Fake news are a type of it, as well as fake reviews, for instance, that are tailored to harm or to promote something.

Although the recent interest growing in the area, there are several available corpora of different types of deception. In [15], the authors present three datasets related to social topics, such as opinions on abortion, death penalty, and feelings about a best friend, containing 100 deceptive and 100 truthful sentences. In [18], the authors build two datasets containing satirical and true news in four different domains (civics, science, business, and "soft" news), totalizing 240 samples. In [14], two datasets are collected on the celebrity news domain. The first one consists in emulating journalistic writing style, using Amazon Mechanical Turk, resulting in 240 fake news. The second one is collected from the web, following similar guidelines to the previous dataset (aiming to identify fake content that naturally occurs on the web), resulting in 100 fake and 100 legitimate news. Other corpora are available in English, such as the Emergent [8] and LIAR [21] corpora. For Portuguese, it is possible to find some websites that compile true and fake news for fact checking (as the ones cited in the previous section), but they often present comments about the news (and not the original texts themselves) and are not ready-to-use corpora for NLP purposes.

Some methods for detecting deceptive content have been investigated, using varied textual features, as commented by [5] and systematized in [24]. [1] studies false declarations in social networks, looking for clues of falsification (lies, contradictions and distortions), exaggeration (modifiers and superlatives), omission (lack of information, half truths) and deception (subject change, irrelevant information and misconception). [22] proposes to look at the amount of verbs and modifiers (adjectives and adverbs), complexity, pausality, uncertainty, non-immediacy, expressivity, diversity and informality features. In [14–16], the authors compare the performance of classifiers using n-grams/bag of words, part of speech tags and syntactic information, readability metrics and word semantic classes.

Despite the efforts already made, as far as we know, there is no public and labeled dataset of fake news written in Portuguese. The absence of representative data may seriously impact the processes of development, evaluation and comparison of automatic detection methods. In what follows, we report our efforts to build the first reliable corpus in this area for Portuguese.

3 The Fake.Br Corpus

In order to create a reliable corpus, we have collected and labeled real samples written in Portuguese. The corpus – simply called "Fake.Br Corpus" – is composed of true and fake news that were manually aligned, focusing only on Brazilian Portuguese. To the best of our knowledge, there is no other similar available corpus for this language.

Collecting texts to the corpus was not a simple task. It took some months to manually find and check available fake news in the web and, then, to semi-automatically look for corresponding true news for each fake one. The manual step was necessary to check the details of the fake news and if they were in fact fake, as we wanted to guarantee the quality and reliability of the corpus.

The alignment of true and fake news is relevant for both linguistic studies and machine learning purposes, as positive and negative instances are important for validating linguistic patterns and automatic learning, depending on the adopted approach. Besides this, the alignment is a desired characteristic of the corpus, as pointed by [17], which also suggests the following for assembling the corpus: news should be in plain text format, as this is usually more appropriate for NLP; the news must have similar sizes (usually in number of words) in order to avoid bias in learning, but, if this is not the case, size normalization (e.g., text truncation) may be carried out when necessary; specification of a time period for collecting the texts, as writing style may change in time and this may harm the corpus purposes; maintenance of pragmatic factors, e.g., the original link to the news, as such information may be useful in the future for fact checking tasks [13].

Overall, we collected 7,200 news, with exact 3,600 true and 3,600 fake news. All of them are in plain text format, with each one in a different file. We kept size homogeneity as much as we could, but some true news are longer than the fake ones. We established a 2 years time interval for the news, from January of 2016 to January of 2018, but there were cases of fake news in this time period that referred to true news of a time before this. We did not consider this as a problem and kept these news in the corpus. Finally, we saved all the links and other metadata information (such as the author, date of publication, and quantity of comments and visualizations, when possible) that was available.

We manually analyzed and collected all the available fake news (including their titles) in the corresponding time period from 4 websites: *Diário do Brasil* (3,338 news[6]), *A Folha do Brasil* (190 news), *The Jornal Brasil* (65 news) e *Top Five TV* (7 news). Finally, we filtered out those news that presented half truths[7], keeping only the ones that were entirely fake.

The true news in the corpus were collected in a semiautomatic way. In a first step, using a web crawler, we collected news from major news agencies in Brazil, namely, *G1*, *Folha de São Paulo* and *Estadão*. The crawler searched in the corresponding webpages of these agencies for keywords of the fake news, which were nouns and verbs that occurred in the fake news titles and the most frequent words in the texts (ignoring stopwords). About 40,000 true news were collected this way. Then, for each fake news, we applied a lexical similarity measure (the cosine measure presented in [19]), choosing the most similar ones to the fake news, and performed a final manual verification to guarantee that the fake and true news were in fact subject-related. It is interesting to add that there were cases in that the true news explicitly denied the corresponding fake one (see,

[6] We could realize that most of the checked sites shared many fake news.

[7] Half truth may be defined as the case in which some actual facts are told in order to give support to false facts [4].

e.g., the first example in Table 1), but others were merely on the same topic (second example in Table 1).

Table 1. Examples of aligned true and fake news.

Fake	True
Michel Temer propõe fim do carnaval por 20 anos, "PEC dos gastos". Michel Temer afirmou que não deve haver gastos com aparatos supérfluos sem pensar primeiramente na educação do Brasil. A medida pretende calcelar o carnaval de 2018	*Michel Temer não quer o fim do Carnaval por 20 anos. Notícias falsas misturam proximidade dos festejos, crise econômica e medidas impopulares do governo do peemedebista*
Acabou a mordomia ! Ingresso mais barato pra mulher é ilegal. Baladas que davam meia entrada para mulher, ou até mesmo gratuidade, esto na ilegalidade agora. Acabou o preconceito com os homens nas casas de show de todo o Brasil	*Ingresso feminino barato como marketing 'não inferioriza mulher', diz juíza do DF. Afirmação consta em decisão sobre preços diferentes para homens e mulheres em festa no Lago Paranoá. 'Prática permite que mulher possa optar por participar de tais eventos sociais', diz texto*

Overall, the collected news may be divided into 6 big categories regarding their main subjects: politics, TV & celebrities, society & daily news, science & technology, economy, and religion. In order to guarantee consistency and annotation quality, the texts were manually labeled with the categories. Table 2 shows the distribution of texts by category. As expected, politics is the most frequent one.

Table 2. Amount of documents per category in the Fake.Br corpus.

Category	Number of samples	%
Politics	4,180	58.0
TV & celebrities	1,544	21.4
Society & daily news	1,276	17.7
Science & technology	112	1.5
Economy	44	0.7
Religion	44	0.7

Table 3 shows a overall comparison of the news, including the average number of tokens and sentences, as well as several other features. It is interesting to notice some differences, e.g., spelling errors were more frequent in the fake news.

Table 3. Basic statistics about the Fake.Br corpus.

Features	Fake news	True news
Avg number of tokens	216.1	1,268.5
Avg number of types (without punctuation and numbers)	119.2	494.1
Avg size of words (in characters)	4.8	4.8
Type-token ratio	0.68	0.47
Avg number of sentences	12.7	54.8
Avg size of sentences (in words)	15.3	21.1
Avg number of verbs (norm. by the avg number of tokens)	14.3	13.4
Avg number of nouns (norm. by the avg number of tokens)	24.5	24.6
Avg number of adjectives (norm. by the avg number of tokens)	4.1	4.4
Avg number of adverbs (norm. by the avg number of tokens)	3.7	4.0
Avg number of pronouns (norm. by the avg number of tokens)	5.0	5.2
Avg number of stopwords (norm. by the avg number of tokens)	31.0	32.8
Percentage of news with spelling errors	36.0	3.0

Finally, we adopted the proposal of [22] to compute other linguistic features that may serve as indications of fake content, to know: (i) pausality, which checks the frequency of pauses in a text, computed as the number of punctuation signals over the number of sentences; (ii) emotiveness, which is an indication of language expressiveness in a message [23], computed as the sum of the number of adjectives and adverbs over the sum of nouns and verbs; (iii) uncertainty, measured by the number of modal verbs and occurrences of passive voices; and (iv) non-immediacy, measured by the number of 1st and 2nd pronouns. Table 4 shows the values of these features in the corpus. The higher differences in uncertainty and

Table 4. Linguistic features of [22] in the Fake.Br corpus.

Features	Fake news	True news
Avg pausality per text	2.46	3.04
Avg emotiveness per text	0.20	0.21
Avg uncertainty per text	4.48	23.24
Avg non-immediacy per text	0.62	4.05

non-immediacy values are due to the size difference of the texts, as these two metrics are not normalized.

In what follows, we present our experiments on fake news detection using the above corpus.

4 Experiments and Results

Motivated to create an automatic classifier of fake news, we run some tests using machine learning over the Fake.Br corpus. To guarantee a fair classification, we have normalized the size of the texts (in number of words) by truncating the longer texts to the size of their aligned counterparts.

Following [16], we run the widely used SVM technique [6] (the LinearSVC implementation in Scikit-learn, with default parameters). We tried different features of [16, 22]:

- bag of words/unigrams (simply indicating whether each word occurred or not in the text, using boolean values), after case folding, stopword[8] and punctuation removal, and stemming;
- the (normalized) number of occurrences of each part of speech tag, as indicated by the NLPNet tagger [9];
- the (normalized) number of occurrences of semantic classes, as indicated by LIWC for Brazilian Portuguese [2], which is an enriched lexicon that associates to each word one or more possible semantic classes (from a set of 64 available classes);
- and the pausality, emotiveness, (normalized) uncertainty and (normalized) non-immediacy features.

Still following the work of [16], we used an evaluation strategy of 5-fold cross-validation. We computed the traditional precision, recall and F-measure metrics for each class, as well as general accuracy. Table 5 shows the average results that we achieved for different feature sets. The first three rows refer to features of [16], while the fourth is a combination of them; the next four rows are the features of [22], also followed by their combination; we then combine the best features of both initiatives; and, finally, we combine all the features (in the last row).

Bag of words alone could (surprisingly) achieve good results (88% of F-measure, for both true and fake news), and other features (including the ones of [22]) did not help to significantly improve this. It is interesting that most of the methods performed similarly for the two classes.

We show in Table 6 the confusion matrix for the bag of words classification. One may see that there is still room for improvements. In our opinion, misclassifying (and, consequently, filtering out) true news is more harmful than not detecting some fake news (the same logic of spam detection), and this must have more attention in the future.

[8] We also remove numeric values in order to help avoiding sparsity.

Table 5. Classification results.

Features	Precision		Recall		F-Measure		Accuracy
	Fake	True	Fake	True	Fake	True	
Part of speech (POS) tags	0.76	0.74	0.73	0.77	0.74	0.76	0.75
Semantic classes	0.73	0.73	0.73	0.73	0.73	0.73	0.73
Bag of words	0.88	0.89	0.89	0.88	0.88	0.88	0.88
POS tags + semantic classes + bag of words	0.88	0.89	0.89	0.88	0.89	0.89	0.88
Pausality	0.52	0.52	0.58	0.46	0.55	0.49	0.52
Emotiveness	0.57	0.56	0.53	0.61	0.55	0.58	0.56
Uncertainty	0.51	0.51	0.46	0.57	0.48	0.54	0.51
Non-immediacy	0.53	0.51	0.16	0.86	0.24	0.64	0.51
Pausality + emotiveness + uncertainty + non-immediacy	0.57	0.56	0.53	0.60	0.55	0.58	0.57
Bag of words + emotiveness	0.88	0.89	0.89	0.88	0.89	0.89	0.89
All the features	0.89	0.89	0.89	0.89	0.89	0.89	0.89

Table 6. Confusion matrix for bag of words classification.

Classified as	Actual classes	
	True	Fake
True	3,192	432
Fake	408	3,168

We checked that the classification errors are correlated with the news categories in the following way: 11.6% of the political texts were misclassified; for TV & celebrities, 10.4%; for society & daily news, 12.3%; for science & technology, 16.1%; for economy, 18.1%; and, for religion, 20.4%. Economy and religion categories appear to be the most difficult ones, but this may have happened due to fewer learning instances that we have for such categories.

We have also run some other machine learning techniques, from different paradigms, as Naïve-Bayes, Random Forest, and Multilayer Perceptron (with the default parameters of Scikit-learn). Additionally, we tried bag of words with different minimum numbers of occurrence in the corpus, as well as other values for the occurrence of words, as their (normalized) frequency (instead of boolean 0 or 1 values). Multilayer Perceptron could achieve 90% of accuracy. Considering words with at least 3 occurrences produced the same results; from 5 to more occurrences, the results start to slightly fall. Using word frequency (instead of boolean values) did not improve the results.

One final test was to run the experiments without truncating the size of the texts. The use of full texts achieved impressive 96% of accuracy with bag of words, but this classification is probably biased, as true texts are significantly longer than the fake ones.

It is interesting that, in our case, differently from [16], part of speech tags did not produce the best results. Such difference is probably explained by the dataset. While our dataset is "spontaneous" (to the extent that such nomenclature makes sense for fake news), collected from the web, [16] used a dataset of a different nature (in fact, the authors used sentences), produced by hired people to the task.

Overall, the achieved results were above our expectations. One factor that may help explaining such good results is that we have filtered out news with half truth, which might make things more complex (and equally interesting). This remains for future work, as we comment below.

5 Conclusions

To the best of our knowledge, we have presented the first reference corpus for fake news detection for Portuguese - the Fake.Br corpus. More than this, we have run some experiments, following some well known attempts in the area, and produced good results, considering the apparent difficulty of the task. We hope that our corpus may foster research in the area and that the methods we tested instigate new ones in the future.

For future work, we hope to identify other features that may help distinguishing the remaining misclassified examples, as well as to test other classification techniques, using, e.g., distributional representations and methods. We also aim at dealing with other deception types, such as satiric texts and fake opinion reviews, and with more complex cases, as the news including half truth.

More information about this work and the related tools and resources may be found at the OPINANDO project website[9].

Acknowledgments. The authors are grateful to FAPESP, CAPES and CNPq for supporting this work.

References

1. Appling, D.S., Briscoe, E.J., Hutto, C.J.: Discriminative models for predicting deception strategies. In: Proceedings of the 24th International Conference on World Wide Web, pp. 947–952 (2015)
2. Balage Filho, P.P., Pardo, T.A., Aluisio, S.M.: An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pp. 215–219 (2013)
3. Bond Jr., C.F., DePaulo, B.M.: Accuracy of deception judgments. Personal. Soc. Psychol. Rev. **10**(3), 214–234 (2006)

[9] https://sites.google.com/icmc.usp.br/opinando/.

4. Clem, S.: Post-truth and vices opposed to truth. J. Soc. Christ. Eth. **37**(2), 97–116 (2017)
5. Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: methods for finding fake news. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, pp. 82:1–82:4 (2015)
6. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
7. Duran, N.D., Hall, C., McCarthy, P.M., McNamara, D.S.: The linguistic correlates of conversational deceprion: comparing natural language processing technologies. Appl. Psycholinguist. **31**(3), 439–462 (2010)
8. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1163–1168. Association for Computational Linguistics (2016)
9. Fonseca, E.R., Aluísio, S.M.: A deep architecture for non-projective dependency parsing. In: Proceedings of the NAACL-HLT Workshop on Vector Space Modeling for NLP (2015)
10. George, J.F., Keane, B.T.: Deception detection by third party observers. In: Paper Presented at the Deception Detection Symposium, 39th Annual Hawaii International Conference on System Sciences (2006)
11. Gimenes, G., Cordeiro, R.L., Rodrigues-Jr, J.F.: Orfel: efficient detection of defamation or illegitimate promotion in online recommendation. Inf. Sci. **379**, 274–287 (2017)
12. Hauch, V., Blandón-Gitlin, I., Masip, J., Sporer, S.L.: Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. Personal. Soc. Psychol. Rev. **19**(4), 307–342 (2015)
13. Musskopf, I.: A ciência da detecção de fake news, September 2017. https://medium.com/data-science-brigade/a-ci%C3%AAncia-da-detec%C3%A7%C3%A3o-de-fake-news-d4faef2281aa
14. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. CoRR abs/1708.07104 (2017)
15. Pérez-Rosas, V., Mihalcea, R.: Cross-cultural deception detection. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 440–445 (2014)
16. Pérez-Rosas, V., Mihalcea, R.: Experiments in open domain deception detection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1120–1125 (2015)
17. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: three types of fakes. Proc. Assoc. Inf. Sci. Technol. **52**(1), 1–4 (2015)
18. Rubin, V.L., Conroy, N.J., Chen, Y., Cornwell, S.: Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 7–17 (2016)
19. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill Inc., New York (1986)
20. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
21. Wang, W.Y.: "Liar, Liar Pants on Fire": a new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada (2017)

22. Zhou, L., Burgoon, J., Twitchell, D., Qin, T., Nunamaker Jr., J.: A comparison of classification methods for predicting deception in computer-mediated communication. J. Manag. Inf. Syst. **20**(4), 139–165 (2004)
23. Zhou, L., Twitchell, D.P., Qin, T., Burgoon, J.K., Nunamaker, J.F.: An exploratory study into deception detection in text-based computer-mediated communication. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences (2003)
24. Zhou, L., Zhang, D.: Following linguistic footprints: automatic deception detection in online communication. Commun. ACM - Enterp. Inf. Integr.: Tools Merging Data **51**(9), 119–122 (2008)

Creating a Portuguese Context Sensitive Lexicon for Sentiment Analysis

Mateus Tarcinalli Machado[1](✉), Thiago A. S. Pardo[2], and Evandro Eduardo Seron Ruiz[1]

[1] Department of Computing and Mathematics – FFCLRP, University of São Paulo, Ribeirão Preto, Brazil
{mateusmachado,evandro}@usp.br

[2] Interinstitutional Center for Computational Linguistics (NILC), University of São Paulo, São Carlos, Brazil
taspardo@icmc.usp.br

Abstract. This work presents and evaluates a new Portuguese context sensitive lexicon for sentiment analysis. Distinctive composition approaches to produce lexicons from established ones were also tested. The experiments were carried out with the corpus ReLi, composed of opinative texts about books, and with the following sentiment lexicons: Brazilian Portuguese LIWC, Opinion Lexicon, and SentiLex.

Keywords: Polarity detection · Lexicon creation

1 Introduction

Sentiment Analysis uses Natural Language Processing techniques to extract and classify opinions, emotions, evaluations, and attitudes related to products, services, organizations, people, events and subjects expressed in free text. A Sentiment Analysis (SA) first application area was product evaluation, an area that had a big boost with the advent of Web 2.0, following the growth of electronic commerce and the more active participation of consumers and users on the web. Since the seminal paper by Pang and Lee [8], literature has presented SA as a rather complex task. The major obstacles may be divided into: (a) analyzing the meaning of sentiments and (b) detecting the suitable sentiment polarity. Some of these tasks are further discussed by Hussein [6].

SA can become more demanding if, for instance, the text mentions various characteristics about one assessed entity. In this case, the evaluator may qualify this single item with different feelings. This way, it may also occur that certain ratings can be positive or negative, depending on the analyzed *aspect* of the entity. Aspect-Based Sentiment Analysis (ABSA) is a fine grained form of SA aiming to identify the aspects of given entities and its related sentiments [15]. Pavlopoulos and Androutsopoulos [9] suggest that most ABSA systems subdivide this deeper SA processing into three subtasks, which are: (i) aspect term

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 335–344, 2018.
https://doi.org/10.1007/978-3-319-99722-3_34

extraction (detection of product's characteristics described in textual format), (ii) aspect term sentiment estimation (identification of the sentiment polarity – usually positive, negative or neutral – related to each aspect), and (iii) aspect aggregation (not always present, performs the grouping of identified aspects that are synonymous or near-synonyms).

Specifically for the aspect term sentiment estimation subtask (ii), there are two main approaches: (a) the lexicon-based approaches, and; (b) machine learning ones [14]. In the first, the polarity of a text is identified by analyzing the semantic orientation of the words and phrases composing the text. This orientation is obtained through dictionaries containing word lists and their polarities. For machine learning approaches it is necessary to build a classifier. This is generally accomplished by using examples of manually classified texts, which labels these methods as supervised classification task.

Although one may find plenty of research related to lexicon-based sentiment analysis for the English language, this paper focuses in the creation of a Portuguese context sensitive lexicon. The suggested approach is tested on opinative review texts about books, found in the only known Portuguese annotated corpus for aspects, the ReLi corpus [4]. The results of this approach are also compared to other methods of lexicon creation based on existing sentiment lexicons.

The rest of this paper is organized as follows: Sect. 2 discusses lexicon-based SA and some Portuguese lexicon creation processes. Section 3 describes the method used to create the proposed context sensitive lexicon and the dataset used to test it. Section 4 compares different SA lexicons performance under lexicon-based approaches. Finally, Sect. 5 draws some conclusions from previous results and outlines directions for future work.

2 Related Work

The Liu book [7] encapsulates a great part of the research on concepts related to the Sentiment Analysis area on its origins. In his book, he presents examples of applications, related problems, methods for sentiment analysis, aspect extraction, generation of lexicons, and opinion summarization, among others.

Souza *et al.* [13] created a new lexicon (OpLexicon) combining lexicons using three different techniques. A first lexicon was created using the analysis of an annotated corpus. Another one was created by searching for synonyms and antonyms in a thesaurus. A final lexicon was obtained through the automatic translation of *Liu's English Opinion Lexicon* [5].

Another important work was the one that created the lexicon Sentilex [11]. The development of the lexicon occurred in two stages: firstly, a dictionary of adjectives was created with their respective polarities. From these adjectives, a set of lexical-syntactic rules was manually created. These rules were applied to a large collection of n-grams. The frequencies of the adjectives and the rules found were used as inputs for a statistical classifier. In a second step, polarities were assigned to the new found adjectives and this list was expanded through the exploration of synonym graphs.

Lexicons play a major role in SA. An important related work is the paper by Taboada *et al.* [14], which presents the Semantic Orientation CALculator (SO-CAL) technique. SO-CAL calculates the polarity of a text using word lexicons marked with their semantic orientation (polarity and force). This system also identifies words that intensify or reverse (negation words) these polarities. Balage *et al.* [1] used SO-CAL to evaluate the Brazilian Portuguese LIWC dictionary for SA, comparing this lexicon with SentiLex and OpLexicon.

The International Workshop on Semantic Evaluation (SemEval) [10] is an important venue for researchers in SA. This workshop always presents new datasets for analysis, providing the basis for comparing results of different methods.

3 Materials and Methods

3.1 Methods

This work is focused in sentiment polarity identification using sentiment lexicons. In the next subsection, **LexReLi**, we introduce our methodology, constructing the proposed new context sensitive lexicon. Then, at **Lexicon combination** subsection, we investigate lexicons produced by different combination techniques that involve three well-known sentiment lexicons: Brazilian Portuguese LIWC, Opinion Lexicon and SentiLex.

LexReLi. Here we describe the construction of the Lexicon ReLi (LexReLi), specialized in identifying the polarity of aspects in opinion texts about books. For this task, we extended the dataset collecting, from the skoob website, more reviews about the books mentioned in ReLi (see more about ReLi in Sect. 3.2). This corpus is composed of 6,698 reviews, 51,148 phrases, and 980,640 words. For the construction of this lexicon, we combined some strategies of aspect identification with polarity detection and applied it to this corpus. The objective was to create a lexicon composed only by adjectives, where their polarities are identified through the context they belong to, *i.e.*, the sentence polarity.

Both corpora, the original ReLi and its extended version, have been submitted to a preprocessing phase. Two tokenizers [2] were applied to the extended ReLi, one to split the reviews into sentences and another to divide the sentences into tokens. After that, both corpora were submitted to a part-of-speech tagger [3]. Here are the tasks pursued to construct the LexReLi lexicon:

Constructing the LexReLi

1. **Aspect identification.** The first step was the identification of sentences that have nouns close to adjectives, as this combination tends to indicate an aspect (noun) close to its characteristic (adjective);
2. **Polarity detection.** The selected sentences went through a process of polarity detection. We then applied the Adjectives Preference method (as explained

in Sect. 3.3) with a lexicon obtained from the combination of LIWC, SentiLex, and OpLexicon (see Sect. 4 for results that justify this preference);
3. The frequency of adjective occurrence in positive and negative sentences was computed;
4. If one adjective appears more often in positive phrases, this polarity was assigned to it; otherwise, a negative polarity was assigned. If the difference between the number of times it appears in positive and in negative sentences is less than two, the adjective was not included in our lexicon.

Now, in order to verify if the combination of well-known lexicons can enhance the performance of SA, we investigated the construction of a new lexicon from a combination of three representative ones.

Lexicon Combination. Three frequently used lexicons in similar researches have been combined to form a new SA lexicon. They are: Opinion Lexicon [12, 13], SentiLex [11] and the Brazilian Portuguese Linguistic Inquiry and Word Count 2007 (LIWC) [1]. We have created combined lexicons from the three aforementioned dictionaries, using two combination methods.

Combined Lexicons. In a first approach, we only combine the dictionaries, one after the other, ignoring possible disagreements between them, prevailing, in these cases, the classification adopted by the first dictionary added to the set. This way, word order brought by the dictionary combination will interfere in the final result. This fact led us to create six possible combinations, altering the order of inclusion of the three dictionaries. In total, 6 lexicons were generated with equal amounts of words, but slight differences in polarities.

Conciliated Lexicon. In a second approach, we constructed a new lexicon based on the previous three aforementioned. It was established that a word is added to the new dictionary if it appears in only one of them, or if the word has the same polarity in at least two dictionaries. This way, eventual polarity disagreements among dictionaries are solved. Table 1 presents the final composition of the sentiment lexicons.

3.2 The Dataset

For our work on aspect-based polarity detection experiments, we worked with the ReLi corpus [4]. The ReLi is a Portuguese book review corpus composed by 1,600 reviews of 14 books collected from the ‘skoob’ website[1]. Skoob is a collaborative social network for book readers. The reviews on skoob were manually annotated for opinion presence, identified aspects, and their respective polarities.

From the ReLi corpus, we selected only phrases that contained at least one aspect and their respective polarity. That way, we worked with a set of 2,675 aspects and respective polarities (2,089 positives and 586 negatives), showing an unbalanced sentiment polarity distribution.

[1] https://www.skoob.com.br/.

Table 1. Polarity distribution found on lexicons.

Lexicon	Positive	Neutral	Negative	Total
OpLexicon	8, 595	8, 974	14, 550	32, 119
SentiLex	20, 478	7, 600	51, 112	79, 190
LIWC	12, 376	0	14, 612	26, 988
LexReLi	1, 091	0	452	1, 543
Conciliated	32, 543	11, 155	62, 676	106, 374
Combined*	34, 433	12, 636	64, 638	111, 707
LexReLi + Combined*	34, 974	12, 244	64, 685	111, 903

*Average of the six obtained lexicons.

3.3 Polarity Detection

We implemented four approaches for polarity detection in a two-step method. We combined the Aspect Based method with three phrasal level methods: Words Polarities, Adjectives Polarities and Adjectives Preference. As a first step we apply the Aspect Based method; if it can not identify the polarity of the sentence the algorithm follows a second step and uses one of the three phrasal level methods, which are: words polarities, adjectives polarities and adjectives preference. We applied our approach using the mentioned sentiment lexicons in Sect. 3.1.

Aspect Based. Demonstrated in Algorithm 1, this method tries to find the polarity related to every aspect in the sentences. The algorithm locates, in each sentence, every aspect marked in the ReLi corpus (line 5). Given the aspect, the algorithm searches for the nearest adjective and verifies its polarity in the lexicon (lines: 13, 14). To deal with negation, a list of negation terms was obtained from the Brazilian Portuguese 2007 LIWC lexicon (line 3). If there is a negation word between the aspect and the adjective, the polarity of the adjective is reversed (lines: 15, 16). The polarities of adjectives found close to aspects are then summed up (line 18). The final result of this sum indicates the polarity of the sentence (line 21).

Words Polarities. This is a simple method of polarity identification using a lexicon. Each word in a sentence is sought in a sentiment lexicon. The polarity value found for each word in a sentence is then summed. If the value obtained from this sum is greater than zero, it indicates a positive polarity; if it is equal to zero, we have a neutral one; and, if less than zero, we have a negative polarity. As in Taboada *et al.* [14] SO-CAL method, negation words modify polarity of nearest words. When the algorithm finds a word from LIWC negation word list, the polarity of the next found word is reversed.

Adjectives Polarities. Similar to the previous method but, in this case, only the adjective's polarities found in a sentence were added. Negation was also taken into account, reversing the polarity of the adjective closest to the negative term.

Algorithm 1. Aspect Based Polarity Detection algorithm

```
function aspectBased(sentence)
    $adjectives \leftarrow POSTagger(sentence).getAdjectives()$
    $negationWords \leftarrow LIWC.getNegationWords()$
    $polaritySum \leftarrow 0$
    for aspect in locateAspects(sentence) do
        $wordPosition \leftarrow 0$
        $aspectLastWordPosition \leftarrow getAspectLastWordPosition(aspect)$
        for word in sentence do
            if $wordPosition <= aspectLastWordPosition$ then
                continue
            if word in negationWords then
                $negationFlag \leftarrow True$
            if word in adjectives and word in lexicon then
                $adjectivePolarity \leftarrow lexicon.getWordPolarity(word)$
                if negationFlag then
                    $adjectivePolarity \leftarrow adjectivePolarity * -1$
                    $negationFlag \leftarrow False$
                $polaritySum \leftarrow polaritySum + adjectivePolarity$
                $break$ // inside for command
            $wordPosition \leftarrow wordPosition + 1$
    return $polaritySum$
```

Adjectives Preference. This method is based on both previous methods. Initially, similar to the second method, only the polarities of the adjectives were taken into account. The difference in this method is that, when no adjective is found in a lexicon, the algorithm analyzes the polarities of every word, just as in the first method. Negation was treated in the same way as in the Adjective Polarities method.

4 Results

In our experiments, we combine the method of detecting aspect polarity with the three above methods for detecting polarity. We apply these methods initially using the lexicons individually and then using the combined ones. Our intention is to analyze both, the performance of the algorithms and, the approaches used to create a lexicon either as in LexReLi or as a combination of other lexicons.

To evaluate the results of the experiments, we compared the polarities found by the implemented methods with those indicated in the ReLi corpus. This was accomplished using an evaluation methodology similar to the one used in the SemEval workshops [10]. We calculated the accuracy for each experiment, defined as the number of correctly predicted polarities divided by the total number of polarity aspects found on ReLi corpus.

4.1 Individual Lexicons

The results obtained with each method, individually using the lexicons, are shown in Fig. 1. These experiments evaluate the methods of identifying polarity and our strategy for the creation of LexReLi lexicon.

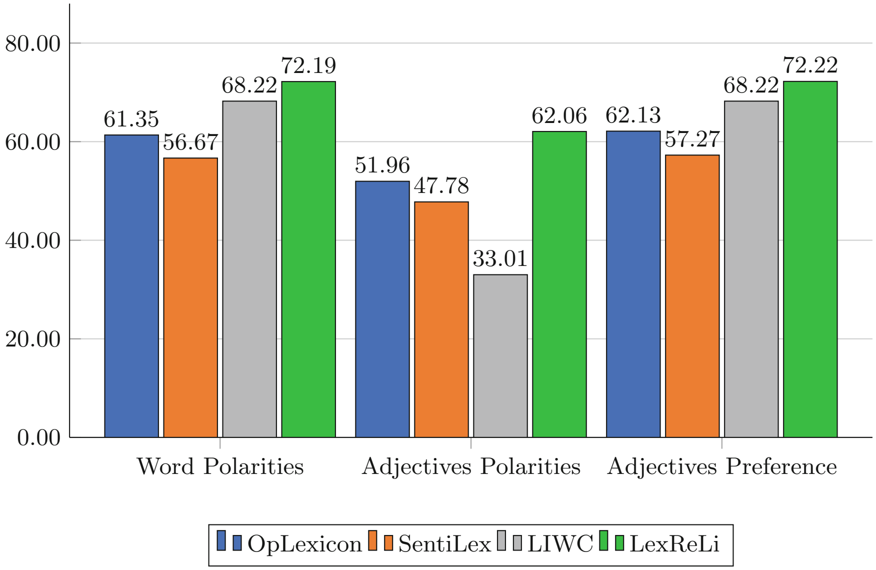


Fig. 1. Accuracy (%) for experiments with individual lexicons.

In the first experiment, we combined the Aspect Based polarity analysis with the phrase polarity analysis identified by the polarity of the words, through the method Words Polarities. Among the three standard lexicons we tested, the LIWC obtained the best result with an accuracy of 68.22%, showing that the size of the lexicon does not define its quality. The largest lexicon, the SentiLex with 79,190 words, obtained the worst result, 56.67% of accuracy. The intermediary size lexicon, OpLexicon, with 32,119 words, obtained 61.35% of accuracy. The proposed LexReLi lexicon, specialized in the literary context, despite the 1,543 words, obtained the best result in this experiment, with 72.19% of accuracy.

In our second experiment, we evaluated, on phrase level, the detection of polarity analyzing only the adjectives, using the Adjectives Polarities method. In this experiment, the LexReLi obtained the best result, reaching 62.06% of accuracy. The Lexicon OpLexicon obtained an accuracy of 51.96%, followed by SentiLex with 47.78%. Finally, came the LIWC with 33.01%.

As a final experiment, we evaluated the Adjectives Preference method that gives priority to adjectives in the sentence-level analysis. The results were slightly higher than the first experiment, with LexReLi obtaining the best result with an accuracy of 72.22%. The LIWC obtained 68.22%, OpLexicon 62.13% and SentiLex 52.27%.

4.2 Combined Lexicons

As explained, we used various approaches to combine lexicons. We have a first lexicon that has undergone a conciliation process where possible conflicts between combined lexicons have been solved. We also have six lexicons formed by the combination of LIWC, OpLexicon, and SentiLex, where what has changed between one and other was the order the words were included in the lexicons. Finally, we presented LexReLi combined with these same six last lexicon combinations. In these experiments, we tried to evaluate these lexicon construction techniques. Figure 2 shows the obtained results.

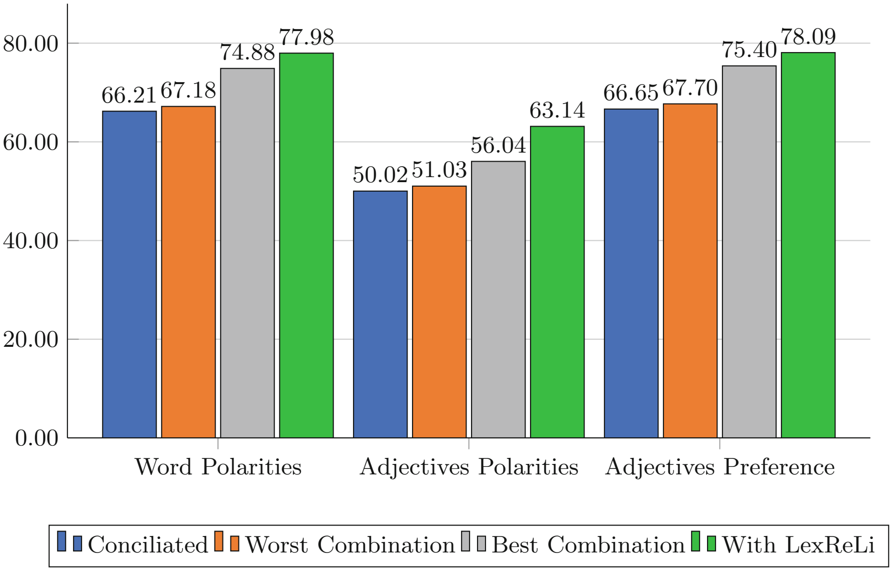


Fig. 2. Accuracy (%) for experiments with the combined lexicons.

Regarding the methods of analysis, the results were very similar to the experiments with the individual lexicons. The adjectives-preferred approach yielded the best results, but these were little better than the approach that analyzes the polarity of words. The method that analyzes only the adjectives obtained the lowest results.

The lexicon obtained in the conciliation process, although this seems to be the correct way to combine lexicons, was the one that obtained the lowest results. It obtained 66.21% of accuracy when analyzed through word polarities, 50.02% only by adjectives, and 66.65% when using the method that gives preference to adjectives.

In the combination of lexicons, we obtained, for each method of analysis, six different results. We present here only the smallest and highest obtained results. For the lexicon combinations with LexReLi, we only present the best result, since the difference between the worst and the best results was not as expressive as the previous combinations.

For word polarities method, the combination made in the order LIWC + SentiLex + OpLexicon obtained the best accuracy with 74.88%. The lowest result was obtained with SentiLex + OpLexicon + LIWC with 67.18% of accuracy. The LexReLi + LIWC + SentiLex + OpLexicon combination achieved the best result of the experiments with the method, reaching an accuracy of 77.98%.

The method that uses only adjectives for analysis, as well as in the experiment carried out with the individual lexicons, obtained the lowest results. The best result, 63.14% of accuracy, was obtained for the combination LexReLi + LIWC + OpLexicon + SentiLex. Without LexReLi, the LIWC + OpLexicon + SentiLex combination achieved 56.04% and the SentiLex + OpLexicon + LIWC combination had the lowest result with 51.03% of accuracy.

In our last experiment, we tested again the preference for adjectives method and obtained the highest results among all the performed experiments, although they were little superior to the method that analyzes only the polarity of words. The LexReLi + LIWC + OpLexicon + SentiLex combination obtained the highest result with 78.09% of accuracy. Without LexReLi, the highest result was obtained by LIWC + SentiLex + OpLexicon with 75.40%, and the lowest by SentiLex + OpLexicon + LIWC with an accuracy of 67.70%.

5 Conclusion

We may conclude that the approaches used to create and combine lexicons were effective, given the evident improvement in the results of the application of the methods using these lexicons. The method used to create LexReLi may be easily applied used for the creation of lexicons for different contexts. Improvements in the methods of polarity analysis used in its creation would also imply improvements in the resulted lexicon.

The lexicon combination method that used a conciliation approach proved to be less effective because, in addition to being more laborious in creation, it obtained inferior results in relation to the methods that used a simple combination of lexicons.

In the combination approaches, we obtained indications that it is a good practice to give priority to the lexicon that obtained the best individual result, since in the experiments we observed that the combinations initiated by the LIWC lexicon achieved better results.

Although slightly superior, the results of the preference to the adjectives method are promising. Improvements may be implemented in the two steps of the approach. In the first stage, one may work with advanced methods for detecting words that qualify aspects. In the second step, it is possible to improve or apply other methods for detecting polarity at the phrase level.

References

1. Balage Filho, P.P.P., Aluísio, S., Pardo, T.T.: An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: 9th Brazilian Symposium in Information and Human Language Technology – STIL, pp. 215–219 (2013)
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Inc., Sebastopol (2009)
3. Fonseca, E.R., Rosa, J.L.G.: Mac-morpho revisited: Towards robust part-of-speech tagging. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pp. 98–107 (2013)
4. Freitas, C., Motta, E., Milidiú, R.L., César, J.: Vampiro que brilha... Rá! Desafios na anotação de Opinião em um corpus de resenhas de livros. In: XI Encontro de Linguística de Corpus (ELC 2012), São Paulo (2012)
5. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
6. Hussein, D.M.E.-D.M.: A survey on sentiment analysis challenges. J. King Saud Univ. Eng. Sci. (2016). https://doi.org/10.1016/j.jksues.2016.04.002. ISSN 1018-3639
7. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
8. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
9. Pavlopoulos, J., Androutsopoulos, I.: Aspect term extraction for sentiment analysis: new datasets, new evaluation measures and an improved unsupervised method. In: Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL, pp. 44–52 (2014)
10. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 Task 12: Aspect based sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 486–495. Association for Computational Linguistics, Denver, Colorado (2015)
11. Silva, M.J., Carvalho, P., Costa, C., Sarmento, L.: Automatic expansion of a social judgment lexicon for sentiment analysis. Technical report TR 1008. University of Lisbon, Faculty of Sciences LASIGE 6694, December 2010. http://repositorio.ul.pt/handle/10455/6694
12. Souza, M., Vieira, R.: Sentiment analysis on Twitter data for portuguese language. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS (LNAI), vol. 7243, pp. 241–247. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28885-2_28
13. Souza, M., Vieira, R., Chishman, R., Alves, I.M.: Construction of a portuguese opinion Lexicon from multiple resources. In: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pp. 59–66 (2011)
14. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
15. Thet, T.T., Na, J.C., Khoo, C.S.: Aspect-based sentiment analysis of movie reviews on discussion boards. J. Inf. Sci. **36**(6), 823–848 (2010)

A Parallel Corpus of Theses and Dissertations Abstracts

Felipe Soares[1,2](✉), Gabrielli Harumi Yamashita[2], and Michel Jose Anzanello[2]

[1] Instituto de Informática, UFRGS, Porto Alegre, Brazil
felipe.soares@inf.ufrgs.br
[2] Escola de Engenharia, UFRGS, Porto Alegre, Brazil
gabrielli.hy@gmail.com, anzanello@producao.ufrgs.br

Abstract. In Brazil, the governmental body responsible for overseeing and coordinating post-graduate programs, CAPES, keeps records of all theses and dissertations presented in the country. Information regarding such documents can be accessed online in the Theses and Dissertations Catalog (TDC), which contains abstracts in Portuguese and English, and additional metadata. Thus, this database can be a potential source of parallel corpora for the Portuguese and English languages. In this article, we present the development of a parallel corpus from TDC, which is made available by CAPES under the open data initiative. Approximately 240,000 documents were collected and aligned using the Hunalign tool. We demonstrate the capability of our developed corpus by training Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) models for both language directions, followed by a comparison with Google Translate (GT). Both translation models presented better BLEU scores than GT, with NMT system being the most accurate one. Sentence alignment was also manually evaluated, presenting an average of 82.30% correctly aligned sentences. Our parallel corpus is freely available in TMX format, with complementary information regarding document metadata.

Keywords: Parallel corpus · Scientific abstracts · Portuguese/English

1 Introduction

The availability of cross-language parallel corpora is one of the basis of current Statistical and Neural Machine Translation systems (e.g. SMT and NMT). Acquiring a high-quality parallel corpus that is large enough to train MT systems, specially NMT ones, is not a trivial task, since it usually demands human curating and correct alignment. In light of that, the automated creation of parallel corpora from freely available resources is extremely important in Natural Language Processing (NLP), enabling the development of accurate MT solutions. Many parallel corpora are already available, some with bilingual alignment, while others are multilingually aligned, with 3 or more languages, such as Europarl [3],

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 345–352, 2018.
https://doi.org/10.1007/978-3-319-99722-3_35

from the European Parliament, JRC-Acquis [9], from the European Commission, OpenSubtitles [12], from movies subtitles.

The extraction of parallel sentences from scientific writing can be a valuable language resource for MT and other NLP tasks. The development of parallel corpora from scientific texts has been researched by several authors, aiming at translation of biomedical articles [6,11], or named entity recognition of biomedical concepts [5]. Regarding Portuguese/English and English/Spanish language pairs, the FAPESP corpus [1], from the Brazilian magazine *revista pesquisa FAPESP*, contains more than 150,000 aligned sentences per language pair, constituting an important language resource.

In Brazil, the governmental body responsible for overseeing post-graduate programs across the country, called CAPES, tracks every enrolled student and scientific production. In addition, CAPES maintains a freely accessible database of theses and dissertations produced by the graduate students (i.e. Theses and Dissertations Catalog - TDC) since 1987, with abstracts available since 2013. Under recent governmental efforts in data sharing, CAPES made TDC available in CSV format, making it easily accessible for data mining tasks. Recent data files, from 2013 to 2016, contain valuable information for NLP purposes, such as abstracts in Portuguese and English, scientific categories, and keywords. Thus, TDC can be an important source of parallel Portuguese/English scientific abstracts.

In this work, we developed a sentence aligned parallel corpus gathered from CAPES TDC comprised of abstracts in English and Portuguese spanning the years from 2013 to 2016. In addition, we included metadata regarding the respective theses and dissertations.

2 Materials and Methods

In this section, we detail the information retrieved from CAPES website, the filtering process, the sentence alignment, and the evaluation experiments. An overview of the steps employed in this article is shown in Fig. 1.

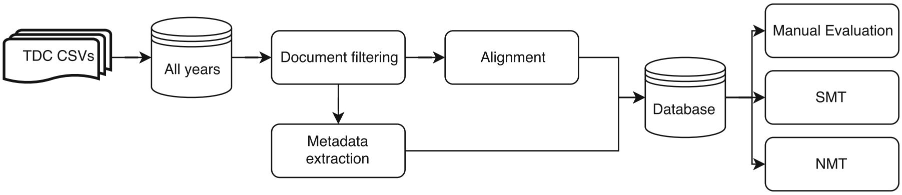


Fig. 1. Steps employed in the development of the parallel corpora.

2.1 Document Retrieval and Parsing

The TDC datasets are available in the CAPES open data website[1] divided by years, from 2013 to 2016 in CSV and XLSX formats. We downloaded all CSV files from the respective website and loaded them into an SQL database for better manipulation. The database was then filtered to remove documents without both Portuguese and English abstracts, and additional metadata selected.

After the initial filtering, the resulting documents were processed for language checking[2] to make sure that there was no misplacing of English abstracts in the Portuguese field, or the other way around, removing the documents that presented such inconsistency. We also performed a case folding to lower case letters, since the TDC datasets present all fields with uppercase letters. In addition, we also removed newline/carriage return characters (i.e \n and \r), as they would interfere with the sentence alignment tool.

2.2 Sentence Alignment

For sentence alignment, we used the LF aligner tool[3], a wrapper around the Hunalign tool [10], which provides an easy to use and complete solution for sentence alignment, including pre-loaded dictionaries for several languages.

Hunalign uses Gale-Church sentence-length information to first automatically build a dictionary based on this alignment. Once the dictionary is built, the algorithm realigns the input text in a second iteration, this time combining sentence-length information with the dictionary. When a dictionary is supplied to the algorithm, the first step is skipped. A drawback of Hunalign is that it is not designed to handle large corpora (above 10 thousand sentences), causing large memory consumption. In these cases, the algorithm cuts the large corpus in smaller manageable chunks, which may affect dictionary building.

The parallel abstracts were supplied to the aligner, which performed sentence segmentation followed by sentence alignment. A small modification in the sentence segmentation algorithm was performed to handle the fact that all words are in lowercase letters, which originally prevented segmentation. After sentence alignment, the following post-processing steps were performed: (i) removal of all non-aligned sentences; (ii) removal of all sentences with fewer than three characters, since they are likely to be noise.

2.3 Machine Translation Evaluation

To evaluate the usefulness of our corpus for SMT purposes, we used it to train an automatic translator with Moses [4]. We also trained an NMT model using the OpenNMT system [2], and used the Google Translate Toolkit[4] to produce

[1] https://dadosabertos.capes.gov.br/dataset/catalogo-de-teses-e-dissertacoes-de-2013-a-2016.

[2] https://github.com/Mimino666/langdetect.

[3] https://sourceforge.net/projects/aligner/.

[4] https://translate.google.com/toolkit/.

state-of-the-art comparison results. The produced translations were evaluated according to the BLEU score [7].

2.4 Manual Evaluation

Although the Hunalign tool usually presents a good alignment between sentences, we also conducted a manual validation to evaluate the quality of the aligned sentences. We randomly selected 400 pairs of sentences. If the pair was fully aligned, we marked it as “correct”; if the pair was incompletely aligned, due to segmentation errors, for instance, we marked it as “partial”; otherwise, when the pair was incorrectly aligned, we marked it as “no alignment”.

3 Results and Discussion

In this section, we present the corpus’ statistics and quality evaluation regarding SMT and NMT systems, as well as the manual evaluation of sentence alignment.

3.1 Corpus Statistics

Table 1 shows the statistics (i.e. number of documents and sentences) for the aligned corpus according to the 9 main knowledge areas defined by CAPES. The dataset is available[5] in TMX format [8], since it is the standard format for translation memories. We also made available the aligned corpus in an SQLite database in order to facilitate future stratification according to knowledge area, for instance. In this database, we included the following metadata information: year, university, title in Portuguese, type of document (i.e. theses or dissertation), keywords in both languages, knowledge areas and subareas according to CAPES, and URL for the full-text PDF in Portuguese. An excerpt of the corpus is shown in Table 2.

Table 1. Corpus statistics according to knowledge area.

Knowledge area	Docs	Sents	Tokens EN	Tokens PT
Health sciences	38, 221	224, 773	5.46M	5.51M
Humanities	38, 493	189, 648	5.63M	5.54M
Applied social sciences	32, 176	160, 131	4.66M	4.60M
Agricultural sciences	26, 740	154, 710	3.92M	3.92M
Engineering	27, 074	149, 888	3.87M	3.92M
Multidisciplinary	26, 502	140, 849	3.84M	3.81M
Exact and earth sciences	19, 630	106, 098	2.64M	2.66M
Biological sciences	16, 465	98, 994	2.33M	2.34M
Linguistic and arts	13, 717	64, 281	1.99M	1.96M
Total	239, 018	1, 289, 372	34.35M	34.28M

[5] https://figshare.com/s/6f760a4f3610a83c2e3f.

3.2 Translation Experiments

Prior to the MT experiments, sentences were randomly split in three disjoint datasets: training, development, and test. Approximately 13,000 sentences were allocated in the development and test sets, while the remaining was used for training. For the SMT experiment, we followed the instructions of Moses baseline system[6]. For the NMT experiment, we used the Torch implementation[7] to train

Table 2. Excerpt of the corpus with document ID.

ID	Portuguese	English
127454	nessa tese apresentamos duas linhas de pesquisa distintas, a saber, na primeira, referente aos capítulos 1 e 3 aplicamos técnicas estatísticas à análise de imagens do satélite de abertura sintética (sar) e, na segunda, referente ao capítulo 2, examinamos problemas relativos à estimação de parâmetros por máxima verossimilhança na distribuição exponencial-poisson	in this thesis we present two distinct research lines, namely, the first, referring to chapters 2 and 3, apply statistical techniques to the analysis of synthetic aperture radar (sar) images, and the second, referring to chapter 4, we examined problems concerning parameter estimation by maximum likelihood in exponential-poisson distribution
1419264	para determinação dessa flora utilizamos os recursos de observação, coleta e identificação	we use the resources of investigation, collection and identification to determine this flora
439358	estimaram-se os benefícios ambientais da reciclagem de veículos com mais de 10 anos de uso, considerando os poluentes na fabricação de um veículo novo	we estimated the environmental benefits of recycling vehicles in use more than 10 years, taking into consideration pollution engendered in the manufacture of a new vehicle
675023	a coleta de dados se deu por meio de entrevista semiestruturada com 12 familiares cuidadores de crianças atendidas em pronto-socorro pediátrico de um hospital de ensino	data collection was through semi-structured interviews with 12 family caregivers of children seen in a pediatric emergency department of a teaching hospital
675023	os dados foram submetidos á análise de conteúdo temático conforme bardin (2011)	the data were subjected to thematic content analysis according to bardin (2011)
1173306	o planejamento e programação do projeto de construção naval têm dois objetivos por base: diminuir o tempo de fabricação e os custos	shipbuilding project planning and scheduling possess two major objectives: manufacturing time and cost reduction

[6] http://www.statmt.org/moses/?n=moses.baseline.

[7] http://opennmt.net/OpenNMT/.

a 2-layer LSTM model with 500 hidden units in both encoder and decoder, with 12 epochs. During translation, the option to replace UNK words by the word in the input language was used, since this is also the default in Moses.

Table 3 presents the BLEU scores for both translation directions with English and Portuguese on the development and test partitions for Moses and OpenNMT models. We also included the scores for Google Translate (GT) as a benchmark of a state-of-the-art system which is widely used.

Table 3. BLEU scores for the translations using Moses, OpenNMT, and Google Translate. Bold numbers indicate the best results in the test set.

Partition	System	PT → EN	EN → PT
Dev	Moses	44.07	41.21
	OpenNMT	44.02	43.36
Test	Moses	43.85	41.05
	OpenNMT	**43.89**	**43.22**
	Google Translate	42.57	38.92

NMT model achieved better performance than the SMT one for EN → PT direction, with approximately 2.17% points (pp) higher, while presenting almost the same score for PT → EN. When comparing our models to GT, both of them presented better BLEU scores, specially for the EN → PT direction, with values ranging from 1.27 pp to 4.30 pp higher than GT.

We highlight that these results may be due to two main factors: corpus size, and domain. Our corpus is fairly large for both SMT and NMT approaches, comprised of almost 1.3M sentences, which enables the development of robust models. Regarding domain, GT is a generic tool not trained for a specific domain, thus it may produce lower results than a domain specific model such as ours. Scientific writing usually has a strict writing style, with less variation than novels or speeches, for instance, favoring the development of tailored MT systems.

Below, we demonstrate some sentences translated by Moses and OpenNMT compared to the suggested human translation. One can notice that in fact NMT model tend to produce more fluent results, specially regarding verbal regency.

> Human translation: *this paper presents a study of efficiency and power management in a packaging industry and plastic films.*
>
> OpenNMT: *this work presents a study of efficiency and electricity management in a packaging industry and plastic films.*
>
> Moses: *in this work presents a study of efficiency and power management in a packaging industry and plastic films.*
>
> GT: *this paper presents a study of the efficiency and management of electric power in a packaging and plastic film industry.*

Human translation: *this fact corroborates the difficulty in modeling human behavior.*

OpenNMT: *this fact corroborates the difficulty in modeling human behavior.*

Moses: *this fact corroborated the difficulty in model the human behavior.*

GT: *this fact corroborates the difficulty in modeling human behavior.*

3.3 Sentence Alignment Quality

We manually validated the alignment quality for 400 sentences randomly selected from the parsed corpus and assigned quality labels according Sect. 2.4. From all the evaluated sentences, 82.30% were correctly aligned, while 13.33% were partially aligned, and 4.35% presented no alignment. The small percentage of no alignment is probably due to the use of Hunalign tool with the provided EN/PT dictionary.

Regarding the partial alignment, most of the problems are result of segmentation issues previous to the alignment, which wrongly split the sentences. Since all words were case folded to lowercase letters, the segmenter lost an important source of information for the correct segmentation, generating malformed sentences. Some examples of partial alignment errors are shown in Table 4, where most sentences were truncated in the wrong part.

Table 4. Examples of partial alignment errors.

Portuguese	English
os dados foram comparados entre os grupos por anova de medidas repetida	data were compared by repeated measures anova. results: we found a significa
o estudo utilizará um software commercial para simular a peça	the study will use commercial software to simulate the piece with a number of different crack sizes and the
buscamos subsídios teóricos em autores que veem na reflexão e na pesquisa um grande potencial para o desenvolvimento d	we seek theoretical support in authors who see in reflection and research a great potential for

4 Conclusion and Future Work

We developed a parallel corpus of theses and dissertations abstracts in Portuguese and English. Our corpus is based on the CAPES TDC dataset, which contains information regarding all theses and dissertations presented in Brazil from 2013 to 2016, including abstracts and other metadata.

Our corpus was evaluated through SMT and NMT experiments with Moses and OpenNMT systems, presenting superior performance regarding BLEU score than Google Translate. The NMT model also presented superior results than the SMT one for the EN → PT translation direction. We also manually evaluated sentences regarding alignment quality, with average 82.30% of sentences correctly aligned.

For future work, we foresee the use of the presented corpus in mono and cross-language text mining tasks, such as text classification and clustering. As we included several metadata, these tasks can be facilitated. Other machine translation approaches can also be tested, including the concatenation of this corpus with other multi-domain ones.

References

1. Aziz, W., Specia, L.: Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In: STIL 2011, Cuiabá, MT, October 2011
2. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: Open-Source Toolkit for Neural Machine Translation. arXiv e-prints (2017)
3. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: MT Summit, vol. 5, pp. 79–86 (2005)
4. Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. Association for Computational Linguistics (2007)
5. Kors, J.A., Clematide, S., Akhondi, S.A., Van Mulligen, E.M., Rebholz-Schuhmann, D.: A multilingual gold-standard corpus for biomedical concept recognition: the mantra GSC. J. Am. Med. Inform. Assoc. **22**(5), 948–956 (2015)
6. Neves, M., Yepes, A.J., Névéol, A.: The scielo corpus: a parallel corpus of scientific publications for biomedicine. In: Chair, N.C.C., et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France, May 2016
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
8. Rawat, S., Chandak, M.B., Chauhan, N.: An approach for efficient machine translation using translation memory. In: Unal, A., Nayak, M., Mishra, D.K., Singh, D., Joshi, A. (eds.) SmartCom 2016. CCIS, vol. 628, pp. 285–291. Springer, Singapore (2016). https://doi.org/10.1007/978-981-10-3433-6_34
9. Steinberger, R., et al.: The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages. arXiv preprint arXiv:cs/0609058 (2006)
10. Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., Tron, V.: Parallel corpora for medium density languages. Amsterdam Studies in The Theory and History of Linguistic Science, p. 247 (2007)
11. Wu, C., Xia, F., Deleger, L., Solti, I.: Statistical machine translation for biomedical text: are we there yet? In: AMIA Annual Symposium Proceedings, vol. 2011, p. 1290. American Medical Informatics Association (2011)
12. Zhang, S., Ling, W., Dyer, C.: Dual subtitles as parallel corpora. In: Calzolari, N., et al. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, 26–31 May. European Language Resources Association (ELRA) (2014). ISBN 978-2-9517408-8-4

Development of a Brazilian Portuguese Hotel's Reviews Corpus

Joana Gabriela Ribeiro de Souza, Alcione de Paiva Oliveira(✉), and Alexandra Moreira

Universidade Federal de Viçosa, Viçosa, MG 36570-900, Brazil
joana.souza@ufv.br, alcione@gmail.com

Abstract. The provision of voluntary textual information mediated by the Internet, and particularly by Web 2.0, provided an opportunity for the creation of large linguistic corpora. These corpora can serve as a fundamental resource for the development of applications focused on natural language, especially those using deep learning techniques that require big datasets. One type of application that benefits from these resources is the ones that perform sentiment analysis. This article describes the creation of corpus aimed to support sentiment analysis applications. It consists of reviews hotels located in the Brazilian capitals and the Federal District, written in Brazilian Portuguese language. The reviews that make up the corpus have been taken from TripAdvisor and have undergone normalization and POS tagging. The primary goal is to make it available to the community to be used in machine learning tasks geared toward natural language.

Keywords: Linguistic corpus · Portuguese corpus · Hotel's reviews
Sentiment analysis

1 Introduction

The Web allows users to interact, collaborate with each other and share information on a variety of subjects. There are digital spaces such as social networks, forums, e-commerce sites among others where people usually express their opinions. In these environments the users usually make use of an informal language, being very common the use of slang, abbreviations, different spellings of words existing in the grammar, besides the creation of new terms. These are some of the challenges encountered by researchers in compiling corpus with data scraped from the Internet.

Corpora are key resources for the training and testing of applications focused on natural language processing. Nevertheless the creation of these features can be quite complex and time-consuming. There are several compromise decisions and processing steps, such as normalization and annotation. According to [4] the Web has been widely used as a corpus, due to the amount of data available in several languages and textual genres, free of charge and easy to access. [6] presented the

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 353–361, 2018.
https://doi.org/10.1007/978-3-319-99722-3_36

difficulties and effort involved in extracting information from the Web. In order to extract and store textual data from the Internet, in order to become useful for NLP tools, it is necessary the adoption of Web scraper tools and to apply several pre-process steps to the data to remove undesirable "noises". Occurrences such as the use of acronyms, terms in different languages, emojis, non-formal use of the Portuguese language among other situations are common in texts written by people on the Internet. Patil [6] also points out the challenges for extracting Web texts due to the variety of content and formats available that vary from one genre to another (from a social network to a government website, for instance). Meyer et al. [5], listed important details to consider when deciding the use of the Web as a corpus, such as selecting the appropriate search tool, the necessary pre-processing, among others. [2,3] also point out interesting normalization that were taken in account in our development of a corpus composed of reviews in Brazilian Portuguese.

This work was motivated due to the need for Brazilian Portuguese corpora for performing sentiment analysis. For this task, we choose hotel reviews to make up our corpus considering its availability on the Web. The textual information was collected from a well-known tourist attraction review site and went through several stages of processing, from the adjustments to the text to the addition of syntactic annotation, which will be described in this article. The corpus developed was made available on the Internet in order to help meet the need for corpus in Brazilian Portuguese.

This article describes the steps performed in the development of a corpus of hotel reviews in Brazilian Portuguese for performing sentiment analysis. Section 2 presents some related work. Section 3 describes the compilation of the corpus composed of texts written by people who have stayed in hotels in the 26 Brazilian capitals and the Federal District. Section 4 presents the process of analysis and annotation of the types of occurrences that originate terms outside the vocabulary in this type of text. The procedures involved in pre-processing and normalization are described in Sect. 4, and finally we present a brief conclusion along with the indication of future work in Sect. 5.

2 Related Works

The construction of corpus is one of the fundamental tasks of the natural language processing area. Here we will present some work related to the development corpus written in the Portuguese language.

[2] points out some questions about the standardization of a product reviews corpus in Portuguese. On this corpus the authors performed the semantic role labeling, sentiment analysis, classification and summarization. They used the MXPOST tagger for the part-of-speech (POS) annotation, and used a small portion of the corpus for measuring the accuracy. Afterwards, they manually created 4 golden corpora following the spelling normalization (including foreign words and named entities); case; punctuation; and use the of Internet slang. They observed that case information got the highest correction rate, albeit they

expected that slang would have greater impact (only 0.24% of the corpus sample). They selected four tasks to be performed: (1) true casing normalization using Named-entity recognition (NER) as one of the main strategies; (2) punctuation problems correction; (3) spelling correction using Unitex and then manually checking common words to evaluate if the word had been accurately corrected, and; (4) Internet slang normalization, where the words were categorized as written different from cultured norm and abbreviations, repetition of signs and letters, and sequences related to emotions (such as emoticons).

[3] described how the corpus used in [2] has been compiled and the strategies used to normalize it. They defined 8 categories based on the types of noise found in the corpus: misspellings, acronym, proper name, abbreviation, Internet slang, foreign word, unit of measure, and unrated or doubtful tokens. As a result, they made lists with the most common terms for each category and concatenated it in a tool where it is possible to carry out the normalization of a text (especially in the context of product reviews). In the list they identified and corrected acronym, proper name, abbreviation, Internet slang, as well as the spelling according to the spell checker Aspell.

[1] presented two types of normalization (destructive and non-destructive) and an architecture developed by them in order to normalize a corpus in German, without losing information that, for a POS tagger, can be considered "noises", but may give important clues about non-standard language. The architecture was based on the use of two normalization techniques. First a destructive normalization was performed, using HyDRA that unite a rule pattern with n-grams frequency to define when a word actually contains a hyphen, correcting it. Subsequently a non-destructive normalization that aims to maintain the "noise" of the corpus by rewriting words that are emphatically written ("looooooove") or with typing errors. It uses an annotation layer, so it doesn't lose information that may be useful for sentiment analysis or for discovering new aspects of the language, for instance.

In [7] it is described how the corpus CETEM/Público was created (Corpus of Extracts of Electronic Texts MCT/Public). A corpus developed with the support of the Ministry of Science and Technology (MCT). The CETEM/Público was created from journalistic texts from newspaper "O público" founded in 1990. The newspaper is written in Portuguese, being almost exclusively from European Portuguese texts, with exceptions of texts written by Brazilians and Africans. The creation steps involved cleaning of texts of images, subject classification, and sentence separation using the program library developed in the AC/DC project. The result was a corpus of 180 million words. The main difference is from the present work is that the corpus is not aimed to sentiment analysis.

3 The Corpus Development

As the source of textual information to build up our corpus of hotel reviews we have chosen the site TripAdvisor[1], since it is one of the most used sites,

[1] https://www.tripadvisor.com.br.

by travelers, to evaluate not only hotels, but also several other types of tourist attractions and related services. Many tourist related establishment presents at their front door or reception desk the seal of recommendation and/or the quality certificate giving by TripAdvisor. Our corpus is only composed of hotel reviews, so we will always talk about this evaluation context in this paper. On TripAdvisor when making a review, users should enter the number of circles (meaning similar to stars) corresponding to the overall evaluation of the hotel, give a title to the evaluation (which can be understood as a summary of the review), the evaluation (with at least 200 characters where one can give more details about the stay), choose the month and year of the visit, as well as other non-mandatory information. We collected four information from the evaluations: the number of circles, the title, the evaluation and instead of the date of the stay, we collected the date on which the evaluation was performed.

The data were collected only from accommodations classified as hotels. The reviews were collected from February to March 2018, so the most recent review dates from March 20, 2018. Reviews were taken from hotels of the 26 capitals of the Brazilian states and the Federal District as well. We chose to collect reviews of hotels in the capitals in order to have a clear criteria of delimitation of the number of cities and to cover all Brazilian states as well. We gathered a total of 730,069 reviews. Until this moment the corpus had not gone through any linguistic pre-processing, being just removed the HTML tags. Using the NLTK (Natural Language Toolkit)[2], a Python library for NLP, it was verified that the corpus contained 55,950,007 tokens and 457,337 types. Among the most common words are: hotel (hotel), não (no), bem (well or good), manhã (morning), bom (good), quarto (bedroom), café (coffee), localização (location), atendimento (attendance or service) and excelente (excellent). What is expected given the context of the evaluations. In contrast, the less used terms refer to writing errors, very common due to the different levels of literacy of the users.

The Fig. 1 presents in summary form the steps followed by us for the development of the corpus. Firstly, the capture of hotel reviews yielded four files (dates, notes, titles and comments) for each hotel of that capital. After this process, we gathered all the files by city, by region and later in only four files containing all the data collected row by row sequentially, in such way that the i-th line of the dates file corresponds to the evaluation date that is on i-th line in the comments file and so on. After this junction we performed the first normalization where we remove all the HTML tags and in the case of the dates we converted from the format "2 de janeiro de 2018" (January 2, 2018) to "2/01/2018" and for grades we converted the internal code used by TripAdvisor (from "bubble_10", to "bubble_50") to numbers from "1" to"5". Later we did the second normalization where we tried to make several non-destructive corrections according to [1] resulting two versions of the comments: one with the normalizations for tokenization described in the later section and another also without the stopwords.

[2] https://www.nltk.org.

We provided four files (comments normalized, dates, grades and titles) with free access for the community[3].

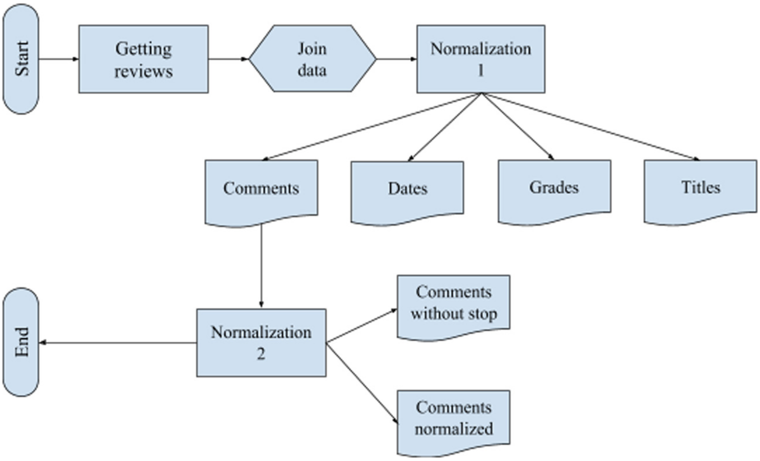


Fig. 1. The steps of the process of the corpus creation and normalization.

4 The Corpus Analysis and Normalization

We used NLTK to get a general idea of the size of the corpus and we studied the related works in order to establish a basis for normalizations that could be applied. The first normalization we did in the corpus was the removal of the excesses of punctuation and sequence of repeated letters that did not form words. As TripAdvisor requires the user to write a comment of at least 200 characters, in several occasions the users completed the comments with punctuation sequences and random characters that did not form words, so we removed several of these occurrences. We kept reticence and sequences of up to three exclamations or questions marks (which may have some meaning when it comes to the sentiment analysis). We have developed a lexical dataset to help us to reduce the number of words that were linked to each other (e.g. "Bomcafédamanhã"). We separated terms such as numbers or hyphens (preceded and followed by spaces) linked to words (e.g. "8 Limpeza" to "8 Limpeza" and "-Gostei" to "- Gostei"). In addition, we kept the words the way they were written, even if incorrectly, due to typing errors or intentionally. So we kept terms with "adoreeeeei" and "lliiiixxxoooo" (similar to "I loveeeeed it" and "trrraaaaaassshh", respectively). On the Internet it is common to write uppercase terms as a way of emphasizing something either positively or negatively, for this reason we also kept the texts capitalization intact.

[3] The corpus files are available at: https://bit.ly/2JVRJbI.

We observed that the corpus had several types of errors in the formatting of the words that prevented them from being tokenized correctly, and those errors did not fit the types of errors or "noises" mentioned previously. Table 1 indicates examples of occurrence that were very common in the corpus and the correction made the tokenizer more efficient. After this normalization the number of tokens increased (even with previously destructive normalization) and the number of types was considerably reduced since the corrections of these occurrences produced more words that could be recognized and counted correctly. By doing some empirical tests we noticed that NLTK tokenizer fails in some common cases, as in the examples shown in Table 1. We also tested the Spacy[4] tokenizer which is an API also developed in Python for NLP which according to developers is the fastest tool and provides the most features up to date. In addition, Spacy supports deep learning, which is a hot topic these days. However, considering the cases in Table 1 Spacy does not separate all the tokens, even though it separates a few more cases that the NLTK is not able to handle ("calçada.." tokenized for ["calçada",".","."] by Spacy and ["calçada.."] by the NLTK), but we have corrected many of these problems with the standardization process. The reason why we continue to use NLTK as a NLP tool in this work is due to its approach of treating words that contain hyphens. NLTK keeps the term as a single token ("wi-fi" tokenized to ["wi-fi"]) while Spacy treats it as distinct tokens ("wi-fi" tokenized to ["wi", "-","fi"]). As one of the objectives of the work is to make the corpus available to the community, we opted to keep these terms in this way. After picking the tool and carrying out the normalizations described here, we have obtained a count of 56,743,114 tokens and 246,307 types. Considering stop-words and punctuation signs, we can see that each review, on average, consists of about 77 tokens, with the largest review having 2,857 tokens and the lowest having only 2 tokens.

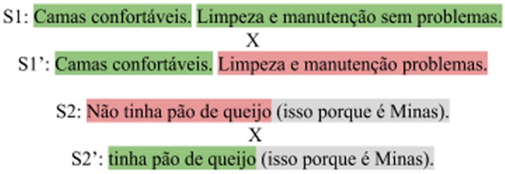
Table 1. Some occurrences found in the corpus

Occurrence	Correction
Muito.Porém	Muito.Porém
Residência..	Residência...
*apartamentos	*apartamentos
Custo/benefício	Custo/benefício
2 km	2 km

As them main purpose of the corpus is to be used in sentiment analysis applications, words such as "não" and"sem" (no and without, respectively) can change the meaning of the phrase by inverting its polarity, for example (in sentiment analysis the polarity of a term is basically its classification between the classes: positive, negative or neutral). Figure 2 contains examples of phrases found in the

[4] https://spacy.io.

corpus that when removing these words have their polarity altered. Due to this type of event, stopwords that could indicate change of polarity, intensity or clues to the next sentence classification were maintained in the normalized corpus.


S1: Camas confortáveis. Limpeza e manutenção sem problemas.
X
S1': Camas confortáveis. Limpeza e manutenção problemas.

S2: Não tinha pão de queijo (isso porque é Minas).
X
S2': tinha pão de queijo (isso porque é Minas).

Fig. 2. Examples of polarity changes by removing "sem" and"não". Where green, red and gray colors mean positive, negative and neutral polarity respectively. (Color figure online)

Thus, by manipulating the set of Portuguese stopwords incorporated in NLTK we kept the terms "não","mais", "mas","muito", "sem" and"nem" (no, more/plus, but, much/very, without and nor/neither, respectively). After this normalization we created a corpus of reviews without stopwords with 39,165,169 tokens. After these normalizations in the reviews, we produced a set of four files that can be used by the community for various purposes (we do not provide the stopwords file since we do not remove all of them from our corpus), as well as serve as a resource for our own research in sentiment analysis.

After some standardization, some analysis was done on the content of the corpus. The reviews are divided into five classes: horrible, bad, reasonable, very good and excellent. The Table 2 presents the distribution of the reviews in the five available classes, showing that there is a considerable imbalance between the negative, neutral and positive classes, whereas the negative classes (horrible and bad) together correspond to only 7.1% of the corpus, 16.6% of the reviews are neutral (reasonable) and 76.2% are positive (very good and excellent). Depending on the application, it would be interesting to balance these classes or make some kind of compensation in the machine learning algorithm.

Table 2. The table shows the unbalanced distribution of the reviews among the 5 classes

Class distribution	
Class	Percentage
Horrible	2.8%
Bad	4.3%
Reasonable	16.6%
Very good	40.2%
Excellent	36.0%

Although we have selected only Portuguese comments on the TripAdvisor site, the resulting corpus still contains emojis and many words in other languages, mainly terms in English as well as Chinese and Spanish. Several of these reviews date back to the 2014 World Cup, which brought tourists from all over the world to host cities. We emphasize that several of these occurrences mix Portuguese with other languages, and we chose to keep these terms as they were written, with no translations.

The Fig. 3 displays a graph with the corpus most common words, after eliminating the punctuation marks. Not surprisingly, prepositions are among the most common words. Furthermore, other frequent words are highly related to the

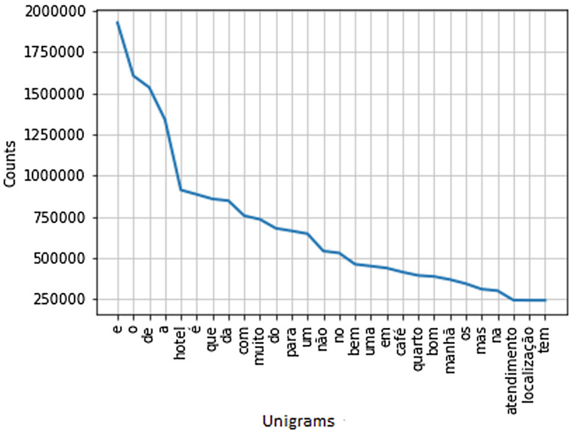


Fig. 3. Graph with the most common words/unigrams in the corpus.

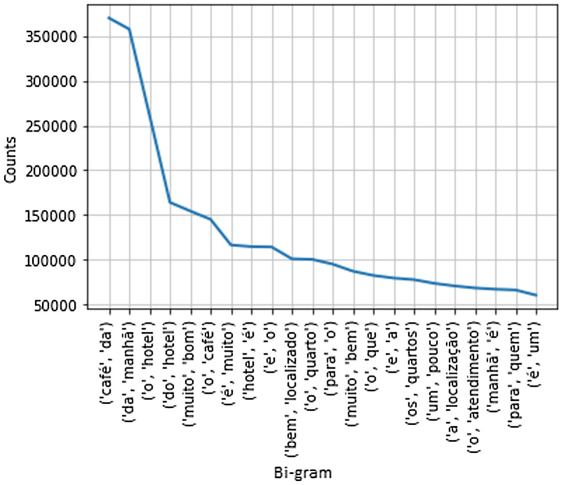


Fig. 4. Graph with the bi-gram distribution in the corpus.

hotel context, such as: hotel, coffee, room, service and location. Figure 4 shows the most common bi-grams in the corpus pointing out that terms like "caf/'e da manhã" (breakfast) and manhã (morning) are the most common terms, words that refers directly to the reviews' context.

5 Conclusions

The main contribution of this work is the production of a hotel review corpus with considerable size that will serve as a dataset for future work in sentiment analysis. We are making it free available to the community. We carried out a semi-automatic normalizations to reduce noise present in the corpus, but with the intention of keeping it as accurate as possible, considering that it is a corpus made up of texts extracted from the web. The corpus also can be used to aid in the tasks of extracting information, identifying patterns and new aspects present in the context of hotel evaluations among others. As future work, techniques such as [1] can be applied to spell checking the "noisy" terms while still keeping its original meaning, as well as normalizing the titles archive. Moreover, it is possible to adopt the use of multilingual methodologies to address cases of reviews written partially or entirely in languages other than Portuguese.

Acknowledgments. This research is supported in part by the funding agencies CAPES, FAPEMIG, and CNPq.

References

1. Bildhauer, F., and Schfer, R.: Token-level noise in large Web corpora and non-destructive normalization for linguistic applications. In: Proceedings of Corpus Analysis with Noise in the Signal (CANS 2013) (2013)
2. Duran, M.S. et al.: Some issues on the normalization of a corpus of products reviews in Portuguese. In: Proceedings of the 9th Web as Corpus Workshop (WaC-9), pp. 22–28 (2014)
3. Hartmann, N.S., et al.: A large corpus of product reviews in Portuguese: tackling out-of-vocabulary words. In: 9th European Language Resources Association-ELRA International Conference on Language Resources and Evaluation, pp. 3865–3871 (2014)
4. Kilgarriff, A., Grefenstette, G.: Introduction to the special issue on the web as corpus. Comput. linguist. **29**(3), 333–347 (2003)
5. Meyer, C.: The world wide web as linguistic corpus. Lang. Comput. **46**, 241–254 (2003)
6. Patil, P.: Application for data mining and web data mining challenges. Int. J. Comput. Sci. Mob. Comput. **6**(3), 39–44 (2017)
7. Rocha, P.A., Santos D.: CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Proceedings of V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000) (2000)

Sentiment Analysis & Opinion Mining

Aspect Clustering Methods for Sentiment Analysis

Francielle Alves Vargas and Thiago Alexandre Salgueiro Pardo(✉)

Interinstitutional Center for Computational Linguistics (NILC),
Institute of Mathematical and Computer Sciences, University of São Paulo,
São Carlos, Brazil
franciellealvargas@gmail.com, taspardo@icmc.usp.br

Abstract. Automatic aspect identification and clustering are critical tasks for opinion mining/sentiment analysis, as users employ varied terms (explicitly or not) to evaluate objects of interest and their characteristics. In this paper, we focus on aspect clustering methods and present a new approach to group implicit and explicit aspects from online reviews. We evaluate four linguistic methods inspired in the literature and one statistical method (using word embeddings), and also propose a new one, based on varied linguistic knowledge. We test the methods in three commonly used domains and show that the method that we propose significantly outperforms the other methods by a large margin.

Keywords: Clustering · Aspect-based sentiment analysis
Opinion mining

1 Introduction

The expansion of the social networks and e-commerce services resulted in the growth of online reviews in the web. Websites as Amazon and Buscape encourage users to write reviews for products, where users may do objective or subjective descriptions for a product and its aspects or properties. Subjective descriptions are characterized by a personal language, with opinions, sentiments, emotions and judgments. The research area in charge of identifying, extracting and summarizing subjective information in texts is called opinion mining or sentiment analysis [17,21,22]. According to [27], this area is different from the traditional text mining area, which is mostly based on objective topics rather than on subjective perceptions.

According to [24], due to the huge number of reviews in unstructured and varied formats, it is impractical for interested users to fully read and understand what other users comment. Therefore, the semantic organization of such information is mandatory, and the area of sentiment analysis consists in a first step towards this, enabling "mining" and synthesizing the relevant information, which may be employed by final users and companies for supporting their decision making process. Although its usefulness, according to [13], sentiment analysis represents a

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 365–374, 2018.
https://doi.org/10.1007/978-3-319-99722-3_37

"delicious challenge", as natural languages are very rich and allow to express subjectivity in different ways.

In this paper, our particular interest lies on how users refer to the aspects of the products that they evaluate in their reviews, as the area has struggled with the way users employ several different terms to refer to the same aspects. For example, in the review passage "she considered the camera price very expensive", the consumer employed the term "price" to evaluate an aspect of the camera; however, consumers might also use the terms "cost", "value", "investment", etc. In addition, consumers may use implicit or explicit aspects to refer to the same aspect, e.g., the sentences "she got calls at the São Francisco river" and "working anywhere" have been employed in actual reviews to evaluate the (implicit) "signal" aspect of a smartphone. It is also interesting to notice that, in some domains, proper names may be employed to refer to the aspects. For instance, the proper names "Sony" and "Nikon" may be used to evaluate the "product brand" aspect of digital cameras.

As there are some previous work on aspect identification for the Portuguese language (see, e.g., [3]), we focus our efforts on the next step of aspect clustering, which aims at automatically grouping aspect terms that refer to the same thing. Such process is a core step for several sentiment analysis tasks, as aspect-based polarity classification and opinion summarization.

In this paper, we investigate six aspect clustering methods for both explicit and implicit aspects in product reviews. We test four linguistic-based methods inspired in the literature, a statistical method (based in word embeddings), and a new (linguistic) method that we propose, which was motivated by an empirical study of the relevant linguistic phenomena in Portuguese reviews. We compared the six methods on three different domains - smartphones, digital cameras, and books - and demonstrate that our method significantly outperforms the others.

The rest of this paper is organized as follows. Section 2 introduces the essential related work. Section 3 presents the clustering methods, while Sect. 4 describes their evaluation. Some final remarks are made in Sect. 5.

2 Related Work

According to [25], two kinds of similarity measures are usually applied in the aspect clustering task: (i) those relying on knowledge resources (e.g., thesauri and semantic networks) [2,10,23], and (ii) those relying on distributional properties of the words in corpora [4,19,26].

In the knowledge source approach, ontologies are frequently explored, using WordNet lexical relations [15] or the categories of Wikipedia[1], as in [8,18]. In [18], for instance, the hypernymy/is-a relation is the basis for clustering aspects. However, such approach did not perform very well. The authors reported that specific domain aspects were not found on ontologies and lexicons, e.g., in smartphone domain, relevant aspects as "gps", "3G", "wap" and "hit" are too specific to happen in the above ontologies.

[1] http://wiki.dbpedia.org/.

The distributional similarity methods generally employ measures such as Cosine, Jaccard, Dice and PMI (Pointwise Mutual Information) [12], as in [1,4,26,28]. In this line, [28] also makes use of the widely adopted word embeddings produced by *word2vec* algorithm [14] to find aspect categories (which refer to the type of entity being evaluated, and not the groups that we look for). Some good results were produced, but difficulties to group domain specific aspects were also reported.

Overall, we could not find proposals for clustering implicit aspects. Only explicit aspects are tackled.

In this paper, we have developed and tested aspect clustering methods that were inspired in such previous attempts. We also try to overcome some of their limitations, as we describe in what follows.

3 Clustering Methods

We compared six aspect clustering methods: 4 linguistic methods inspired in the literature, 1 statistical method and 1 new method that we propose. All the methods receive as input a list of implicit and explicit aspects in their user reviews and produce as output clusters of similar aspects, i.e., aspects that refer to the same property or feature of an object. As we commented before, we consider that the task of aspect identification in the reviews was already performed (e.g., by one of the methods of [3]). We only focus on the aspect clustering task.

The implicit aspects are represented by their indicative terms in the reviews, for instance, the "working anywhere" n-gram to refer to the signal aspect of smartphones.

The general overview of the 4 linguistic methods inspired in the literature is shown in Fig. 1. The methods were incrementally implemented in order to evaluate the results obtained at each level of increment. For example, the first implemented method creates clusters of aspects using only synonymy relations (i.e., aspects that happen to be in a synonymy relation are clustered together). The second method creates clusters using synonymy and is-a relations. The third method uses synonymy, is-a, and part-of relations. The fourth method uses synonymy, is-a, part-of, and coreference relations. As we work for the Portuguese language, we have used the Onto-PT lexical ontology [16] for extracting synonymy, is-a and part-of relations. To find coreference relations, we have employed the CORP coreference resolution system for Portuguese [6,7]. These resources are widely used for this language.

For our statistical method, we adopted the ready-to-use trained word embedding models proposed by [9] and available in the NILC word embeddings repository[2]. These models have been widely used for Portuguese. We used the word2vec version [14] with 300 dimensions. The idea is that aspects with "similar" corresponding vectors should be clustered together.

[2] http://www.nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc.

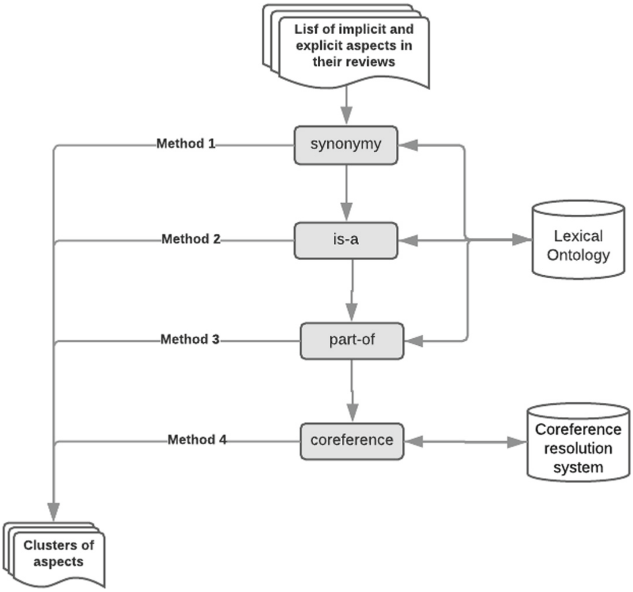


Fig. 1. Aspect clustering methods

The last method is the one that we propose, which we refer by OpCluster-PT. The OpCluster-PT algorithm was motivated by a linguistic empirical study, in which we studied the aspect-related linguistic phenomena that happen in product reviews. This study showed that, besides the traditional lexical relations (such as synonymy, hypernymy/is-a and meronymy/part-of), it is necessary to use causative, deverbal, diminutive (or augmentative), foreignism, and substring relations to find aspect clusters. To find these new relations, we have used Onto-PT [16], CORP [6,7] and the Portuguese foreignism and deverbal iLteC lexicons [5,11]. A list of diminutive/augmentative words was compiled for this proposal, as we could not find an available one. It is important to say that, as we use CORP, we also take advantage of this tool to (indirectly) find the hypernymy relations in our new method, as CORP proved do be better than Onto-PT in our domains for identifying aspects grouped by such relations.

The relevance of the new relations may be easily exemplified. Causative relations help finding that “to finish” may indicate the implicit aspect “end of story” (of a book); deverbal relations may indicate that “to write” (in a specific manner) refers to the implicit aspect “writing style” (in book domain); the diminutive “little book” (*livrinho*, in Portuguese) refers to the aspect “book”; the aspect “display” is a foreignism in Portuguese language that is similar to the aspect “screen” (*tela*, for some electronic product); and substring relation allows to detect that “image” and “image quality” may refer to the same aspect “image” (for some electronic product again).

The proposed method works as the previous linguistic methods. It incrementally forms groups with the aspects that show some of the predicted relations. The full method is shown in Algorithm 1. It receives as input a set of reviews R and the list of implicit and explicit aspects A that have occurred in the reviews. It then performs three main steps. In the first one (starting in the first *repeat*), for each aspect a, it looks for other aspects that show some relation with the previous one. In the second step (starting in the second *repeat* inside the first one), the method looks for other aspects that show some specific relations[3] with the ones that were clustered together with a. This second step is, in fact, a recursive step, which looks for any other related aspect that was left outside the cluster. In the last step (in the third *repeat*), we look for any remaining unitary clusters that might be joined with other clusters. This happens if the unitary aspect is in a substring relation with some aspect of other cluster. As output, the algorithm provides a list of clusters G.

In what follows, we report the evaluation of the methods.

4 Evaluation

To evaluate the aspect clustering methods, we have manually annotated a corpus of smartphone, digital camera, and book reviews. We selected three commonly used domains to test how robust and generalizable the methods are. Each domain counted with 60 reviews.

In each review, we marked and clustered the explicit and implicit aspects. The implicit aspects were indicated by the clue terms that signaled them. Such data consisted in the reference annotation to which the automatic output of the clustering methods was compared to.

Table 1 shows the relevant numbers of the reference annotation. One may see that there is a significant difference among the domains. In the smartphone and digital camera domains, we have identified more domain specific aspects, maybe due to the popularity of such devices, which allows users to comment about their technical details. Books, otherwise, are usually not evaluated on their technical details (as the type of paper and weight), but on more prototypical aspects in this domain (as characters and story).

Table 1. Reference annotation

Domain	Book	Camera	Smartphone
Total number of aspects	103	132	180
Number of explicit aspects	91	109	142
Number of implicit aspects	12	23	38
Number of clusters	21	36	48

[3] We only look for coreference, foreignism and diminutive-augmentative relations, because we empirically observed that they were the most accurate ones in this step.

Algorithm 1. OpCluster-PT

Input: List of aspects $A = \{a_1, a_2, ..., a_n\}$ sorted by frequency (in decreasing order) and their corresponding reviews $R = \{r_1, r_2, ..., r_m\}$ preprocessed by CORP
Output: Clusters of aspects G = $\{g_1, g_2, ..., g_p\}$, where each g_i contains a subset of aspects of A

Let **B** = $\{b_{syn}, b_{part}, b_{caus}, b_{devb}, b_{fore}, b_{dim-augm}, b_{coref}, b_{subs}\}$, where each $b_{relation}$ in **B** contains the result of searching for aspects in synonymy, part-of, causative, deverbal, foreignism, diminutive-augmentative, coreference, and substring relations (for example, b_{syn} contains the synonymous aspects)
Let **j** and **k** be integers initiliazed with zero

```
repeat
    Consider $a_i$ as the most frequent aspect in A
    if $a_i$ in A contains synonymous words in Onto.PT then
        Add such words to $\mathbf{b_{syn}}$
    end if
    if $a_i$ in A contains part-of related words in Onto.PT then
        Add such words to $\mathbf{b_{part}}$
    end if
    if $a_i$ in A contains causative related words in Onto.PT then
        Add such words to $\mathbf{b_{caus}}$
    end if
    if $a_i$ in A contains deverbal construction related words in iLteC lexicon then
        Add such words to $\mathbf{b_{devb}}$
    end if
    if $a_i$ of A contains foreignism related words in iLteC lexicon then
        Add such words to $\mathbf{b_{fore}}$
    end if
    if $a_i$ in A contains diminutive or augmentative related words in our compiled list then
        Add such words to $\mathbf{b_{dim-augm}}$
    end if
    if $a_i$ in A has related coreference terms in the corresponding reviews, as indicated by CORP
    then
        Add such terms to $\mathbf{b_{coref}}$
    end if
    if $a_i$ in A contains substring relations with other aspects in A then
        Add such aspects to $\mathbf{b_{subs}}$
    end if
    Remove duplicate items from $\mathbf{B} = \{b_{syn}, b_{part}, b_{caus}, b_{devb}, b_{fore}, b_{dim-augm}, b_{coref}, b_{subs}\}$
    Increment j
    Create cluster $\mathbf{g_j}$ and add to it the aspects of intersection(A,B)
    Remove from A the aspects of intersection(A,B)
    Empty B
    repeat
        Consider $a_k$ as each aspect in $\mathbf{g_j}$, ignoring $a_i$, which was already processed
        if $\mathbf{a_k}$ in $\mathbf{g_j}$ has related coreference terms in the corresponding reviews, as indicated by CORP
        then
            Add such terms to $\mathbf{b_{coref}}$
        end if
        if $\mathbf{a_k}$ in $\mathbf{g_j}$ contains foreignism related words in iLteC lexicon then
            Add such words to $\mathbf{b_{fore}}$
        end if
        if $\mathbf{a_k}$ in $\mathbf{g_j}$ contains diminutive or augmentative related words in our compiled list then
            Add such words to $\mathbf{b_{dim-augm}}$
        end if
        Remove duplicate items from $\mathbf{B} = \{b_{syn}, b_{part}, b_{caus}, b_{devb}, b_{fore}, b_{dim-augm}, b_{coref}, b_{subs}\}$;
        Add to $\mathbf{g_j}$ the aspects of intersection(A,B)
        Remove from A the aspects of intersection(A,B)
        Empty B
    until every $\mathbf{a_k}$ in $\mathbf{g_j}$ is tested
until A is empty
repeat
    Consider each unitary cluster $\mathbf{g_j}$ in G
    if the aspect in this unitary cluster $\mathbf{g_j}$ has a substring relation with some aspect in other cluster
    $\mathbf{g_k}$ in G then
        Add this aspect in $\mathbf{g_j}$ to the cluster $\mathbf{g_k}$
        Eliminate $\mathbf{g_j}$
    end if
until every unitary cluster $\mathbf{g_j}$ in G is tested
```

For evaluating the methods, we have computed the traditional clustering evaluation measures of Precision, Recall, F-measure and Global F-measure (as defined in [20]) over the reference clusters. Precision indicates the proportion of aspects of each automatic cluster that is correctly clustered (according to the reference clusters). Recall indicates the proportion of aspects of the reference clusters that was covered by the automatically generated clusters. As these measures are complementary, we also compute the F-measure score, which represents the harmonic average between precision and recall. The global F-measure of each automatically generated cluster, relative to the entire set of clusters, is based on the cluster that best describes each reference cluster. The achieved results are shown in details in Tables 2, 3, 4 and 5.

One may see that the synonymy-based method (the simplest one, under the linguistic point of view) presented the best Precision results (in Table 2) for the task. However, we observed a very high number of unitary clusters (in relation to the reference annotation), which increases Precision (as the Precision of each unitary cluster is equal to 100%), but seriously harms Recall, which is confirmed in Table 3. The method that we propose here produced significantly higher recall numbers for all the domains. Overall, looking at F-measure values, the method we propose was the best one, outperforming all the others.

It is also interesting to notice that our method achieved the best results in book domain, probably because there are more prototypical aspects and, therefore, less domain specific aspects, which are more difficult to find in the linguistic repositories that we adopt.

Surprisingly, the word embeddings performed very poorly. We believe that this happened because we have used a widely used model trained on general corpora, and not corpora of reviews. However, one might argue that most of the aspects are general enough to be used in general language corpora too. This remains as an open question to investigate in the future.

As illustration, our method automatically generated a cluster composed by the aspects "cost benefit", "price", "value", "investment" and "cheap" (which is an implicit aspect), which is very good. An example of problematic cluster is the one composed by "enterprise", "lg", "nokia", "sony", "sony_ericson", "program", "design", "system" and "model", in which the 4 last words are clearly misplaced in this cluster.

Table 2. Precision results

	Methods	Book	Camera	Smartphone
1	Synonymy	**0.974**	**0.987**	**0.973**
2	Synonymy + is-a	0.916	0.967	0.940
3	Synonymy + is-a + part-of	0.916	0.967	0.943
4	Synonymy + is-a + part-of + coreference	0.945	0.963	0.953
5	Word embeddings	0.953	0.962	0.956
6	OpCluster-PT	0.925	0.933	0.947

Table 3. Recall results

	Methods	Book	Camera	Smartphone
1	Synonymy	0.231	0.281	0.296
2	Synonymy + is-a	0.242	0.287	0.314
3	Synonymy + is-a + part-of	0.242	0.287	0.310
4	Synonymy + is-a + part-of + coreference	0.321	0.307	0.364
5	Word embeddings	0.231	0.292	0.300
6	OpCluster-PT	**0.748**	**0.687**	**0.550**

Table 4. F-measure results

	Methods	Book	Camera	Smartphone
1	Synonymy	0.374	0.438	0.454
2	Synonymy + is-a	0.383	0.442	0.471
3	Synonymy + is-a + part-of	0.383	0.442	0.466
4	Synonymy + is-a + part-of + coreference	0.480	0.466	0.527
5	Word embeddings	0.372	0.448	0.457
6	OpCluster-PT	**0.827**	**0.792**	**0.702**

Table 5. Global F-measure results

	Methods	Book	Camera	Smartphone
1	Synonymy	0.300	0.351	0.347
2	Synonymy + is-a	0.249	0.319	0.333
3	Synonymy + is-a + part-of	0.244	0.319	0.333
4	Synonymy + is-a + part-of + coreference	0.399	0.409	0.508
5	Word embeddings	0.280	0.336	0.350
6	OpCluster-PT	**0.711**	**0.605**	**0.583**

We have checked that, for correcting the remaining errors and improving the results, we might also incorporate knowledge about proper names and slangs, using, e.g., Wikipedia data and specialized lexicons. This remains for future work.

5 Final Remarks

According to [17], the aspect-based sentiment analysis task requires deep understanding of natural language characteristics and textual context. Therefore, in this paper, we present the OpCluster-PT algorithm, designed to cluster explicit and implicit aspects in product reviews. We achieved the best results when comparing to other four linguistic-based methods and one statistical method.

As a side effect of this work, a reference dataset was produced, with indicated explicit and implicit aspects, as well as manually produced aspect clusters. Additionally, we have also produced aspect ontologies for the investigated domains.

More information about this work and the related tools and resources may be found at the OPINANDO project website[4].

Acknowledgments. The authors are grateful to FAPESP, CAPES and CNPq for supporting this work.

References

1. Abu-Jbara, A., King, B., Diab, M.T., Radev, D.R.: Identifying opinion subgroups in arabic online discussions. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, pp. 829–835 (2013)
2. Alvarez, M., Lim, S.: A graph modeling of semantic similarity between words. In: Proceedings of the Conference on Semantic Computing, Irvine, United States, pp. 355–362 (2007)
3. Balage Filho, P.P.: Aspect extraction in sentiment analysis for Portuguese. Ph.D. thesis, University of São Paulo, São Carlos, Brazil (2017)
4. Chen, Y., Zhao, Y., Qin, B., Liu, T.: Product aspect clustering by incorporating background knowledge for opinion mining. PLOS One **11**(8), 1–16 (2016)
5. Ferreira, J.P., Janssen, M.: Dicionário de Formas Não Adaptadas, 1ª edn. Instituto de Linguística Teórica e Computacional (2017)
6. Fonseca, E., Sesti, V., Antonitsch, A., Vanin, A., Vieira, R.: CORP: Uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências. Linguamática **9**(1), 3–18 (2017). https://doi.org/10.21814/lm.9.1.241. http://linguamatica.com/index.php/linguamatica/article/view/v9n1p1
7. Fonseca, E.B., Vieira, R., Vanin, A.A.: CORP: coreference resolution for portuguese. In: Proceedings of the 12th International Conference on the Computational Processing of Portuguese, Tomar, Portugal, pp. 9–11 (2016)
8. García, A., Cuadros, M., Rigau, G., Gaines, S.: V3: unsupervised generation of domain aspect terms for aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland, pp. 833–837 (2014)
9. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: evaluating on word analogies and natural language tasks. In: Proceedings of the Symposium in Information and Human Language Technology, Uberlandia, Brazil, pp. 122–131 (2017)
10. Hughes, T., Ramage, D.: Lexical semantic relatedness with random graph walks. Comput. Linguist. **7**(1), 581–589 (2007)
11. Janssen, M., Ferreira, J.P.: Dicionário de nomes deverbais, 1ª edn. Intituto de Linguística Teórica e Computacional (2007)
12. Lee, L.: Measures of distributional similarity. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics ACL 1999, pp. 25–32. Association for Computational Linguistics, Stroudsburg (1999). https://doi.org/10.3115/1034678.1034693

[4] https://sites.google.com/icmc.usp.br/opinando/.

13. Liu, B.: Sentiment Analysis and Opinion Mining, 1st edn. Morgan & Claypool Publishers, San Rafael (2012)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Computing Research Repository **1301.3781**(1) (2013)
15. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: WordNet: an on-line lexical database. Int. J. Lexicograph. **3**, 235–244 (1990)
16. Oliveira, H.G.: Beyond the automatic construction of a lexical ontology for Portuguese: resources developed in the scope of Onto.PT. In: Proceedings of the Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish, São Carlos, Brazil, pp. 64–68 (2014)
17. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Stroudsburg, United States, pp. 79–86 (2002)
18. Patra, B.G., Mandal, S., Das, D., Bandyopadhyay, S.: Ju_cse: a conditional random field (CRF) based approach to aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland, pp. 370–374 (2014)
19. Pereira, F., Tishby, N., Lee, L.: Distributional clustering of English words. In: Proceedings of the 31st Annual Meeting on Association for Computational Linguistics, Stroudsburg, United States, pp. 183–190 (1993)
20. Seno, E.R.M.: Um mátodo para fusão automática de sentenas similares em português. Ph.D. thesis, University of São Paulo, São Carlos, Brazil (2010)
21. Taboada, M.: Sentiment analysis: an overview from linguistics. Ann. Rev. Linguist. **2**(1), 325–347 (2016). http://www.annualreviews.org/doi/full/10.1146/annurev-linguistics-011415-040518
22. Wu, C.W., Liu, C.L.: Ontology-based text summarization for business news articles. In: Proceedings of the 3th International Symposium on Computer Architecture, Honolulu, United States, pp. 389–392 (2003)
23. Yang, H., Callan, J.: A metric-based framework for automatic taxonomy induction. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Suntec, Singapore, pp. 271–279 (2009)
24. Yu, J., Zha, Z., Wang, M., Wang, K., Chua, T.: Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, United Kingdom, pp. 140–150 (2011)
25. Zhai, Z., Liu, B., Xu, H., Jia, P.: Clustering product features for opinion mining. In: Proceedings of the 4th International Conference on Web Search and Data Mining, New York, United States, pp. 347–354 (2011)
26. Zhang, S., Jia, W., Xia, Y., Meng, Y., Yu, H.: Product features extraction and categorization in Chinese reviews. In: Proceedings of the 6th International Multi-Conference on Computing in the Global Information Technology, Nice, France, pp. 38–42 (2011)
27. Zhao, L., Li, C.: Ontology based opinion mining for movie reviews. In: Karagiannis, D., Jin, Z. (eds.) KSEM 2009. LNCS (LNAI), vol. 5914, pp. 204–214. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10488-6_22
28. Zhou, X., Wan, X., Xiao, J.: Representation learning for aspect category detection in online reviews. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Texas, United States, pp. 417–423 (2015)

Finding Opinion Targets in News Comments and Book Reviews

Leonardo Gabiato Catharin(✉) and Valéria Delisandra Feltrim(✉)

State University of Maringá, Maringá, PR, Brazil
leo.catharin@gmail.com, vfeltrim@din.uem.br

Abstract. In this paper we approach the extraction of opinion targets in user-generated texts written in Portuguese in two different domains: political news articles and books. The former was represented by SentiCorpus-PT, and the latter by ReLi. We implemented and tested five prototypes for extracting explicit opinion targets from the corpora: one based on Centering, one based on morphosyntactic patterns, and three based on syntactic and morphosyntactic heuristics, which were used as baselines. Experimental results proved target extraction on ReLi to be a more difficult task than on SentiCorpus-PT, probably due to its low proportion of targets per sentence. Also, the baseline based on proper nouns performed better on SentiCorpus-PT, showing that superficial heuristics may yield good results in this domain.

Keywords: Opinion targets · Centering · Morphosyntactic patterns

1 Introduction

The most common type of sentiment analysis is the classification of polarity, in which the analyzed text is classified as positive or negative. However, to obtain a more refined analysis, it is necessary to extract other aspects of the opinion in addition to its polarity, such as the entities on which the opinion is expressed [7]. These entities (or entities' aspects) are usually called opinion targets.

Opinion target extraction has been generally performed as a part of the opinion mining process for products' reviews [2,15]. It is, however, less common in other domains. In this paper we address the extraction of targets in opinionated texts written in Portuguese in two different domains: political news articles and books. For the former we have used SentiCorpus-PT [1], a corpus of comments about political news articles. For the latter we used the corpus ReLi [3], which consists of book reviews from an on-line social network of readers. SentiCorpus-PT and ReLi are both user-generated content corpora which have been manually annotated regarding opinions' polarities and targets. They are, however, very different in relation to their annotated targets. In SentiCorpus-PT, targets are human entities, namely politicians. Also, most of its sentences have at least one annotated target. In ReLi, targets are either a book or one of its parts, and most of its sentences do not have annotated targets.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 375–384, 2018.
https://doi.org/10.1007/978-3-319-99722-3_38

Considering the corpora characteristics, we implemented and evaluated five prototypes for extracting explicit opinion targets: one based on Centering Theory, which was adapted from Ma and Wan [9]; one based on morphosyntactic patterns, adapted from Turney [16]; and three heuristics that we call baselines: one based on syntactic information and two based on the occurrence of proper nouns. All five prototypes were applied to SentiCorpus-PT and three of them were applied to ReLi.

Experimental results proved the target extraction on ReLi to be a more difficult task than on SentiCorpus-PT, probably due to its low proportion of targets per sentence. Also, the baselines based on proper nouns performed better than the other approaches on SentiCorpus-PT, showing that superficial heuristics may yield good results in some domains.

The remaining of this paper is organized as follows. Section 2 presents previous works related to the approaches implemented in our prototypes. Section 3 describes the used corpora and details their processing. The prototypes are described in Sect. 4. Experimental results are presented in Sect. 5. Finally, Sect. 6 presents the conclusions of this study.

2 Related Work

2.1 Centering Theory

Proposed by Grosz et al. [4], the Centering Theory was developed to evaluate coherence in a discourse by analyzing transitions among centers of attention (entities) in each utterance. According to the theory, each utterance U_i has an ordered set of associated centers called Forward-Looking Centers - $\mathrm{Cf}(\mathrm{U}_i)$. This set contains all the potential centers of attention of the current utterance, as well as the potential centers of the next utterances, assuming that the text is coherent. The centers in $\mathrm{Cf}(\mathrm{U}_i)$ are ranked according to their salience, which is usually measured based on the center's syntactic function and following the scale "subject > object > other". The best ranked center of Cf is its most salient and is called Preferred Center - $\mathrm{Cp}(\mathrm{U}_i)$, or next preferred center. The highest ranked element of $\mathrm{Cf}(\mathrm{U}_{i-1})$ realized in U_i constitute U_i's Backward-Looking Center - $\mathrm{Cb}(\mathrm{U}_i)$. Thus, in a coherent segment, $\mathrm{Cp}(\mathrm{U}_i)$ is likely to be the $\mathrm{Cb}(\mathrm{U}_{i+1})$.

Centering Theory was first used for extracting opinion targets by Ma and Wan [9]. From the analysis of a corpus of news articles comments written in Chinese, the authors concluded that information about attention centers could be useful for identifying targets. Since a center is the focus of attention of an utterance, it would be likely to be the target. The proposed approach aims at extracting both explicit targets (target explicitly mentioned) and implicit targets (targets not explicitly mentioned, but that can be inferred from the context). Explicit targets are identified by an algorithm that uses the sets Cf and Cb. The identification of implicit targets uses, in addition to the Centering sets, information from the news article from which the comments were collected. The authors evaluated their proposal on a corpus of user comments composed of 1,597 sentences extracted from nine news articles related to economics, sports, and technology. For each sentence a single target was extracted and the overall accuracy (for both explicit and implicit targets) was 43.2%.

Oliveira and Feltrim [11] adapted Ma and Wan's approach to identify explicit targets from Portuguese news comments extracted from SentiCorpus-PT. The authors reported 60.4% accuracy on a random subset of the SentiCorpus-PT composed of 100 sentences, for which coreference has been manually resolved.

2.2 Pattern Matching

In the context of textual analysis, patterns are units of information that occur repeatedly in the text. Information from different levels of language processing can be used to find and represent such patterns, and approaches based on pattern matching have been applied in a variety of information extraction tasks.

For the extraction of opinion targets, it is common to base the patterns on morphosyntactic information provided by a part-of-speech (POS) tagger. For example, Turney [16] used patterns based on POS to extract opinion phrases in Epinions reviews. In a similar manner, Htay and Lynn [5] proposed a set of POS patterns to identify product aspects and opinion words from customers reviews extracted from Epinion and Amazon. Maharani et al. [10] combined Turney's [16] and Htay and Lynn's [5] patterns, and proposed new ones, to identify product aspects in reviews extracted from the corpus of Hu and Liu [6]. In a more general scenario, Rocha et al. [13] used POS patterns to extract named entities from a book written in Portuguese. Besides POS information, Qiu et al. [12] and Liu et al. [8] used syntactic patterns that characterize relations between opinionated words and opinion targets.

3 Corpora

As mentioned in Sect. 1, we used two corpora as basis for the development and evaluation of our target extraction prototypes: SentiCorpus-PT [1] and ReLi [3]. Both corpora are composed of user-generated content written in Portuguese and have been manually annotated regarding opinions' polarities and targets.

SentiCorpus-PT is composed of comments about a series of news articles covering TV debates on the 2009 election of the Portuguese Parliament [1]. Each sentence in the corpus may have different opinion targets, and targets are human entities, namely politicians, political organizations (generally used for referring its members), media personalities, or users. The version of SentiCorpus-PT used in this study[1] comprises 1,082 comments, totalizing 3,888 annotated sentences. 94.3% of the sentences has at least one annotated target, and most of them (79%) has exactly one target. Also, targets may be referred by expressions of up to 8 words, but most of them are referred by one word (75.7%).

ReLi consists of book reviews extracted from an on-line social network of readers [3]. As in SentiCorpus-PT, each sentence may have different opinion targets, but in ReLi targets are either a book or one of its parts, such as chapters,

[1] http://dmir.inesc-id.pt/project/SentiCorpus-PT_01_in_English.

characters, etc. The available version of ReLi[2] contains 1,600 reviews from 13 different books, totalizing 12,514 sentences. For each sentence, words are annotated with its POS tag, whether it is a target, and polarity information. Only 17.8% of the sentences in the corpus have annotated targets, which is a small proportion compared to SentiCorpus-PT. Of this percentage, 81.4% have exactly one annotated target. In ReLi targets may be referred by expressions of up to 20 words, but most of them are referred by one word (84.5%). Table 1 summarizes the observed numbers of targets per sentence and words per target in both corpora.

Table 1. Targets per sentence and words per target in SentiCorpus-PT and ReLi

SentiCorpus-PT				ReLi			
Sentences with	#	Targets of	#	Sentences with	#	Targets of	#
No targets	221	1 word	3331	No targets	10281	1 word	2337
1 target	3071	2 words	597	1 target	1818	2 words	102
2 targets	494	3 words	370	2 targets	320	3 words	108
3 targets	77	4 words	69	3 targets	72	4 words	84
4 targets	18	5 words	20	4 targets	20	5 words	47
5–6 targets	7	6–8 words	13	5 targets	3	6–20 words	89

4 Prototypes

4.1 Based on Centering

We implemented an adaptation of Ma and Wan's approach [9] that considers only explicit targets, since the identification of implicit targets requires the acquisition of information that is not contained in the comments/reviews, according to the authors' original proposal. Figure 1 presents the general pipeline implemented by this prototype.

To build the Cf set for a sentence, it is necessary to filter the noun phrases (NP) and to order them according to their salience, which in this case was inferred by means of the NP's syntactic function. We used the parser available as part of the COGrOO API[3] for identifying both NPs and their syntactic function. After the NPs filtering, they are ranked according to the scale "Subject (SUBJ) > Object (ACC) > other". Ranked Cfs are estimated for all sentences in a comment/review and used as input to the Centering algorithm.

Our Centering algorithm is a fragment of the original algorithm of Ma and Wan [9] that accounts for explicit targets. All NPs (centers c_i) in Cf(s_i) are considered as candidate targets for sentence s_i. If the sentence is the first in the

[2] https://www.linguateca.pt/Repositorio/ReLi/.

[3] http://ccsl.ime.usp.br/redmine/projects/cogroo/wiki/API%20CoGrOO%204x.

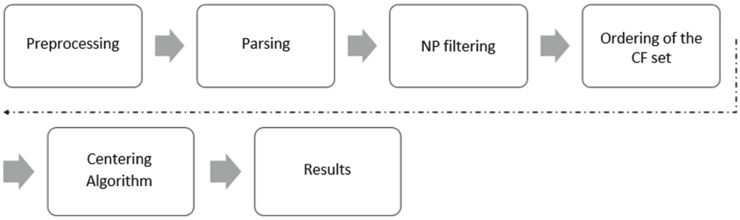


Fig. 1. Pipeline implemented by the centering based prototype

text (s_1), Cb(s_1) set will be empty, so the best ranked candidate of Cf(s_1) is chosen as the target for the sentence. Otherwise, Cb(s_i) will be composed by candidates of Cf($s_i - 1$) that are referred in s_i. In this case, the best ranked candidate in Cb(s_i) (probably Cp($s_i - 1$)) is chosen as the target for sentence s_i. Since the algorithm search for the most salient center in Cf/Cb, it identify only one target per sentence.

4.2 Based on Pattern Matching

As in Turney [16], we used POS information to represent a pattern, and patterns were extracted from a training corpus. Since ReLi has few targets in relation to its number of sentences, it was difficult to find patterns that occur in a relevant frequency. For this reason, we decided to use this approach only on SentiCorpus-PT. It worth noticing that while ReLi's target/sentence ratio is 0.22, it is 1.8 for SentiCorpus-PT.

Figure 2 presents the two-step process implemented in this prototype. For Step 1 - pattern extraction, half of SentiCorpus-PT's sentences was randomly selected and used for training. The remaining of the corpus was used for testing in Step 2 - target extraction. For both steps, we used the CoGrOO API for POS tagging. In Step 1, all the annotated targets with their respective POS tags were extracted from the training sentences, resulting in 53 patterns that range from one word (17 patterns) to five words (2 patterns) length. In Step 2, these patterns were joined in sets, and used to identify targets on the test sentences.

4.3 Based on Heuristics

Based on the notion that targets would be salient entities in the sentences, we expected many targets to have the function of subject. In addition, especially in SentiCorpus-PT, we expected many targets to be proper nouns. Based on that intuition, we implemented three heuristics as baselines. As in the previous prototypes, we used the CoGroo API for parsing and tagging.

Baseline 1 extracts the subject of the sentence as the target. If the parser identifies more than one subject in a sentence, they are ranked by POS as follows: proper noun > noun > others. If more than one subject-proper noun is found,

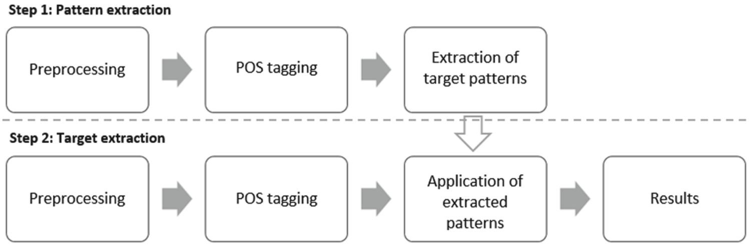

Step 1: Pattern extraction
Preprocessing
POS tagging
Extraction of target patterns
Step 2: Target extraction
Preprocessing
POS tagging
Application of extracted patterns
Results

Fig. 2. Pipelines implemented by the prototype based on pattern matching

then the one that occurs first in the sentence is selected. Thus, this baseline extracts at most one target per sentence.

Baseline 2 extracts the first proper noun found in the sentence without considering its syntactic function. As in baseline 1, baseline 2 also extracts at most one target per sentence and, for this reason, both baselines have a low recall.

Baseline 3 extracts all proper nouns in the sentence, regardless of their syntactic function. Differently from baselines 1 and 2, it may extract more than one target per sentence. This baseline increases recall at the cost of lowering precision.

5 Experiments and Results

In all evaluation experiments, we considered that a target was correctly extracted when the output of the prototype was equal to or contained within a reference target for the processed sentence. Experiments were conducted by approach and results were measured in terms of precision, recall, and f-measure.

5.1 Based on Centering

The Centering based prototype (CT) selects a target by analyzing the sets Cf and Cb, and the accuracy in the estimation of Cb depends on information about coreferent entities. However, our current prototype does not include coreference resolution. To evaluate how this limitation would impact the results, we manually resolved coreference for SentiCorpus-PT. Due to the high cost of this task and the size of corpus ReLi, we did not perform it for this corpus. Tables 2 and 3 present, respectively, the results obtained for SentiCorpus-PT and ReLi by the CT prototype, as well by baselines 1 and 2. As baseline 3 can extract more than one target per sentence, thus differentiating itself from the other baselines and the CT prototype, it was not included in the aforementioned tables.

As shown in Table 2, baseline 2, which extracts the first proper noun of the sentence as target, achieved the best results for SentiCorpus-PT. Since the targets in this corpus are human entities, we expected this baseline to perform well. Also, as expected, CT's results improved when applied to the corpus with coreference resolution; nevertheless, it remained below baseline 2. Despite reaching

Table 2. Centering based prototype results for SentiCorpus-PT

Prototype	Precision	Recall	F-measure
Baseline 1	0.61	0.31	0.41
Baseline 2	0.79	0.38	0.51
CT without correference	0.47	0.29	0.36
CT with correference	0.58	0.36	0.44

Table 3. Centering based prototype results for ReLi

Prototype	Precision	Recall	F-measure
Baseline 1	0.11	0.22	0.15
Baseline 2	0.06	0.0	0.06
CT without correference	0.10	0.27	0.15

reasonable precisions, all prototypes had a low recall. One contributor factor to this is that these prototypes identifies at most one target per sentence.

Results for ReLi (Table 3) were very discrepant from SentiCorpus-PT. As expected, baseline 2 performed poorly in this corpus, due to the low occurrence of proper nouns as targets. Baseline 1 and CT performed equally in terms of f-measure, but recall was slightly higher for the latter. Based on these results, we believe that CT with coreference resolution would overcome the baselines for this corpus. Even so, results would be worse than for SentiCorpus, since CT always identifies one target per sentence, and most sentences in ReLi do not have targets.

5.2 Based on Pattern Matching

As explained in Sect. 4.2, the pattern matching prototype was built and tested for SentiCorpus-PT only. The pattern matching approach can extract many targets per sentence, so this prototype was compared to baseline 3, which behaves similarly. In addition, since baseline 3 extracts all proper nouns of the sentence as targets, it was expected to improve recall of baseline 2. We experimented with several sets of patterns. The configuration of the sets was done empirically and guided by frequency and accuracy of patterns, as well as known features of the corpus, such as the frequent use of proper nouns to mention targets.

The results achieved by the best sets are presented in Table 4. Among the experiments that selected patterns by frequency, the best result was achieved using the 5 most frequent patterns (Set 1). Sets based on this selection strategy shows high recall and low precision, since it retrieves too many expressions. It was the other way around (high precision and low recall) for experiments in which the patterns were selected by precision. The best results were achieved using the 11 most precise patterns, and, to improve recall, we added the pattern "prop" – proper noun to it (Set 2). Since none of this sets reached baseline 3 results, we experimented with sets composed of patterns that include proper nouns. Set 3 includes all patterns (16) that include "prop" and achieved the best results among the tested pattern sets. As an attempt to improve its recall, we added patterns based on personal pronouns - "pron-pers" - to the top 5 "prop" patterns of Set 3 (Set 4). The recall of set 4 was higher, but its accuracy was lower, lowering its f-measure as well.

Table 4. Pattern matching based prototype results for SentiCorpus-PT

Set of patterns	Precision	Recall	F-measure
Baseline 3	0.70	0.61	0.65
Set 1 (5-most frequent)	0.25	0.91	0.39
Set 2 (most precise + prop)	0.56	0.63	0.59
Set 3 (all prop)	0.63	0.62	0.63
Set 4 (5 prop + 3 pron-pers)	0.52	0.71	0.60

Table 5 presents the patterns that compose the sets mentioned in Table 4. The tagset used to describe the patterns is the same one used by the CoGroo API [14].

Table 5. The four best sets of patterns

Sets	Patterns
Set 1	{prop, n, pron-pers, prop prop, prop prop prop}
Set 2	{pron-det prp+pron-pers, prop prp+art prop, n prop prop, n prp n prp+art prop, prp+pron-pers, n prp+art prop, n prp prop prop, prop prp prop, pron-pers pron-det, prop prp+art n, prop prop, prop}
Set 3	{prop, prop prop, prop prop prop, prop prp prop, prop prop prop prop, n prp+art prop, n prop, prop adj, n prop prop, prop prp+art prop, prop prp+art n, n prp n prp+art prop, n prp+art prop prop, n prop prop prop, n prp prop prop, n prp prop}
Set 4	{pron-det prp+pron-pers, prop prp+art prop, n prop prop, n prp n prp+art prop, prp+pron-pers, n prp+art prop, n prp prop prop, prop prp prop, pron-pers pron-det, prop prp+art n, prop prop, prop, prop prop prop, prop prop prop prop, pron-pers pron-det, pron-pers}

6 Conclusion

We evaluate the performance of five prototypes for extracting explicit opinion targets in two different user-generated content corpora: SentiCorpus-PT and ReLi. The prototypes implemented three different approaches: one based on

Centering, one based on morphosyntactic patterns, and one based on syntactic and morphosyntactic heuristics.

The experimental results were better for SentiCorpus-PT in all tested approaches. Among the reasons for this, we can highlight the fact that SentiCorpus-PT has a higher target/sentence ratio than ReLi. For instance, since the CT prototype always extracts one target per sentence, the performance in ReLi was much lower than in SentiCorpus-PT. The low proportion of targets per sentence in Reli also impaired the search for relevant morphosyntactic patterns, and therefore this approach was not tested in this corpus.

Another aspect that influenced the results, especially for the extraction based on the heuristics, is the different forms in which mentions to targets occur in the corpora. Targets in SentiCorpus-PT are human entities, and an analysis of the corpus showed that 60.4% of the annotated targets could be recognized by a tagger/parser as a proper noun, since they contain names of persons or organizations, or acronyms. This characteristic improved the results of the heuristics and patterns based on proper nouns for this corpus. The Reli corpus has very different characteristics in relation to its targets and the experimental results reflected this. Further analysis of the ReLi corpus would be necessary to guide the implementation of new approaches and the proposal of better heuristics for this corpus.

Acknowledgements. We would like to thank Araucária, State of Paraná Research Foundation, for the financial support.

References

1. Carvalho, P., Teixeira, J., Sarmento, L., Silva, M.J.: Liars and saviors in a sentiment annotated corpus of comments to political debates. In: 49th Annual Meeting of The Association for Computational Linguistics, pp. 564–568. ACL, Portland (2011)
2. Filho, P.B., Pardo, T.: NILC_USP: aspect extraction using semantic labels. In: 8th International Workshop on Semantic Evaluation, pp. 533–568. BDPI, Dublin (2014)
3. Freitas, C., Motta, E., Milidiú, R.L., Cesar, J.: Sparkling vampire... lol! Annotating opinions in a book review corpus. In: Aluísio, S.M., Tagnin, S.E.O. (eds.) New Language Technologies and Linguistic Research: A Two-Way Road, pp. 128–146. Cambridge Scholars Publishing (2014)
4. Grosz, B.J., Winstein, S., Joshi, A.K.: Centering: a framework for modeling the local coherence of discourse. Comput. Linguist. **2**(21), 203–225 (1995)
5. Htay, S.S., Lynn, K.T.: Extracting product features and opinion words using pattern knowledge in customer reviews. Sci. World J. **2013**, 1–5 (2013)
6. Hu M., Liu B.: Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
7. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, San Rafael (2012)
8. Liu, K., Xu, L., Zhao, J.: Syntactic patterns versus word alignment: extracting opinion targets from online reviews. In: 51th Annual Meeting of The Association for Computational Linguistics, pp. 1754–1763. ACL, Sofia (2011)

9. Ma, T., Wan, X: Opinion target extraction in Chinese news comments. In: 23th International Conference on Computational Linguistics, pp. 782–790. ACL, Beijing (2010)
10. Maharani, W., Widyantoro, D.H., Khodra, M.L.: Aspect extraction in customer reviews using syntactic pattern. Proc. Comput. Sci. **2**(59), 244–253 (2015)
11. Oliveira, F.W.C., Feltrim, V.D.: Extração de Alvos em Comentários de Notícias em Português baseada na Teoria da Centralização. In: 10th Brazilian Symposium in Information and Human Language Technology, pp. 63–67. STIL, Natal (2015)
12. Qiu, G., Liu, B., Bu, J., Chen, C.: Opinion word expansion and target extraction through double propagation. Comput. Linguist. **37**(1), 9–27 (2011)
13. Rocha, C., Jorge A.M., Sionara, R.A., Brito, P., Pimenta, C., Rezende, S.O.: PAMPO: using pattern matching and pos-tagging for effective named entities recognition in Portuguese. arXiv preprint arXiv ARXIV:1612.09535 **2**(23) (2016)
14. Silva, W.D.C.M.: Aprimorando o corretor gramatical CoGrOO. Master's Dissertation, Instituto de Matemática e Estatística da Universidade de São Paulo (2013)
15. Siqueira, H.B.A.: PairClassif - Um Método para Classificação de Sentimentos Baseado em Pares. Master's Dissertation, Federal University of Pernambuco (2013)
16. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: 40th Annual Meeting of The Association for Computational Linguistics, pp. 417–424. ACL, Philadelphia (2002)

Semi-supervised Sentiment Annotation of Large Corpora

Henrico Bertini Brum(✉) and Maria das Graças Volpe Nunes

Núcleo Interinstitucional de Linguística Computacional,
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo,
São Paulo, Brazil
henrico.brum@usp.br, gracan@icmc.usp.br

Abstract. Huge annotated corpora are relevant for many Natural Language Processing tasks such as Sentiment Analysis. However, a manual and more precise annotation is always costly and becomes prohibitive when the corpus is too large. This paper presents a semi-supervised learning based framework for extending sentiment annotated corpora with unlabeled data, named CasSUL. The framework was used to extend in eight times TTsBR, a corpus of 15.000 tweets in Brazilian Portuguese manually annotated in three polarity classes. The extended annotated corpus was used to train several polarity classifiers and the results show that some combinations of classifier and features can preserve the annotation quality of the original corpus in the resulting corpus.

Keywords: Corpus annotation · Semi-supervised learning
Sentiment analysis

1 Introduction

Several tasks in Natural Language Processing (NLP) require annotated datasets, or corpora, for training and evaluating methods and comparing different systems. The process of manual annotation of corpora, although precise, is usually costly and becomes prohibitive when scaled to larger datasets. For popular tasks on NLP, such as Sentiment Analysis (SA), the study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities [10], we can find widely used corpora that serve as baselines for novel methods and approaches proposed for the task. Two examples are the Stanford Sentiment Treebank [21] and SemEval datasets [13] that are used for evaluating state-of-the-art models due to the reliability of the annotation and to the large number of labeled documents.

For Brazilian Portuguese we can find in the literature several corpora for the same task, but the high costs with manual annotation limit the resources to be either small or obtained trough entirely automatic methods such as user scores in websites or using semantic tips such as presence/absence of emoticons. Furthermore the data presented in these corpora may become outdated, incomplete or

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 385–395, 2018.
https://doi.org/10.1007/978-3-319-99722-3_39

do not be large enough for machine learning approaches, such as Deep Learning architectures, that rely on larger datasets for generalizing classification.

In this paper we introduce CasSUL, a semi-supervised framework annotating sentiment corpora using a small portion of annotated data and extending it with a large set of unlabeled documents, reducing human effort on annotation and providing a middle-ground between manual and automatic labeling of a large dataset.

2 Related Work

The creation of datasets for SA was largely addressed in the literature. Several authors have made efforts in using automatic labeling methods for creating SA corpora due to the challenges involved in the manual annotation – hiring and training annotators; measuring agreement between annotators; writing guidelines; developing interfaces; and more.

In some domains, such as product reviews, users give scores (sometimes grades or stars) for evaluating the target material. Previous works in SA used these scores for automatic identifying positive and negative reviews [16,22]. For Portuguese, some authors followed same approach for creating large datasets, such as Buscapé and Mercado Livre [2,9].

With the growth of social networks and blogs, the number of available data for NLP studies has grown as well, but despite the high volume of data these websites usually do not contain score based systems as the previous. One alternative is to use Distant Supervision, which is the selection of a consistent feature for identifying semantic orientation in texts. [8] adopted emoticons for this task on Twitter, using a selection of *positive emoticons* and *negative emoticons*. This approach performs well for two classes and has been used by other researches [5,15]. One of the weaknesses of this method is that it removes an important feature from the dataset, the emojis; and it is also limited to binary classification only, since defining any sentence without emojis neutral would add too much noise and it is hard to form a list of *neutral* emoticons.

On the other hand, semi-supervised approaches have been shown effective for improving classifiers using unlabeled data for sentiment analysis [6,19,20]. The results using self-training, an iterative method of adding classified documents to the training subset, have improved the classification. This has motivated the framework described below.

3 The CasSUL Framework

We propose a semi-supervised approach for labeling sentiment corpora, named CasSUL. This was implemented as a self-training iterative framework following Fig. 1. CasSUL input is a small dataset of manual labeled documents and a larger set of unlabeled documents. The main goal is to use machine learning classifiers for training models and iteratively increase the labeled documents using reliable classified data.

Initially we use a manually annotated dataset for training a sentiment classification model. This model is used for predicting labels for the unlabeled documents. We use the probability of the predicted labels for ranking the unlabeled documents by the most reliable. In each iteration we add a subset of these data (the most reliable) to the initial dataset, thus increasing its size and re-classifying the remaining unlabeled data each time. We manually set a percentage of data (the threshold shown in Fig. 1) defining how many documents will be added to the training set in each iteration. For example, a threshold of 0.05 will add five percent of the documents in each iteration resulting in twenty iterations.

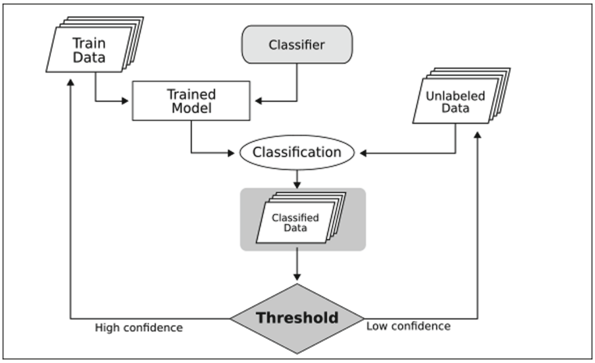


Fig. 1. CasSUL: a semi-supervised framework for corpus annotation.

The intuition behind CasSUL is that manual annotation is necessary in subjective tasks such as SA and should be part of the process. Besides that the iterative addition of new samples provide new information for the classifiers, thus resulting in better classification models for labeling the remaining unlabeled data.

CasSUL differs from other semi-supervised approaches presented in the literature [6,19,20]. Usually the authors use a percentage based on the confidence level (eg. only documents with 90% of confidence will be added to the training dataset). We have modified that since we addressed sentiment classification in three classes using unbalanced corpora as main resource. Preliminary experiments showed that classifiers trained with skewed data learned biased models and the majority class usually overcame the others, propagating the bias even more. Moreover, the use of this approach in three class classification was challenging since high values of confidence resulted in iterations adding just a few documents and finishing the run in five or less steps. The same issue remained even with lower confidence values.

3.1 Datasets

The experiments demanded a SA corpus in Brazilian Portuguese manually annotated and even though the literature offers several datasets for the task, they are either too small, or present strong unbalanced class distribution, or did not go through a reliable evaluation of the annotation process. We compiled and manually annotated a Brazilian Portuguese dataset of TV show commentaries from Twitter for our main experiments – the *TweetSentBR* (TTsBR) [4]. The corpus contains 15.000 tweets labeled in three classes and is split in train/test, 12.999 and 2.001 tweets, respectively.

For the experiments we used the same queries for creating TTsBR and extracted new data (117.050 documents) to be used as the unlabeled dataset. Since this dataset does not have manual annotation, we compared results achieved by models trained with TTsBR train set to the ones obtained with the CasSUL extended corpus (TTsBR train + unlabeled data) both predicting the TTsBR test set.

We also evaluated CasSUL framework using other corpora: *Buscape*, *Mercado Livre* and *Brazilian Elections*. Since we do not have a complete definition of the data we could not extract new documents for experiments so we only used the available data. *Buscape* is a product review corpus that contains 13.685 documents labeled with positive or negative tags [9]. *Mercado Livre* is also a product review corpus containing 43.818 documents also labeled in two classes [2]. Both of them are balanced corpora automatically labeled using scores provided by users. *Brazilian Elections* is a two-set tweet corpus about two running candidates for presidency in 2010 Brazilian Elections, Dilma and Serra [17]. The first set contains 66.643 documents and the second, 9.718. Both corpora are annotated in two classes and unbalanced.

Another dataset we used in our experiments is Pelesent [5]. The corpus was created through distant supervision using emojis and emoticons. It contains 980.067 tweets automatically labeled in two classes and unbalanced. We performed experiments evaluating models trained with Pelesent and with the CasSUL extended corpus.

3.2 Data Representation and Classification

Before running the experiments we preprocessed all the data tokenizing the documents, removing punctuation marks and symbols (we kept emojis and emoticons), replacing user names and urls (for the tags USER and URL) and mapping character repetitions, such as in "*I looooove Star Wars.*" normalized to"*I loove Star Wars.*" (keeping some clue of repetition is importante since it carries semantic information). The preprocessing was made using Enelvo [3], an NLP tool for normalization developed for Brazilian Portuguese.

For handling the data two forms of representation were used: word embeddings and a feature approach. The word embeddings used in this work are shown in [5] and were trained with 14 million tweets using word2vec [11] in two approaches - a 50-dimension skip-gram trained model and a 600-dimension c-bow trained model. The feature representation is based on six literature approaches: **(a) Bag-of-words:** a unigram bag-of-words with occurrence of terms. This method is quite usual in SA [1,2]. **(b) Negation words:** using a negation word list for Brazilian Portuguese, as proposed in [2]. We count the number of negation (such as "*no*","*never*" and "*nothing*") and append it to the data representation. **(c) Emoticons:** Emoticons are an enriched representation of emotions on Twitter [8,15]. They are composed of characters grouped to resemble emotion faces, such as angry or love. We used a list of positive and negative emoticons presented in [1,2] to identify the occurrence of polarity emoticons in the document. **(d) Emojis:** Emojis are similar to emoticons, but instead of groups of characters, they are composed by special characters that represent draws [5]. We used Emoji Sentiment Ranking [14] for obtaining positive, neutral and negative scores for each emoji, and the average of emojis score in the each document is used as feature. **(e) Sentiment lexicon:** We also used the sentiment lexicon Sentilex [18] for identifying sentiment words in the documents. The number of positive and negative words in each document is used as a feature. **(f) Part-of-speech tag:** The tagger from NLPnet [7] was used for extracting the PoS tags of the words, and the numbers of verbs, nouns, adjectives and adverbs of each document are used as features. They can be useful specially for the identification of the neutral class, since the number of adjectives may indicate polarity documents.

4 Experiments and Results

For our experiments with CasSUL we used six machine learning classifiers - Support Vector Machines, Bernoulli Naive Bayes, Logistic Regression, Multilayer Perceptron, Decision Trees and Random Forest. We used scikit-learn, a Python machine learning library, for the classifiers implementation and for calculating the prediction probabilities (used for ranking the most reliable labels in each iteration). Due to space constraints, the results presented in the following subsections were summarized, and details of every experiment can be found in https://bitbucket.org/HBrum/tweetsentbr/ as well as the framework itself.

4.1 Hyperparameter Optimization

We performed a grid-search scheme using different hyperparameter for each classifier. The polarity classification was evaluated by using a 10% subset of the training corpus (if the test corpus were used for evaluation, the parameters would be biased). For every classifier we combined each representation (as shown in Subsect. 3.2) and classifier parameters. We also added a machine learning based feature selection using an SVM as presented in [1].

Every model was trained five times for each classifier using the training set of TTsBR (removing the 10% validation subset). The averages of the results were taken in order to rank the executions and get the best sets of representations and hyperparameters. The combinations of classifiers and hyperparameters are listed below.

(a) SVM: With c values varying as 0.001, 0.01, 0.1, 1 and 10. The hyperparameters obtained were *bag-of-words + emoticons + emojis* and feature selection using *c=1.* **(b) Naive Bayes:** With alpha values varying as 0.1, 0.5 and 1. The best results were obtained with *bag-of-words + emoticons + emojis + sentiment lexicon + PoS tagging* and feature selection using *alpha=0.1.* **(c) Logistic Regression:** We did not explore any hyperparameters for logistic regression, only the representations. The best fit was *bag-of-words + emoticons + emojis + PoS tagging* and feature selection. **(d) Multilayer Perceptron:** Relu was used as activation function in our experiments with MLP. We varied the number of layers (1 and 2), the number of neurons with 30, 60, 100 and 200 (using always the same number even with two layers), the alpha with 0.001, 0.001 and 0.01 and the learning rate with 0.001, 0.01 and 0.1. The best representation was *bag-of-words + negation words + emoticons + emojis + sentiment words + PoS taggin* and feature selection. The best results were obtained with two layers, 200 neurons, alpha=0.001 and learning-rate=0.001. **(e) Decision Trees:** We varied the criterium of the tree split using gini and entropy, the maximum depth of the tree with 4, 5, 8 and with no limit. The best fit was *bag-of-words + negation words + emoticons + emojis* without feature selection using gini as criterium and not establishing a maximum depth. **(f) Random Forest:** We varied the number of estimators (30, 60, 100 and 200), the criterium (gini or entropy) and the maximum depth (4, 5, 8 and without limit). The best results were obtained with *bag-of-words + negation words + emoticons + emojis* and feature selection using 200 estimators, entropy as criterium and without maximum depth.

4.2 TTsBR + CasSUL Extension Using Unlabeled Data

For evaluating CasSUL framework we ran the framework using TTsBR train set (12.999 documents) as the manually labeled input and extended it with the unlabeled data extracted (117.090) documents. At the end of the last iteration the full corpus will contain 130.089 documents, being a combination of the manual labeled data with the unlabeled classified set. After we obtained the final corpus, we trained six models (using the six classifiers presented before) using it and predicted the TTsBR test set labels (2.001 documents), measuring the F1-measure of each model and averaged the value.

We repeated the process using seven thresholds of confidence empirically set (40%, 30%, 25%, 20%, 10%, 5% and 1%) and changing the classifier responsible for the classification model used by CasSUL. Table 1 presents the results obtained for each corpus generated (each cell represents the average F1-Measure of each corpus). For example, the first line of the Table presents the evaluation of a CasSUL extended corpus obtained with SVM classifier ran with each of the seven thresholds of confidence.

The intuition behind the thresholds is that using a lower threshold we ensure more iterations and less documents being added to the final corpus more slowly. Using a higher threshold we add more documents in each iteration (possibly adding more noise to the dataset).

Table 1. Average F1-measure obtained by each classifier using different thresholds in three polarity classes on TTsBR.

Classifier	**40%**	**30%**	**25%**	**20%**	**10%**	**5%**	**1%**
Linear SVM	59,58	58,73	59,47	58,73	55,80	54,15	52,12
Naive Bayes	54,97	53,69	52,73	52,41	50,18	49,45	47,09
Logistic regression	59,91	58,41	58,12	57,04	53,2	50,86	48,76
MLP	**62,14**	**61,65**	**61,74**	**61,40**	**61,19**	**61,02**	**61,04**
Decision tree	57,85	57,54	58,39	56,44	58,29	58,45	57,93
Random forest	57,72	55,99	54,33	53,31	49,61	49,23	49,06

The same process was applied to the original dataset obtaining 61.01 on average F1-Measure. The extended corpus obtained the best results using MLP classifier. Since we are using a held-out subset, our main goal was to achieve values close to the obtained with the original manually annotated corpus.

One phenomena observed during the experiments was the skewing of the majority class in TTsBR (positive). This skewing caused the positive documents to be added early and far more than the others. Some of the final corpora generated had only 7% of the documents labeled as neutral, while 63% were labeled positive. In order to reduce this skewing we used under-sampling [12], removing documents for the majority class to balance the corpora, and repeated the experiments. The results are shown in Table 2.

Table 2. Average F1-measure obtained by each classifier using different thresholds in three class sentiment analysis on balanced TTsBR.

Classifier	**40%**	**30%**	**25%**	**20%**	**10%**	**5%**	**1%**
Linear SVM	60,57	60,84	60,69	60,81	60,91	59,19	56,60
Naive Bayes	57,16	56,08	55,27	54,36	49,06	46,78	45,01
Logistic regression	61,45	**61,71**	**61,48**	61,55	58,93	53,56	50,30
MLP	**62,13**	60,64	61,10	**61,68**	**61,60**	**61,50**	**61,64**
Decision tree	58,36	57,72	58,00	57,58	57,68	57,96	58,31
Random forest	59,34	58,04	57,29	55,52	52,01	50,66	48,99

Although the results did not improve, the final corpora kept the balance. Neutral documents rose from 7% to 15% when using under-sampling. All the distributions (with and without under-sampling) had positive as majority class in the extended corpora. We believe this class may be the easiest to set apart from the neutral and negative, achieving the best confidence levels and being added more frequently to the corpora.

4.3 Using CasSUL on Fully Labeled Corpora

We also evaluated CasSUL using only manually labeled data. For this experiment we used Buscapé Corpus, Mercado Livre Corpus and Brazilian Elections (Elections-Dilma and Elections-Serra). For each corpus we used a 10% sample of the data observing the labels and the remaining 90% of the documents as unlabeled data. All the experiments on these corpora were performed on binary classification, since none of them contains neutral documents. We used the same hyperparameters as the experiments presented in Subsect. 4.1. The only difference is the absence of a test set for evaluation since in this scenario we had the correct labels for the data automatic annotated.

Our best extended corpora with Buscapé obtained 84.74% of F1-Measure using Logistic Regression with 1% threshold. Without extension we achieved 87.66% of F1-Measure on the same 10-fold scheme. For Mercado Livre we achieved 93.17% of F1-Measure also with Logistic Regression with threshold 10%, but still under the 94.76% obtained without extension. Using Elections-Dilma corpus we obtained 83.69% with a MLP extended corpus using 1% threshold as the best value. Without the extension the same experiment resulted in 93.15% F1-Measure. For Elections-Serra we achieved 88.23% with a Random Forest extension using 30% threshold, but 93.63% without extension. Both of the corpora are unbalanced and we can see how this skewing affects our framework.

4.4 Comparison Between Our Approach Vs Distant Supervision

Distant Supervision is a popular method for annotating a large scale corpus for SA, since it demands almost no human effort and can be scalable for hundreds of thousands of documents, even on social networks. Pelesent [5] is a distant supervision corpus for Brazilian Portuguese created using emojis and emoticons. It contains 980.067 tweets automatically labeled through this approach. We have compared Pelesent with the extended TTsBR generated by CasSUL, with 117.090 documents, on the polarity classification task.

Three classifiers, Linear SVM, MLP and Logistic Regression, were used for training models using Pelesent and TTsBR extended (the corpus obtained with MLP using threshold 30%) and the evaluation was carried on cross domain corpus (Buscapé, Mercado Livre and Brazilian Elections). In this scenario we did not used the optimized hyperparameters, since it was designed for other domain. Instead we used the word embeddings trained in [5] with word2vec, 600-dimensions/c-bow.

Since Pelesent is only labeled in two classes, we used only pos/neg TTsBR documents, reducing its size to 128.030 documents total. In Table 3 we can see the results of general F1-Measure for each class. Although seven times smaller, the corpus annotated via CasSUL framework achieved better results than Pelesent in almost every experiment.

Table 3. Comparison between TTsBR extented using MLP with 30% threshold and Pelesent on cross-domain polarity classification.

Evaluation corpus	Classifier	Extended TTsBR			Pelesent		
		F-pos	F-neg	F-Measure	F-pos	F-neg	F-measure
Elections-Dilma	**Linear SVM**	**52,97**	**73,90**	**63,45**	80,8	42,6	61,69
	Log.	**56,03**	**75,27**	**65,66**	79,8	40,3	60,06
	MLP	51,90	72,90	62,39	**79,2**	**46,3**	**62,78**
Elections-Serra	**Linear SVM**	**79,33**	**28,80**	**54,07**	20,4	36,5	28,45
	Log. Regression	**79,70**	**28,27**	**53,99**	20,7	35,0	27,81
	MLP	**78,47**	**27,00**	**52,72**	20,3	39,6	29,90
Mercado livre	**Linear SVM**	**84,90**	**82,97**	**83,93**	77,5	62,6	70,01
	Log. Regression	**85,07**	**82,90**	**83,97**	77,6	62,5	70,04
	MLP	**85,07**	**83,60**	**84,34**	79,3	69,8	74,54
Buscapé-1	**Linear SVM**	**66,53**	**73,47**	**69,99**	70,1	56,8	63,46
	Log. Regression	**69,87**	**73,87**	**71,86**	70,3	57,3	63,80
	MLP	**65,90**	**73,10**	**69,49**	70,6	62,3	66,41
Buscapé-2	**Linear SVM**	**77,40**	**79,73**	**78,55**	72,9	53,9	63,39
	Log. Regression	**78,63**	**79,77**	**79,18**	73,1	54,2	63,63
	MLP	**77,47**	**79,50**	**78,45**	73,8	57,0	65,38

5 Discussion and Future Work

In this paper we presented CasSUL, a framework for annotation of sentiment corpora using self-training. Using CasSUL, the size of the manually annotated corpus TTsBR was extended eight times. The extended TTsBR, when used in a polarity classification task, achieved similar results when compared to the original corpus. This performance has to be confirmed in other NLP tasks, but this evidence of preservation of the annotation quality encourages us to create more representative annotated corpora, as new annotated examples are added to them, with no additional annotation cost. CasSUL was also superior to the most popular alternative for automatic labeling a large datasets for SA (Distant Supervision), even with a corpus ten times bigger.

CasSUL is limited by the self-training known weaknesses as error propagation and skewed class distributions, but this could be reduced by the use of other techniques such as under-sampling or of other semi-supervised alternatives like co-training, for example. Despite of the results with annotated corpora had been inferior to the ones of non-extended corpora, we believe that the size of the datasets may be a key factor on this issue.

Improvements in CasSUL could include: the addition of a step where manual annotators could revise the machine labels (Active Learning), as successfully reported in [6]; more classifiers could be added in order to improve even more the confidence on the final corpus. The use of Deep Learning methods are very recommended since neural architectures have been achieving state-of-the-art values regularly and this approach can be easily inputed in CasSUL.

Acknowledgement. We acknowledge financial support from CNPq and CAPES for the financial support during the experiment that originated this research paper.

References

1. Avanço, L.V., Brum, H.B., Nunes, M.: Improving opinion classifiers by combining different methods and resources. XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), pp. 25–36 (2016)
2. Avanço, L.V.: Sobre normalização e classificação de polaridade de textos opinativos na web (2015)
3. Bertaglia, T.F.C., Nunes, M.G.V.: Exploring word embeddings for unsupervised textual user-generated content normalization. In: WNUT 2016, p. 112 (2016)
4. Brum, H., Nunes, M.G.V.: Building a sentiment corpus of tweets in brazilian portuguese. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), May 2018
5. Correa Jr., E.A., Marinho, V.Q., dos Santos, L.B., Bertaglia, T.F., Treviso, M.V., Brum, H.B.: Pelesent: cross-domain polarity classification using distant supervision. arXiv preprint arXiv:1707.02657 (2017)
6. Dasgupta, S., Ng, V.: Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL 2009, pp. 701–709. Association for Computational Linguistics, Stroudsburg (2009)
7. Fonseca, E.R., Rosa, J.L.G., Aluísio, S.M.: Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. J. Braz. Comput. Soc. **21**(1), 2 (2015)
8. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, **1**(2009) 12 (2009)
9. Hartmann, N.S., et al.: A large corpus of product reviews in Portuguese: tackling out-of-vocabulary words. In: 9th International Conference on Language Resources and Evaluation (2014)
10. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
12. Monard, M.C., Batista, G.E.: Learning with skewed class distrihutions. Adv. Log. Artif. Intell. Robot. LAPTEC **85**(2002), 173 (2002)
13. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: Semeval-2016 task 4: sentiment analysis in Twitter. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016) (2016)
14. Novak, P.K., Smailović, J., Sluban, B., Mozetič, I.: Sentiment of emojis. PloS one **10**(12), e0144296 (2015)
15. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc, vol. 10 (2010)
16. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP 2002, pp. 79–86. Association for Computational Linguistics, Stroudsburg (2002)

17. Silva, I.S., Gomide, J., Veloso, A., Meira Jr, W., Ferreira, R.: Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 475–484. ACM (2011)
18. Silva, M.J., Carvalho, P., Sarmento, L.: Building a sentiment lexicon for social judgement mining. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS (LNAI), vol. 7243, pp. 218–228. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28885-2_25
19. Silva, N.F.F.D., Coletta, L.F.S., Hruschka, E.R.: A survey and comparative study of tweet sentiment analysis via semi-supervised learning. ACM Comput. Surv. **49**(1), 15:1–15:26 (2016)
20. da Silva, N.F.F., Coletta, L.F., Hruschka, E.R., Hruschka Jr., E.R.: Using unsupervised information to improve semi-supervised tweet sentiment classification. Inf. Sci. **355**, 348–365 (2016)
21. Socher, R., et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642 (2013)
22. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, pp. 417–424. Association for Computational Linguistics, Stroudsburg (2002)

Speech Processing

What Weighs for Word Stress? Big Data Mining and Analyses of Phonotactic Distributions in Brazilian Portuguese

Amanda Post da Silveira[1,2](✉), Eric Sanders[3], Gustavo Mendonça[4], and Ton Dijkstra[1]

[1] Donders Centre for Cognition (DCC), Radboud University Nijmegen, Nijmegen, The Netherlands
psyphon.ap@gmail.com
[2] Potiguar University – Laureate International Universities, Natal, Brazil
[3] Centre for Language and Speech Technology (CLST), Radboud University Nijmegen, Nijmegen, The Netherlands
[4] Departments of Computer Science and Computer Mathematics, University of São Paulo, São Paulo, Brazil

Abstract. For about four decades, phonological theories have claimed that word stress assignment depends on the word's syllabic phonotactic complexity in relation to syllabic position. This study analyzes the phonotactic implications for word stress Brazilian Portuguese. After creating a phonotactic corpus and applying Random Forest modeling, phonotactic distributions for word stress were found to be bound to stress pattern and word length in number of syllables. To account for these observations, models of word naming must be extended with aspects of word stress.

Keywords: Phonotactics · Word stress · Brazilian Portuguese

1 Introduction

How acoustic properties of vowels and suprasegmental acoustic features affect word stress assignment has been increasingly debated and theoretically explored in phonetic and psycholinguistic studies [1–6]. At the same time, relatively few empirical studies have considered how the segment sequence within a syllable and the syllabic distributions in words of different lengths affect word stress assignment. Although many studies have been dedicated to this topic in theoretical linguistics, available models are controversial and lack systematic empirical testing. For instance, studies have sometimes used restricted samples of convenience that meet proposed theoretical generalizations, but leave the empirical coverage of the addressed phenomena unexplored and unresolved [7]. In the present study, we created a phonotactic corpus for Brazilian Portuguese for which the word stress system is hypothesized to be weight-sensitive, and then subjected this corpus to a big data analysis to investigate the relationship between phonotactic distributions and word stress.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 399–408, 2018.
https://doi.org/10.1007/978-3-319-99722-3_40

1.1 Phonological Claims About Weight Sensitivity

Sequences of speech sounds in spoken words are considered to be highly language specific. The distribution of sounds (e.g., vowels and consonants) within syllables and words, including the abstract allocation of vowels and consonants to syllables, is called *phonotactics* [8]. Depending on their phonotactics, syllables vary in ‘weight’: They can be light, heavy, and even superheavy given language-specific requirements [9–11].

Formal phonological approaches consider the initial consonants of syllables as irrelevant to weight. In some languages, there is an opposition between short and long vowels, making structures such as (C)V light compared to (C)VV and VC, which are considered heavy. Sequences such as VVC and VCC are called superheavy syllables in languages that call for the distinction heavy versus superheavy. Syllable weight is considered an important factor in word stress assignment in languages that are referred to as weight or quantity sensitive stress systems.

The notion of syllable quantity comes from the counting of syllabic moras - which are metrical units of time. If a syllable contains a long vowel or a diphthong (VV) or a short vowel followed by a consonant in coda (VC), it is assumed to have two moras. In terms of syllable quantity, a syllable is heavy when it has more than one mora [12–14].

In Brazilian Portuguese, [15] claimed that light syllables, such as V and CV, are the most frequent. Heavy syllables, such as VC and CVC or VV/CVV (diphthongs) are the least frequent, and few consonants are accepted in coda position. Weight-sensitivity is supposed to be conditioned by syllabic position: Antepenultimate syllables and final syllables are stressed if they have a heavy syllable, such as in *CÔmoda* (‘dresser’) and *jacaRÉ* (‘alligator’)[1] [16].

In sum, the syllabic position of word stress must be taken into account when specifying the relationship between phonotactics and word stress. Importantly, phonotactic patterns are assumed to relate to word stress patterns in a bottom-up fashion, i.e. from phonotactics to word stress. In the next section, we will review phonological claims for weight sensitivity in the light of empirical findings on the acoustic of vowels and syllables. This will allow us to check the assumption that phonotactics motivates stress rather than the reverse.

1.2 Empirical Evidence Motivating Effects from Syllabic Weight to Stress

Phonetics Studies

In syllabic stress contrasts, the concept of vowel length plays an important role. In this context, vowel length may be partly determined by inherent properties of the vowel, word length in number of syllables, and word stress. Vowels are phonemically distinguished in terms of vowel spectra - vowel formant frequencies [17] - and intrinsic values of F0 and duration [18]. However, vocalic durations are also relative values, because at the word level they can only be interpreted relative to the durations of neighboring syllabic vowels. By computing a vowel duration ratio, we can then

[1] Antepenultimate and final stress patterns are considered to be exceptional in Portuguese and have diacritics on the stressed vowel to mark stress in orthography.

determine that one vowel is longer than the other [19, 20]. A meaningful distinction between long and short versions of a particular vowel (assuming the full versus reduced vocalic status of syllabic nuclei) in stressed and unstressed syllabic positions has been consistently observed. However, observed differences in vowel duration are not large enough to reliably define duration as a phonological contrast for stressed and unstressed vowels independently from spectral features [21–23]. Consequently, although vocalic length is supposed to provide a distinction between long and short vowels and affects word stress assignment depending on syllabic position within a word, it is not a sufficient determinant of word stress assignment.

Phonetic studies have systematically found evidence that phrasal and word stress affect the syllabic structure of a word, but not the reverse. Syllables in stressed position do not show segmental deletion or syllabic reduction, while unstressed syllables frequently suffer from processes such as vocalic reduction, phonotactic simplification, liaison, and elision in word naming and in connected speech. For example, in the BP word *dificuldade*, two unstressed simple syllables are often deleted, /dZif.kuw $^{\mathrm{I}}$dadZ/instead of / dZi.fi.kuw.da.dzi/; and, in English examples, cases of elision such as *peer review /p*ɪ.rɪ ˈvju/, and liaison such as *car audio* /ka.rɑ.di.oʊ/are motivated by phrasal stress. Syllabic reduction and syllabic merging processes seem to happen across all lexica in connected and spontaneous speech, independently of whether the lexica are defined as mainly stress-timed or syllable-timed. In the next session, we will provide a method of corpus creation and data mining to empirically predict the phonological representation of word stress in Brazilian Portuguese.

2 Method

2.1 Building a Phonotactic Corpus for Brazilian Portuguese

In order to create the Brazilian Portuguese phonotactic corpus, we made use of other freely available corpora and tools: for Brazilian Portuguese, the Avaliação Sonora do Português Atual (ASPA) [24], the machine-readable pronunciation dictionary for Brazilian Portuguese, Aeiouadô [25] and The Bank of Portuguese corpus [26].

The phonotactic parsing method in this study used the X-SAMPA transcriptions generated by or for (in the case of Brazilian Portuguese) the corpora cited above. Phonotactic transcriptions were automatically generated from the X-SAMPA transcriptions of each word, by mapping each phonemic segment onto its corresponding phonotactic segment and composing the corresponding phonotactic word. For instance, the BP word *refletir* is transcribed in X-SAMPA as /xe.fle'tSiR/[2] and its phonotactic transcription is CV-CCV-CVC. Each phonotactic word was generated using syllable identification and later sorted for our inventories. Vocalic segments were phonotactically transcribed as: V for vowels, D for diphthongs, and GD or DG for triphthongs. Consonantal segments were transcribed as C. For the ASPA corpus, we will now describe specific adaptations to make it useful for our purposes. Next, we will analyse

[2] The sequence /tS/ in X-SAMPA corresponds to one sound, the voiceless palato-alveolar affricate, which in IPA is represented by the symbol /ʧ/.

the resulting phonotactic inventory with respect to properties related to word stress and syllable position. Specifically, the following variables are considered to play a role in the phonotactics-to-stress relationship: (i) Word Length: As there is evidence for syllabic effects on lexical access [26–28], we chose disyllabic and trisyllabic words for our analyses; (ii) Stress Status of syllables; (iii) Stress Pattern; (iv) Syllabic Position; (v) Phonotactic Pattern; and (vi) Language.

2.2 Phonotactic Corpus for Brazilian Portuguese

We generated a list of Brazilian Portuguese orthographic words and their respective frequencies of occurrence from the ASPA (ASPA: Avaliação Sonora do Português Atual) corpus [24], which is based on The Bank of Portuguese corpus [26]. The Bank of Portuguese is a large cumulative, open source corpus of texts (newspapers, magazines, literature books, academic and business written samples) and oral transcriptions (conversations, meetings, lectures, phone chats, interviews), which comprise 228,766,402 tokens and 607,392 types.

Next, we automatically generated phonetic transcriptions in Aeiouadô [25], which is a hybrid grapheme-to-phoneme converter (G2P) for Brazilian Portuguese that makes use of both rules and machine learning algorithms (see Fig. 1).

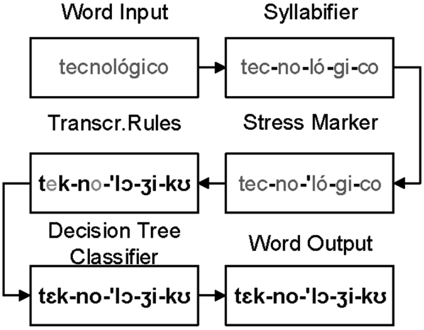


Fig. 1. Algorithm for obtaining phonetic transcription in AEIOUADÔ for the example word *tecnológico* (‘technological’) [25].

The syllabification algorithm follows a rule-approach and is based straightforwardly on the syllabification rules described in the Portuguese Language Orthographic Agreement. As for the stress marker, once the syllable structure is known in Brazilian Portuguese, one can predict where stress falls. Stress falls: (1) on the antepenultimate syllable if it has an accented vowel: <á, â, é, ê, í, ó, ô, ú>; (2) on the ultimate syllable if it contains the accented vowels <á, é, ó> or <i, u> or if it ends with one of the consonants <r, x, n, l, z>; (3) otherwise, on the penultimate syllable. The Aeiouadô provides IPA transcriptions. The tool was modified so that automatic transcriptions were provided in SAMPA. From the SAMPA transcription of each word, phonotactic patterns were extracted.

The resulting Brazilian Portuguese Phonotactic Corpus provides information on the number of graphemes and phonemes in a word, as well as on its number of syllables, word stress pattern, phonotactic distribution per Syllabic Position, and the word's frequency of occurrence in the corpus. The included words vary in length from 1 to 10 syllables. In total, the corpus contains 123,826 lexical transcriptions (lexical types) and 228,766,402 tokens. For instance, in a search for the word *abacaxi* ('pineapple'), the corpus provides the following information: It has 7 graphemes, 7 phonemes, 4 syllables, stress is assigned to the last syllable, its SAMPA transcription corresponds to /a. ba.ka'Si/, its phonotactic transcription corresponds to V-CV-CV-CV, and its token frequency in the corpus is of 874 occurrences.

3 Results

3.1 Statistical Analyses of the Phonotactic Properties

Next, we applied two different methods to analyze the phonotactic inventories of our target language: Conditional Inference Trees and Random Forest modeling. We could have used Multiple Regression Analyses to test to what extent token frequencies of phonotactic structures have a significant effect on stress assignment with respect to a particular syllable in words that differ in number of syllables. However, although the method has the advantage of clearly indicating significant phonotactic patterns, it does not consider hierarchical dependencies among the included variables, which are our main interest in this study.

Following innovations in the statistical modeling of linguistic database analysis by [27] we therefore also applied a Random Forest technique available in the party package R [28, 29], which implements forests of conditional inference trees. Conditional Inference Trees and Random Forest models vary in their explanatory power of factor interactions.

Conditional Inference Trees are constructed based on series of binary decisions that are made with respect to the values of the predictor variables (in our study, Phonotactic Pattern, Stressed Syllable, Syllabic Position, and Word Stress Pattern). The model provides likelihood estimates for predictor variables based on response variable values (in our study, Log token frequency). For instance, for the factor Stressed, it considers whether splitting the data into one of the three possible stress patterns affects the frequency of use of certain Phonotactic Patterns. The Conditional Inference Tree model represents it as a first significant split (or node) based on Stressed Syllable and a second significant split (or node) based on Phonotactic Patterns. Thus, the Conditional Inference Tree algorithm considers all predictors in the analysis and splits the data into subsets whenever the data likelihood allows it. This algorithm is applied in recurrent loops over all the subsets of the model, until no further partitioning is needed, providing an exhaustive and homogenous analysis of predictor interactions based on the data.

The Random Forest approach constructs a large number of these conditional inference trees. Each of these trees contributes a vote based on what it proposes as the most likely outcome response variable (the 'importance measure', implemented in the

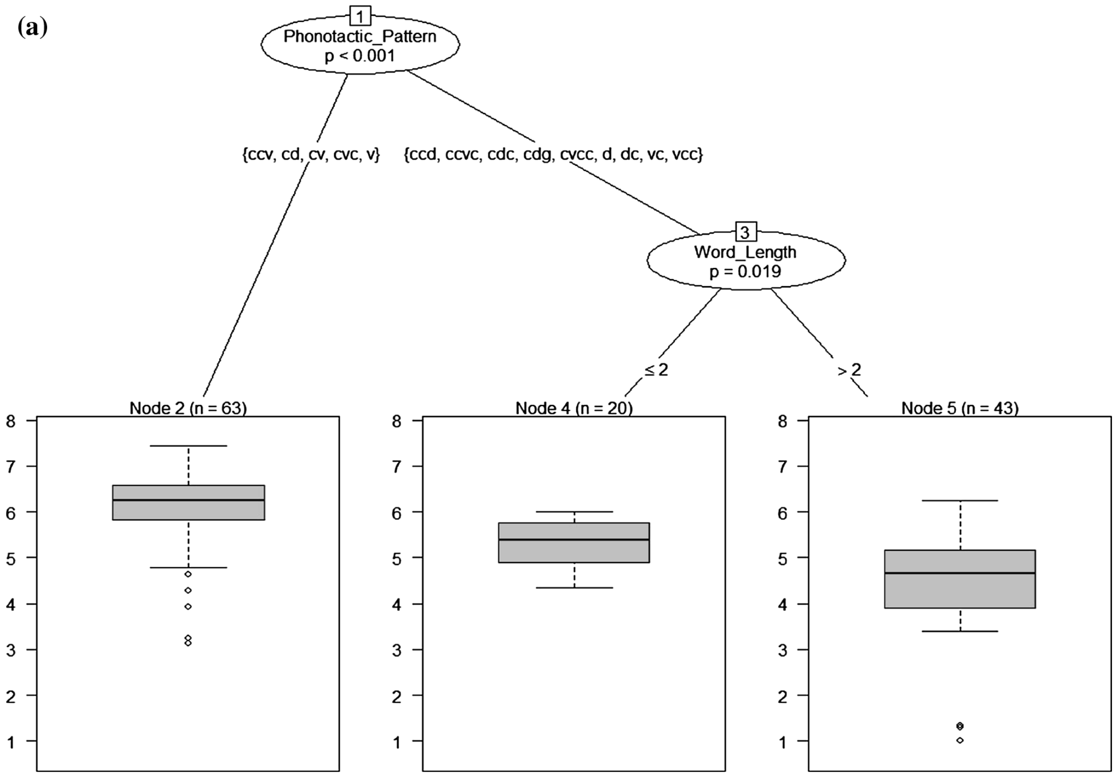


Fig. 2. (a) Conditional Inference Tree for Log Frequency of Phonotactic Patterns to Word Stress of Brazilian Portuguese. The X-axis shows the conditional relationships among the variables and the Y-axis shows the proportion of observations. (b) Conditional permutation variable importance for the Random Forest of Brazilian Portuguese (X-axis) with five predictors (Y-axis).

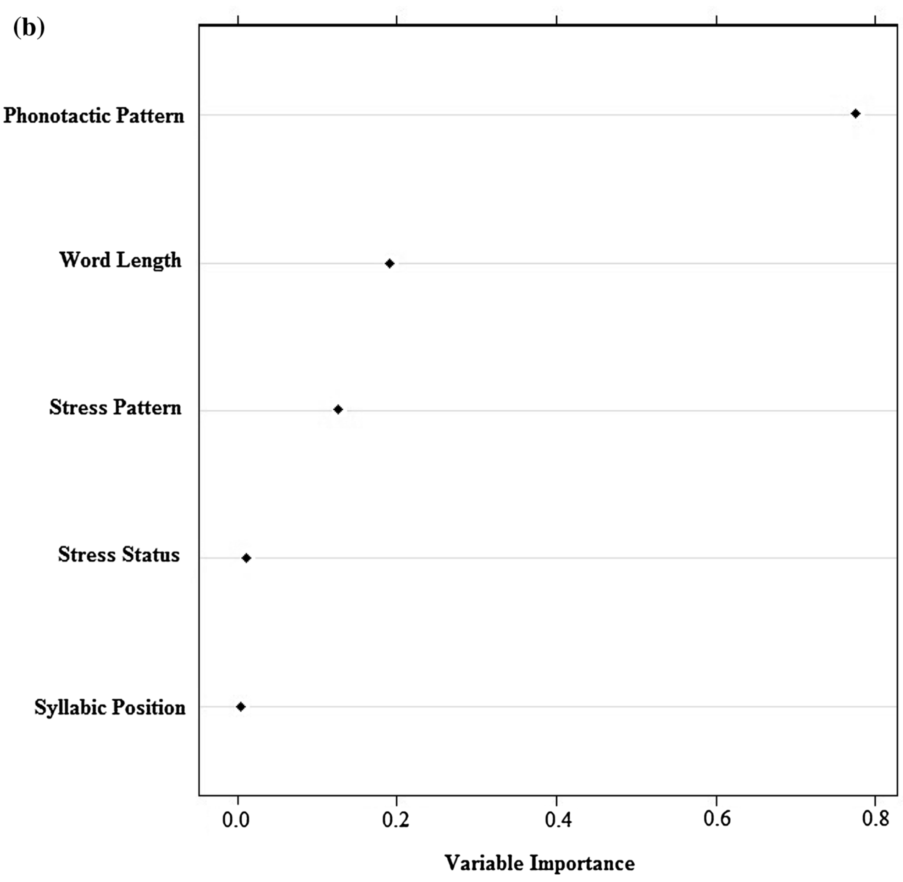


Fig. 2. (*continued*)

Random Forest function of the party package, [28, 29]. The advantage of the Random Forest approach is that all predictor variables are exhaustively analyzed with respect to their importance for the response variable, as multiple trees are created by exclusion and inclusion of predictors via the generation of multiple possible trees. A disadvantage is that the hundreds of conditional inference trees created by the Random Forest algorithm are difficult to describe and display in research papers. The sum of the multiple trees may be accessed via a Variable Importance graph, but the multiple variable interactions that the Random Forest model provides are impossible to be visualized. For this reason, [38] suggested to combine the Confidence Inference Trees with the variable Importance from Random Forest models in linguistic corpus analyses. These analyses are shown in Figs. 2(a) and 2(b).

The Conditional Inference Tree in Fig. 2(a) shows relatively simple interactions in the data. The index of concordance of this model is C = 0.60, which accounts for approximately 36% of the prediction accuracy. The predictors related to word stress

(Stressed Syllable, Syllabic Position, and Word Stress Pattern) did not make it into the tree, which indicates that Phonotactic Patterns in Brazilian Portuguese are weak predictors of Word Stress. An alternative explanation is that Word Stress factors are strongly correlated to Phonotactic Patterns and/or Word Length [30]. In this model, the most important subset is Phonotactic Pattern (node 1). In general, for disyllabic and trisyllabic words the following patterns are preferred (node 2): CCV, CD, CV, CVC, V. Word Length is affected by Phonotactic Pattern (node 3) and the CCD, CCVC, CDC, CDG, CVCC, D, DC, VC, VCC phonotactic patterns are the second favorites for disyllabic words and trisyllabic words (node 5). Figure 2(b) shows the variable importance graph based on the evidence from 126 conditional inference trees generated by the random forest model of BP. It has a concordance of $C = 0.76$, which corresponds to a prediction accuracy of 58% and to 18% improvement over the performance of the single conditional inference tree. The most important variable according to the Random Forest analysis is Phonotactic Pattern, followed by Word Length and Stress Pattern in decreasing order of importance. The two least important predictors are Stress Status and Syllabic Position.

The current analyses provide no indications that there is a straightforward relationship between word stress and phonotactic distributions in this language. On the other hand, the agreement between the analyses involving Conditional Inference Trees and Random Forest Variable Importance were reasonably high for Brazilian Portuguese. In sum, the models indicate that the relationship between phonotactics and word stress was top-down in the sense that word length and word stress determined the phonotactic distributions in Brazilian Portuguese.

4 Discussion

The motivating question for the present study was: What weighs for word stress? In other words, which phonological factors contribute to the assignment of word stress in different languages? To investigate this issue, we decided to use a corpus analyses, because the factors involved were too many and too novel (such as the factor Word Length in number of syllables) to be tested via experimental methods. We created a phonetic corpus for words of Brazilian Portuguese, based an existing corpus, but including the phonotactic transcription of words. We then analyzed the phonotactic properties of this corpus in detail by means of a number of statistical approaches.

In Brazilian Portuguese, Phonotactic Patterns were important predictors of Word Length, but they were not very important for word stress related predictors such as word Stress Pattern and Stress Status. This suggests that for this language, phonotactics do not motivate stress patterns or play only a marginal role in stress assignment. However, phonotactics seems to change because of word length in number of syllables. The finding can be explained primarily by word length – short words reduce less than long words - and only secondarily by word stress. Note that in long words unstressed syllables are deleted; segments remaining after syllable deletions then cluster as codas of stressed neighbouring syllables.

These findings have implications for psycholinguistic models for word stress assignment in word production and word recognition. First, the relationship between phonotactics and word stress is not linear or uni-directional. Second, word stress seems to come first in the process of encoding segmental sequences into syllabic nodes prior to speech production. Therefore, word stress cannot be integrated only later in word production, after the phonotactic distribution is already encoded, as most psycholinguistic models suggest. Third, word length may change the relationship that other predictors hold with word stress, such as that between phonotactic patterns and word stress patterns. Thus, word length should be included as a factor of analysis in psycholinguistic models.

References

1. Cutler, A.: Forbear is a homophone: lexical prosody does not constrain lexical access. Lang. Speech **29**, 201–220 (1986)
2. Cooper, N., Cutler, A., Wales, R.: Constraints of lexical stress on lexical access in English: evidence from native and non-native listeners. Lang. Speech **45**(3), 207–228 (2002)
3. van Heuven, V.J.J.P., Sluijter, A.M.C.: Effects of focus distribution, pitch accent and lexical stress on the temporal organisation of syllables in Dutch. Phonetica **55**, 71–89 (1995)
4. Braun, B., Galts, T., Kabak, B.: Lexical encoding of L2 tones: the role of L1 stress, pitch accent and intonation. Second Lang. Res. **30**(3), 323–350 (2014)
5. Post da Silveira, A., van Heuven, V., Caspers, J., Schiller, N.O.: Dual activation of word stress from orthography: the effect of the cognate status of words on the production of L2 stress. Dutch J. Appl. Linguist. **3**(2), 170–196 (2014)
6. Post da Silveira, A., van Leussen, J.W.: Generating a bilingual lexical corpus using interlanguage normalized Levenshtein distances. In: Proceeding of the 18th International Conference of Phonetic Sciences (XVII ICPhS), Glasgow, UK (2015)
7. Domahs, U., Plag, I., Carroll, R.: Word stress assignment in German, English and Dutch: quantity-sensitivity and extrametricality revisited. J. Comp. German. Linguist. **17**(1), 59–96 (2014)
8. Vitevitch, M.S., Luce, P.A., Charles-Luce, J., Kemmerer, D.: Phonotactics and syllable stress: implications for the processing of spoken nonsense words. Lang. Speech **40**, 47–62 (1997)
9. Hayes, B.: A metrical theory of stress rules. [Doctoral thesis MIT, US]. Revised version distributed by IULC, published by Garland Press, New York (1981)
10. Hyman, L.: A Theory of Phonological Weight. Foris, Dordrecht (1985)
11. Kager, R.: A metrical theory of stress and distressing in english and dutch. [Doctoral thesis, Utrecht University, NL] (1989)
12. Kiparsky, P.: From cyclic phonology to lexical phonology. Struct. Phonol. Represent. **1**, 131–175 (1982)
13. Hayes, B.: Extrametricality and English Stress. Linguist. Inquiry **13**, 227–276 (1982)
14. Trommelen, M., Zonneveld, W.: Klemtoon en metrische fonologie. Dick Coutinho, Muiderberg (1989)
15. Mattoso Câmara Jr., J.: Problemas de Lingüística Descritiva. Vozes, Petrópolis (1969)
16. Bisol, L.: Mattoso Câmara Jr. e a palavra prosódica. DELTA **20**, 59–70 (2004)
17. Ladefoged, P.: A Course in Phonetics. Harcourt Brace Jovanovich, Fort Worth (1975)
18. Peterson, G.E., Lehiste, I.: Duration of syllable nuclei in English. J. Acoust. Soc. Am. **32**, 693 (1960)
19. Major, R.: Stress and rhythm in Brazilian Portuguese. Language **61**, 259–282 (1985)
20. Barbosa, P.A.: Incursões em torno do ritmo da fala. Pontes, Campinas (2006)

21. Lehiste, I.: Suprasegmentals. The MIT Press, Cambridge (1970)
22. Wang, X., Pols, L.C.W., ten Bosch, L.F.M.: Analysis of context-dependent segmental duration for automatic speech recognition. In: Proceedings of 4th International Conference on Spoken Language, pp. 1181–1184 (1996)
23. Ciszewski, T.: Stressed vowel duration and phonemic length contrast. Res. Lang. **10**(2), 215–223 (2012)
24. Cristófaro-Silva, T., de Almeida, L.S., Fraga, T.: ASPA: A Formulação de um Banco de Dados de Referência da Estrutura Sonora do Português Contemporâneo. In: Proceedings XXV Congress of Brazilian Society of Computing Science, São Leopoldo, RS, Brazil (2005)
25. Mendonça, G., Aluísio, S.: Using a hybrid approach to build a pronunciation dictionary for Brazilian Portuguese. In: Proceedings of the 15th INTERSPEECH, pp. 1278–1282 (2014)
26. Berber Sardinha, T.: The Bank of Portuguese, DIRECT Papers, 50, São Paulo/Liverpool: LAEL, PUCSP/AELSU, University of Liverpool (2003)
27. Tagliamonte, S., Baayen, H.: Models, forests and trees of York English: was/were variation as a case study for statistical practice. Lang. Var. Change **24**, 135–178 (2012)
28. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. BMC Bioinform. **9**, 307 (2008)
29. Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform. **8**, 25 (2007)
30. Bernaisch, T., Gries, S.T., Mukherjee, J.: The dative alternation in South Asian English(es): modelling predictors and predicting prototypes. Engl. World-Wide **35**(1), 7–31 (2014)

Sentence Segmentation and Disfluency Detection in Narrative Transcripts from Neuropsychological Tests

Marcos Vinícius Treviso(✉) and Sandra Maria Aluísio

Interinstitutional Center for Computational Linguistics (NILC),
Institute of Mathematical and Computer Sciences, University of São Paulo,
São Paulo, Brazil
marcostreviso@usp.br, sandra@icmc.usp.br

Abstract. Natural Language Processing (NLP) tools aiming at the diagnosis of language impairing dementias generally extract several textual metrics of narrative transcripts. However, the absence of sentence boundary segmentation in transcripts prevents the direct application of NLP methods which rely on these marks to work properly, such as taggers and parsers. We present a method to segment the transcripts into sentences and another to detect the disfluencies present in them, to serve as a preprocessing step for the application of subsequent NLP tools. Our methods use recurrent convolutional neural networks with prosodic, morphosyntactic features, and word embeddings. We evaluated both tasks intrinsically, analyzing the most important features, comparing the proposed methods to simpler ones, and identifying the main hits and misses. In addition, a final method was created to combine all tasks and it was evaluated extrinsically using 9 syntactic metrics of Coh-Metrix-Dementia. In the intrinsic evaluations, we showed that our method achieved (i) state-of-the-art results for the sentence segmentation task on impaired speech, and (ii) results that are similar to related works for the English language for disfluency detection tasks. Regarding the extrinsic evaluation, only 3 metrics showed a statistically significant difference between manual MCI transcripts and those generated by our method, suggesting that our method is capable to preprocess transcriptions to be further analyzed by NLP tools.

Keywords: Sentence segmentation · Disfluency detection
Impaired speech

1 Introduction

In recent years, Mild Cognitive Impairment (MCI) has received great attention because it may represent a preclinical stage of Alzheimer's Disease (AD). Several studies have shown that speech production is a sensitive task to detect aging effects and to differentiate individuals with MCI from healthy ones. Automated

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 409–418, 2018.
https://doi.org/10.1007/978-3-319-99722-3_41

linguistic analysis tools have been applied to transcripts of narratives in English [10,11] and also in Brazilian Portuguese (BP) [1]. The latter study used a publicly available tool, Coh-Metrix-Dementia[1], to extract 73 textual metrics of narrative transcripts, comprising several levels of linguistic analysis from word counts to semantics and discourse. However, the absence of sentence boundary information and the presence of disfluencies in transcripts prevent the direct application of Natural Language Processing (NLP) methods that depend on well-formed texts, such as taggers and parsers.

To enable the correct functionality of NLP tools that analyze potentially impaired speech transcripts, we propose a method to automatically enrich these transcripts. The automatic enrichment in this case is related to the Sentence Segmentation (SS) task, which attempts at finding sentence boundaries, and the Disfluency Detection (DD) task, which is concerned with finding regions of disfluencies and categorizing them into their types. Disfluencies types are usually divided in: (i) fillers, which are used by an interlocutor to indicate hesitation or to keep control of a conversation, e.g. "ah, hm, bom, ent ao, digo"; and (ii) edit disfluencies, which occur when the interlocutor makes a statement that is not complete or correct and therefore corrects or changes his statement, e.g. "ela vai pro castelo pro castelo na verdade ela vai trabalhar no castelo né" in Fig. 1.

The paper is organized as follows. Section 2 presents related work on SS and DD tasks; Sect. 3 describes the dataset used in the evaluations; Sect. 4 presents our proposed model based on recurrent convolutional neural networks which was evaluated in both tasks (SS and DD). Section 5 presents findings and discussions of four evaluations (the first three are intrinsic ones and the last is extrinsic): sentence segmentation, filler detection, edit distance detection and whether the pipeline of tasks proposed in Fig. 3 can be used to automatically process narratives to be evaluated with 9 syntactic metrics of Coh-Metrix-Dementia. Finally, Sect. 6 concludes the paper and outlines some future work.

2 Related Works

The first methods proposed in the literature to deal with SS and DD tasks were based on generative models to deal with textual information, such as HMMs and language models [4,7], and on decision trees to treat prosodic information [15,16]. Discriminative methods that take into account the prediction structure, such as Conditional Random Fields (CRFs), replaced previous methods and improved results for both tasks [12,13], along with variations of CRFs and neural networks [5,14]. In recent years, deep neural networks are getting more and more visibility among machine learning methods, and with them, state-of-the-art results have been reported in various NLP tasks. For the SS task, recent studies make use of convolutional architectures [3], which can learn new representations from the input, and recurrent architectures that can easily model a sequence of words and the dependencies between them [9,18]. Analogously, deep neural models are

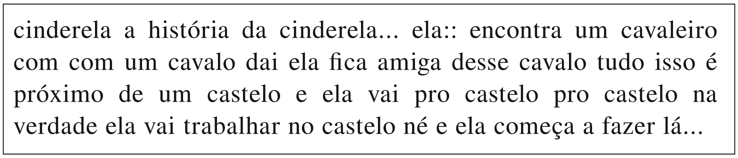
[1] http://nilc.icmc.usp.br/coh-metrix-dementia/.

being widely used in conjunction with decoding mechanisms to detect disfluency regions, with a greater focus on recurrent architectures [9,20].

The method proposed in this dissertation uses both sources of information (textual and prosodic) resulting a recurrent convolutional neural network model [19]. It was evaluated in the SS and DD tasks for the scenario of spontaneous and impaired speech.

3 Cinderella Dataset

The Cinderella dataset consists of spontaneous speech narratives produced in a test to elicit narrative discourse with visual stimuli, using a book of sequenced pictures based on the Cinderella story. In the test, an individual verbally tells the story to the examiner based on the pictures. Then, the narrative is manually transcribed by a trained researcher. Moreover, the narratives are composed of statements that may be impaired due to the inherent nature of patients (advanced age and with cognitive problems). The narratives produced contain many disfluencies. Figure 1 shows a transcript from the Cinderella production task that does not include either capitalization or sentence boundaries, besides presenting disfluencies.



cinderela a história da cinderela... ela:: encontra um cavaleiro com com um cavalo dai ela fica amiga desse cavalo tudo isso é próximo de um castelo e ela vai pro castelo pro castelo na verdade ela vai trabalhar no castelo né e ela começa a fazer lá...

Fig. 1. Narrative excerpt transcribed using the NURC annotation manual (http://www.letras.ufrj.br/nurc-rj/.)

This dataset consists of 60 narrative texts from BP speakers, 20 controls (CTL), 20 with AD, and 20 with MCI, diagnosed at the Medical School of University of São Paulo. This dataset was also used in [1]. This dataset has duration of 4 h and 11 m, an average of 34.4 sentences per narrative, and sentence averages of 11.6 words, counting all patient groups. As for disfluencies, this dataset contains 545 fillers divided into filled pauses, discourse markers and explicit editing terms, and 1778 edit disfluencies categorized into repetition, revisions and restarts.

For SS, using a manual to guide annotations, the narratives were annotated with sentence boundaries by three annotators using lexical, syntactic and prosodic clues. This annotation had an agreement of 0.89 for all patient groups using the Kappa coefficient [2]. For the disfluency annotation process, a well-defined categorization was explained with examples in an annotation manual. Two annotators annotated fillers according to their types, resulting in Kappa values of 0.83 for CTL and 0.84 for MCI patients. With regards to edit disfluencies, a sole annotator carried out the annotation. AD narratives were only used for training the lexical model.

4 Model Description

To automatically extract features from the input and also deal with the problem of long dependencies between words, we propose a model based on recurrent convolutional neural networks (RCNN), which was inspired by the works of [3, 18]. Using this combination of convolutional and recurrent layers, we explored the principle that nearby words have a greater influence in the classification, while distant words may also have some impact. The architecture of our model can be seen in Fig. 2.

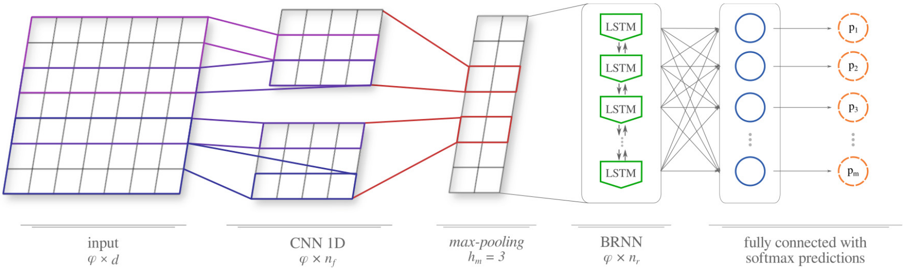


Fig. 2. The architecture is composed by an input layer that has φ input features, and each feature has a dimensionality of d. The convolutional layer is responsible for the automatic extraction of n_f new features depending on 3 neighboring words. Then, a max-pooling operation is applied over time, looking at a region of $h_m = 3$ elements to find the most significant features. The new extracted features are fed into a recurrent bidirectional layer which has n_f units known as Long Short-Term Memory [8], which are able to learn over long dependencies between words. Finally, the last recurrent state output is passed to a fully connected layer, where the softmax operation is calculated, giving us the probability of whether or not the word precedes a boundary or it is part of a disfluency region.

Our final model consists of a combination of two models. The first model deals with lexical information (part-of-speech tags and word embeddings), while the second treats only prosodic information (duration, pitch, energy and pause). Both models have the same architecture as shown in Fig. 2. This strategy is based on the idea that we can train the lexical model with much more data, since textual information is easily found on the web. In order to obtain the most probable class y_k for a word w, a linear combination was created between these two models, where one receives the weighted complement of the other: $P(y_k \,|\, w) = \alpha \cdot P_{lexical}(y_k \,|\, w) + (1 - \alpha) \cdot P_{prosodic}(y_k \,|\, w)$.

Since SS and DD are unbalanced classification problems ($\#NB \gg \#B$), we give different weights for each class in the cost function, where the weight of the minority class (B) is greater than that of the majority (NB). Lastly, we minimize the loss function with respect to all weights by using RMSProp algorithm with backpropagation [17].

5 Evaluations

5.1 Sentence Segmentation

By evaluating different word embedding settings, we found that the embeddings with 600 dimensions induced by Word2vec are more efficient for our RCNN. By adding prosodic information in conjunction with these embeddings and morphosyntactic tags, we found that combining all this information usually results in better results in terms of F_1. We justified the choice of a model based on recurrent neural networks by analyzing the performance of different classification methods based on neural networks, in which our RCNN obtained the best results in the Cinderella Corpus: $F_1 = 0.77$ for CTL and $F_1 = 0.74$ for MCI patients, exceeding the CRF method proposed by [6] by a large margin ($13\% - 15\%$).

We also found that our model achieved good results when tested with narratives whose story is different from the story used for training the narratives. Finally, based on an error analysis, we showed that our RCNN was able to learn lexical, syntactic and semantic evidences.

5.2 Filler Detection

We evaluated our RCNN model for filler detection (filled pauses and discourse markers) with the same approach used for SS. In the first experiments, we found that filled pauses are strongly related to the identities of words. With the evaluation of word embeddings we found that the best technique was FastText with 600 dimensions. By varying the set of features, we noticed that prosodic information has a low impact on classification, and that the textual features of embeddings and part-of-speech tags are, respectively, more impacting in the detection of both types of fillers. However, the best results were obtained when all the features were used together. Our RCNN obtained the best results in the detection of discursive markers for CTL and MCI, and tied with the CRF to detect filled pauses for MCI.

Based on an error analysis, we showed that our RCNN is strongly based on the identity of the words for the detection of both type of fillers. Finally, we showed that the best choice for filler detection is to combine a list of predetermined words for the detection of filled pauses and our RCNN model for the detection of discursive markers. In pursuit of higher performance, we found that it is possible to train another model to decide the case of the filled pause "é", since it is very ambiguous and occurs frequently in impaired speech. And using this strategy, we obtained an average F_1 between 0.91 and 0.92 for filler detection for both CTL and MCI transcripts.

5.3 Edit Disfluencies Detection

With regards to the detection of edit disfluencies, based on a thorough literature review we made two changes in the RCNN: (i) we added new linguistic features; (ii) we included a linear-chain CRF model with the Viterbi algorithm in the last layer of our RCNN (thus forming the RCNN-CRF model).

With the baselines, we verified that lexical information is quite impressive for detection of repetitions. With the evaluation of word embeddings we found that the best technique was Wang2vec with 600 dimensions. But still, the results using only word embeddings fell below the baselines. We observed that the best results were obtained using only the new linguistic features and morphosyntactic tags simultaneously, surpassing the baseline by a large margin. Moreover, our results showed that prosodic and word embeddings do not contribute to the detection of repetitions and revisions. Similarly to the previous tasks, we analyze the performance of different classification methods based on neural networks. With this analysis it was evident that the introduction of the CRF model in our original RCNN model was a good decision, since the RCNN-CRF obtained the best results in the detection of repetitions and revisions for all the groups of patients. In the experiments, we found that the results varied significantly for each type of disfluency: repetitions were detected with $F_1 = 0.85 - 0.89$, revisions were in the range of $F_1 = 0.35 - 0.37$, while restarts were in the range of $F_1 = 0.05$, much lower than the others.

We analyzed the main hits and errors of our RCNN-CRF model for each type of edit disfluency, and we found that repetitions are usually identified when analyzing duplicated word sequences, mainly for short-words repetitions, and that the main errors are due to prefixes or mistakes with revisions. For revisions we verified that the main hits were for sizes greater than 1, and that generally these hits have a lexical structure in which the first word of the original statement is equal to the first word of the correction, and the latter usually reflect a change of verb or noun. The main errors were due to the ambiguity of the phenomenon or because the lexical tips were very distant, indicating a possible increase in the size of the convolution filter to 9 or 11 (we used 7 in our experiments) in the convolutional layer or an increase in the number of neighboring words which should be considered in the manual features. Restarts were clearly the category in which our RCNN-CRF achieved the worst results, then, by disregarding the classification of them and only rephrasing the classification as a binary task (whether it is part of a disfluency region or not), our RCNN-CRF was able to achieve a F_1 of 0.70 for CTL and 0.75 for MCI.

5.4 Extrinsic

After analyzing the best models for each task in the intrinsic evaluations, we built a pipeline to automatically segment sentences and remove disfluencies from transcripts. The full pipeline can be seen in Fig. 3. The pipeline consists of:

1. Using the RCNN to segment the original transcript into sentences;

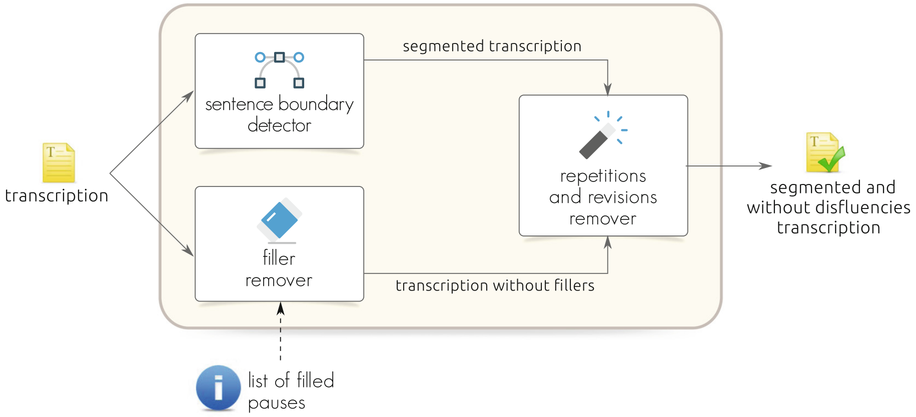

transcription
sentence boundary detector
filler remover
segmented transcription
transcription without fillers
repetitions and revisions remover
segmented and without disfluencies transcription
list of filled pauses

Fig. 3. Full pipeline to automatically segment sentences and to remove disfluencies from transcripts.

2. Combining the word list method to remove filled pauses and the RCNN to remove discursive markers and the filled pause "é";
3. Passing the segmented transcript without filled pauses and discursive markers to the RCNN-CRF, which then removes repetition and revisions.

We evaluated whether this pipeline influences positively or negatively the calculation of Coh-Metrix-Dementia's syntactic metrics. For this, we selected 9 syntactic metrics to verify if there was a statistically significant difference between manual and automatically generated transcripts. The metrics extracted depend on the success of parsing the syntactic tree as a whole; and also depend on the correct identification of verb phrases and noun phrases. The results are shown in Table 1. Such comparisons were analyzed using the non-parametric Wilcoxon rank-sum test, with a significance level of 5% ($p < 0.05$). The null hypothesis is that the metrics have equal averages for manual and automatic transcriptions.

With a significance level of 5% ($p < 0.05$), we show that the results obtained for CTL transcripts have no significant difference in relation to the manual transcriptions. And, as expected, there was a slight degradation of method performance as we moved from a more prepared speech (CTL) to a more impaired speech (MCI), more specifically, three metrics showed a significant difference for MCI.

The words per sentence metric probably was strongly impacted by the automatic tasks as our segmenter tends to put more boundaries than manual annotation does. In addition, our disfluency removal is more conservative and tends to remove fewer words than necessary in the Cinderella dataset. These two phenomena caused the number of words per sentences in automatic transcriptions to be very different from those of manual annotation. The same behavior also affected the metrics mean clauses per sentence and dependency distance, which

Table 1. Extrinsic evaluation on Coh-Metrix-Dementia. Values shown are mean (standard deviation) and p-value. Bold values denote statistical significance at the $p < 0.05$.

Metric	CTL			MCI		
	Auto	Manual	p	Auto	Manual	p
Yngve complexity	2.06 (0.10)	2.11 (0.16)	0.29	2.06 (0.14)	2.15 (0.15)	0.13
Frazier complexity	7.14 (0.24)	7.20 (0.29)	0.68	7.07 (0.30)	7.25 (0.27)	0.10
Mean clauses per sentence	1.88 (0.26)	2.06 (0.37)	0.23	**1.86 (0.28)**	**2.22 (0.41)**	**0.01**
Noun phrase incidence	297.30 (74.18)	296.12 (76.64)	0.84	273.78 (93.44)	311.68 (29.24)	0.19
Modifiers per noun phrase	0.39 (0.07)	0.40 (0.06)	0.95	0.40 (0.08)	0.40 (0.07)	0.87
Dependency distance	34.01 (7.22)	40.26 (11.68)	0.16	**34.29 (7.11)**	**48.01 (14.58)**	**0.00**
Pronouns per noun phrase	0.25 (0.08)	0.23 (0.07)	0.59	0.23 (0.07)	0.23 (0.07)	1.00
Words per sentence	11.25 (1.68)	12.53 (2.55)	0.17	**11.47 (2.09)**	**14.05 (3.04)**	**0.01**
Number of sentences	32.70 (15.57)	30.00 (14.29)	0.47	30.90 (14.92)	26.90 (14.10)	0.24

are also calculated according to the number of sentences in a text. However, this behaviour did not affect syntactic structure, since the values of syntactic metrics Yngve and Frazier complexity did not present statistically significant differences.

6 Conclusions

For SS, we obtained $F_1 = 0.77$ in CTL transcripts and $F_1 = 0.74$ in MCI, achieving the state-of-the-art for this task on impaired speech. For the filler detection task, we obtained, on average, $F_1 = 0.90$ for CTL and $F_1 = 0.92$ for MCI, results that are similar to related works of the English language. When restarts were ignored in the detection of edit disfluencies, $F_1 = 0.70$ was obtained for CTL and $F_1 = 0.75$ for MCI.

In the extrinsic evaluation, only 3 metrics showed a statistically significant difference between the manual transcripts and those generated by our method for MCIs, suggesting that, despite differences in sentence boundaries and disfluency removal, our method is able to generate transcripts to be automatically processed by NLP tools. Our method is publicly available on https://github.com/mtreviso/deepbondd, and the Web interface is available on http://fw.nilc.icmc.usp.br:23680/.

As for future work, we plan to evaluate our method with the output of an automatic speech recognition system for BP, as a high word recognition error rate can greatly affect our results.

Acknowledgments. We thank CNPq for a scholarship granted to the first author.

References

1. Aluísio, S., Cunha, A., Scarton, C.: Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in Portuguese. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 109–114. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_10
2. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. **22**, 249–254 (1996)
3. Che, X., Wang, C., Yang, H., Meinel, C.: Punctuation prediction for unsegmented transcript based on word vector. In: LREC, pp. 654–658 (2016)
4. Chen, J.C.: Speech recognition with automatic punctuation. In: EUROSPEECH, pp. 6–9 (1999)
5. Christensen, H., Gotoh, Y., Renals, S.: Punctuation annotation using statistical prosody models. In: ISCA Tutorial and Research (2006)
6. Fraser, K.C., Ben-david, N., Hirst, G., Graham, N.L., Rochon, E.: Sentence segmentation of aphasic speech. In: NAACL, pp. 862–871 (2015)
7. Heeman, P., Allen, J.: Detecting and correcting speech repairs. In: ACL, pp. 1–8 (1994)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
9. Hough, J., Schlangen, D.: Joint, incremental disfluency detection and utterance segmentation from speech. In: EACL, pp. 326–336 (2017)
10. Jarrold, W.L., Peintner, B., Yeh, E., Krasnow, R., Javitz, H.S., Swan, G.E.: Language analytics for assessing brain health: cognitive impairment, depression and pre-symptomatic alzheimer's disease. In: Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., Huang, J. (eds.) BI 2010. LNCS (LNAI), vol. 6334, pp. 299–307. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15314-3_28
11. Lehr, M., Prud'hommeaux, E.T., Shafran, I., Roark, B.: Fully automated neuropsychological assessment for detecting mild cognitive impairment. In: INTERSPEECH, pp. 1039–1042 (2012)
12. Liu, Y., Shriberg, E., Stolcke, A., Harper, M.P.: Comparing HMM, maximum entropy, and conditional random fields for disfluency detection. In: INTERSPEECH, pp. 3313–3316 (2005)
13. Liu, Y., Stolcke, A., Shriberg, E., Harper, M.: Using conditional random fields for sentence boundary detection in speech. In: ACL, pp. 451–458 (2005)
14. Qian, X., Liu, Y.: Disfluency detection using multi-step stacked learning. In: ACL, pp. 820–825 (2013)
15. Shriberg, E., Bates, R.A., Stolcke, A.: A prosody only decision-tree model for disfluency detection. In: Eurospeech, pp. 2383–2386 (1997)
16. Stolcke, A., et al.: Automatic detection of sentence boundaries and disfluencies based on recognized words. In: ICSLP (1998)
17. Tieleman, T., Hinton, G.: RMSprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw. Mach. Learn. (2012)
18. Tilk, O., Alumäe, T.: LSTM for punctuation restoration in speech transcripts. In: INTERSPEECH, pp. 683–687. ISCA (2015)

19. Treviso, M.V., Shulby, C., Aluísio, S.M.: Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. In: EACL, pp. 1–10 (2017)
20. Wang, S., Che, W., Zhang, Y., Zhang, M., Liu, T.: Transition-based disfluency detection using LSTMs. EMNLP, pp. 2775–2784 (2017)

Imitating Broadcast News Style: Commonalities and Differences Between French and Brazilian Professionals

Plinio A. Barbosa[1](✉) and Philippe Boula de Mareüil[2]

[1] University of Campinas, Campinas, Brazil
pabarbosa.unicampbr@gmail.com
[2] LIMSI, Orsay, France
Philippe.Boula.de.Mareuil@limsi.fr

Abstract. In this paper we asked four Brazilian and four French news announcers, both females and males, to read a text in their native language three times: in a neutral way, imitating broadcast news from memory and imitating the style of another female news announcer after listening to her talking about a different topic. From these 24 recordings, we analysed the proportion of initial stress in accentual phrases, three fundamental frequency (F0) statistical descriptors (median, standard deviation, range), reading and pause duration, as well as spectral emphasis, a correlate of vocal effort. Results pointed out that the main characteristics of the imitated broadcaster style in comparison with the neutral reading are: a higher proportion of initial stress, an increase of at least 2 semitones in F0 median, an increase in F0 range (for Brazilian news announcers, especially) and F0 standard deviation (for French news announcers, especially). In all cases, a higher value for spectral emphasis was found.

Keywords: News announcer style · Prosody · Imitation

1 Introduction

Since speech imitation seems to retain only the most salient characteristics of a speaker or the phonological representation of the sentence [1], it contributes to a better understanding of a speaking style and its perception. Caricature is also involved in this process, because imitation often exaggerates the actual patterns of the imitated style, in particular. Thus, caricature can also allow us to grasp the main features of a particular speaking style. Different strategies for imitation can be used including imitation from inner representations, where caricature often emerge, and consecutive imitation, where mechanisms of convergence and accommodation occur [2–4].

The choice for broadcast news style in this imitation-based study is related to the fact that, in both French and Brazilian Portuguese (henceforth BP), the pronunciation norm is represented by public speech professionals (the radio and

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 419–428, 2018.
https://doi.org/10.1007/978-3-319-99722-3_42

TV, especially) and that we are interested in studying the employment of this norm in imitation tasks that could be easily carried out by non-professional speakers in further research. In the case of France, this norm more or less corresponds to the Paris pronunciation [5], whereas for BP, announcers adopt a kind of trade-off between the pronunciations of coda /R/ from Rio and coda /S/ from São Paulo [6,7].

A tendency towards initial prominence has been observed in the French news announcer style [5,8–14] and is also found in BP broadcasting [16], in words with a strong semantic load especially (e.g., *BIlhões de reais*, BIllions of Reais).

A previous study [17] with a Brazilian AM-radio broadcaster analysed melodic and temporal parameters for five reading tasks. Two of them of interest for this work: the neutral reading of 29 headlines accompanied by another reading as if the professional broadcaster were performing in the studio. Results of acoustical analyses revealed that the main changes concerned F0 median (increase of 12% for the professional reading) and the rate of pitch accents (slightly higher in the neutral reading task).

Since the knowledge of the acoustic and linguistic variables involved in professional speech and its imitation is still fragmented, the goals of the present work are twofold: (1) to compare imitations of journalist speech in two languages/cultures, and (2) to examine the acoustic parameters involved in two types of imitation (imitation from memory and consecutive imitation) by professionals.

Based on general perceptual impressions and the aforementioned research on the prevalence of initial prominence in broadcast news style and acoustic convergence in dialogues, our main hypotheses involving imitation of broadcast news are: (1) an increase of fundamental frequency (F0) median and F0 range of variation; (2) an increase in the proportion of initial prominence; (3) a convergence of acoustic parameters (speech rate, especially); (4) no clear differences across languages are hypothesised.

2 Methodology

2.1 Corpus

The corpus is composed of the reading of the text "The North Wind and the Sun" ("La bise et Le soleil" in French, "O vento sul e o sol" in BP) in three ways by journalists, recorded in France and Brazil: (1) in a neutral way, (2) in the journalist style of their country from memory, using their own representations, and (3) in the style of a TV female broadcaster of either country, just after listening to her talking about another topic. The text is an adaptation of an Aesop fable which has been used for over a century by the International Phonetic Association for comparing languages' phonetic systems [15]. Even if it would be easier to our subjects to use actual transcripts of read news, we preferred to use a more general text for comparison purposes with non-professional speakers in already ongoing research. Two female and two male speakers in each country participated in the experiment. All the French professionals lived in Paris and those from Brazil lived in the state of São Paulo. For accomplishing the third task, the French news announcers listened to an excerpt of circa 1 min from Johanne Portal (BFMTV), whereas the Brazilian news announcers listened to an excerpt of over 1 min. from Salete Lemos (Rede TV News), both commenting on previous news. Portal has a median F0 of 240 Hz (95 st ref 1 Hz) and Lemos has a median F0 of 200 Hz (92 st ref 1 Hz).

Accentual phrases (AP) of at least three syllables were analysed with particular attention due to our interest in initial stress realisation. This also provides more compatible comparisons across the two languages, because parallel constituents can be found if necessary. There were 36 APs of at least 3 syllables in BP and 39 in French: they are shown between square brackets below, in BP and French. Here accentual phrases are understood as phonological words. Their delimitation corresponds to the APs produced by at least 3 out of 4 speakers in each language.

[O vento] sul e o sol [discutiam] qual dos dois era [o mais forte], quando passou [um viajante] [envolto] [num casaco]. [Ao vê-lo], [apostaram] que [aquele] que [primeiro] [conseguisse] [obrigar] [o viajante] [a tirar] [o casaco] [seria] [considerado] [o mais forte]. [O vento] sul [começou] [a soprar] [com muita força], mas quanto [mais soprava], [mais o viajante] [se embrulhava] [no seu casaco], [até que] [o vento] sul [desistiu]. O sol brilhou então [com toda intensidade], e [imediatamente] [o viajante] tirou [o casaco]. [O vento] sul teve assim [de reconhecer] [a superioridade] do sol.

La bise et [le soleil] [se disputaient], chacun [assurant] [qu'il était] [le plus fort], [quand ils ont vu] [un voyageur] [qui s'avançait], [enveloppé] [dans son manteau]. [Ils sont tombés] d'accord [que celui] [qui arriverait] [le premier] [à faire ôter] [son manteau] [au voyageur] serait [regardé] [comme le plus fort]. Alors, la bise s'est mise [à souffler] [de toute sa force] mais [plus elle soufflait], [plus le voyageur] serrait [son manteau] [autour de lui] et [à la fin,] la bise [a renoncé] [à le lui faire ôter]. Alors [le soleil] a commencé] [à briller] et [au bout d'un moment], [le voyageur], [réchauffé], [a ôté] [son manteau]. Ainsi, la bise [a dû reconnaître] [que le soleil] était [le plus fort] des deux.

2.2 Measurements

The following measures were extracted from the 24 (3 styles × 2 languages × 4 subjects) readings: the first two sets of measurements were computed for each AP, whereas the latter two sets for the entire readings.

Each AP was annotated under Praat [18] according to whether or not the peak of the F0 contour was anchored in the nuclear stressed syllable: 0 if anchored in the nuclear syllable, 1 if anchored in the first syllable of the lexical (or phonological word), and 2 otherwise. For each AP, we computed the range (maximum − minimum) of the F0 contour in semitones (st). According to [19], 3 st may be considered as an appropriate threshold for predicting a prosodically prominent unit. A higher proportion of initial stress with F0 range greater than 3 st is expected to occur in the journalist style, as a correlate of new information.

For each AP, we computed the following F0 descriptors in semitones: F0 median, maximum of F0 estimated by quantile 99.5% to avoid errors in F0 tracking, range of F0, computed from the difference between quantiles 99.5% and 0.5%, absolute and relative F0 standard-deviation, absolute and relative F0 semi-amplitude between quartiles. For relative measurements, the ratio in reference to F0 median was computed.

Total reading duration was computed for each style defined by the interval between the onset of the first vowel and the offset of the last vowel of each reading. This serves as a measure of speech rate, because the text is the same for the three readings for each speaker. In addition, total silent pause duration was measured.

Finally, spectral emphasis was measured: it is a measure of the concentration of energy in high frequencies, defined by [20] as $E - E_0$ where E is the energy in dB up to the Nyquist frequency and E_0 is the energy of the signal low-band filtered up to 400 Hz. The authors showed that this measure is a correlate of vocal effort. It was computed on the basis of the speech signal corresponding to each reading.

3 Statistical Analyses and Results

For statistical analyses, due to the non-normality of residuals, we used the Scheirer Ray Hare (SHR) non-parametric equivalent of a 2-Way ANOVA with the following factors: SPEAKER (four levels per language) and STYLE (three levels). When necessary, the Wilcoxon post hoc test was used with the Bonferroni correction for the level of significance. In all models, the level of significance adopted was 1%. Boxplots are shown by grouping speakers according to gender, for the sake of visibility.

Only measures of F0 median, standard-deviation and range were significantly different for STYLE with some differences across speakers in the two languages in some cases discussed below. Reading and silent pause duration as well as spectral emphasis were compared descriptively and comments on the results for these parameters will be made after the presentation of the results for initial stress for each language.

3.1 BP Speakers

In BP, Table 1 shows increases of initial stress proportion for all speakers in the consecutive imitation task in comparison with the neutral reading, especially for F0 ranges greater than 3 st (figures in parentheses) with increases between 8 and 28%. In the imitation from memory, only speaker CL significantly increases the proportion of initial stress with more than 3 st of range. For all speakers, the initially stressed APs with an F0 range greater than 3 st have a proportion between 95 and 100% for the two imitation tasks.

As for reading duration, all BP speakers slowed down their total reading duration in the imitation tasks, in comparison with the neutral reading: from

Table 1. Proportion of accentual phrases labeled as initially stressed (AP1) in BP. In parentheses, proportions where F0 range is greater than 3 st are tabulated. Last column is the average for the four speakers.

Task	LC	JL	PL	CL	All
Neutral	42 (36)	53 (36)	63 (50)	69 (50)	57 (43)
Imitation/memory	42 (39)	43 (36)	61 (50)	78 (75)	56 (50)
Imitation/consecutive	53 (44)	56 (50)	86 (78)	72 (58)	67 (58)

3 to 10% in the imitation from memory (except speaker LC, who sped up the reading), and from 6 to 31% in consecutive imitation. Total pause duration ranges between 10 and 21% for all readings; only male speakers LC and CL exhibit higher amounts of pause duration in imitation (from circa 16 to 20%). In BP, imitations are 1 to 2 dB higher in both imitations for male speakers; they are 1 to 3 dB higher (in imitation from memory) and 3 to 7 dB higher(in consecutive imitation) for female speakers.

As for F0 median, Fig. 1 shows the boxplots related to an overall SHR test with $H_{3,402} = 267.0$ for SPEAKER, $H_{2,402} = 35.6$ for STYLE and no significant interaction. This parameter is not significantly different between neutral reading and imitation from memory, and increases by 2 st in consecutive imitation.

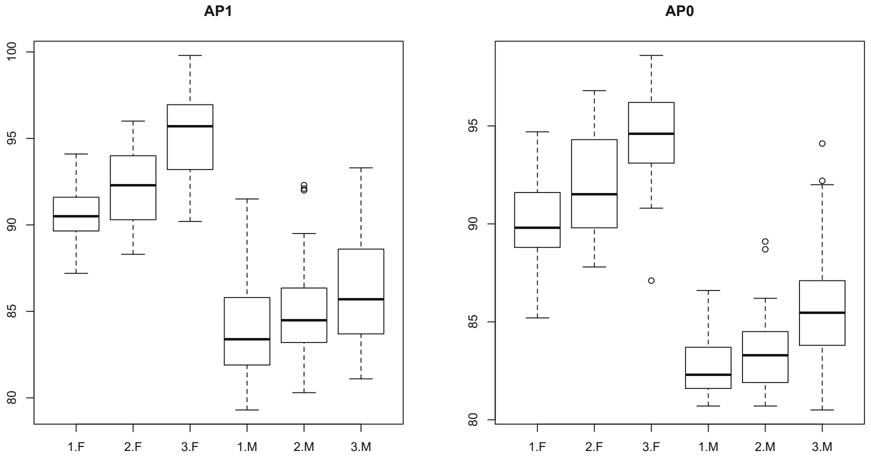


Fig. 1. F0 median in semitones with a 1 Hz reference for (1) neutral reading, (2) imitation by memory and (3) consecutive imitation for male (LC, CL) and female (PL, JL) BP speakers in accentual phrases labeled as initially stressed (AP1) or not (AP0).

In the case of F0 range (Fig. 2), there is a significant difference for STYLE in accentual phrases labeled as AP1 ($H_{2,218} = 10.9$) and for SPEAKER in accentual phrases labeled as AP0 ($H_{3,172} = 8.2$) with mean values of 8 st in the neutral reading task and 10 st in the imitation tasks in accentual phrases labeled as AP1. By contrast, in accentual phrases labeled as AP0, speaker CL exhibits a higher mean (8 st) than does JL (6 st).

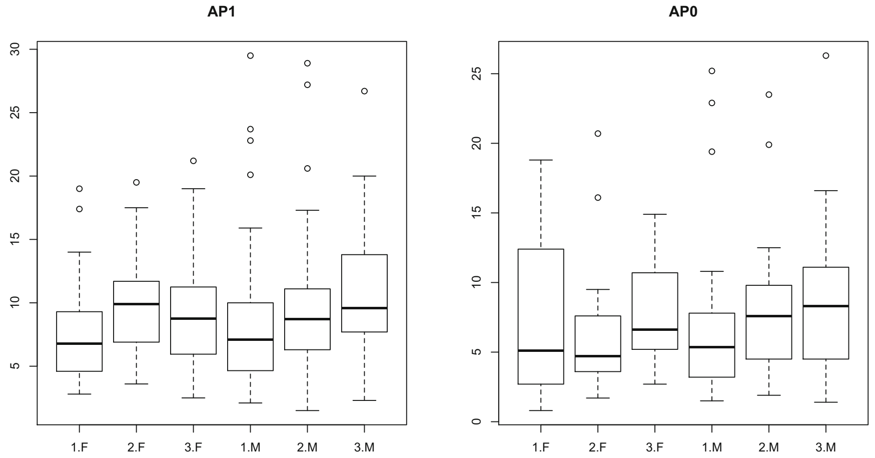


Fig. 2. F0 range in semitones in (1) neutral reading, (2) imitation from memory and (3) consecutive imitation for male (LC, CL) and female (PL, JL) BP speakers in accentual phrases labeled as initially stressed (AP1) or not (AP0).

3.2 French Speakers

In French, Table 2 shows increases of initial stress proportion for all speakers for the imitation tasks except speaker BH. For all speakers, the proportion of APs with initial stress and F0 ranges greater than 3 st is between 89 and 100% for the two imitation tasks. Figures are not available for speaker MC, who did not succeed in performing the consecutive imitation task.

Table 2. Proportion of accentual phrases labeled as initially stressed (AP1) in French. In parentheses, proportions where F0 range is greater than 3 st are tabulated. Last column is the average for the four speakers.

Task	BH	LP	AR	MC	All
Neutral	45 (39)	71 (63)	42 (32)	50 (47)	52 (45)
Imitation/memory	50 (39)	71 (68)	63 (58)	74 (71)	65 (59)
Imitation/consecutive	42 (36)	79 (79)	71 (71)	-	64 (62)

As for reading duration, when imitating the TV news announcer, speaker LP sped up his reading by 17%, whereas speakers BH and AR slowed down their readings between 10 and 32%, even though the BFMTV announcer speaks very fast. Spectral emphasis is 1 to 3 dB higher for both imitations in French, irrespective of speaker gender.

As for F0 median, Fig. 3 shows that, in both types of stress placement, male and female speakers tend to feature higher values in the imitation tasks. This tendency is significant for male speaker BH (overall SHR test with $H_{3,383} = 138.2$ for SPEAKER, $H_{2,383} = 41.9$ for STYLE and $H_{5,386} = 38.0$ for the interaction) who rises from a median of 86 st (in neutral reading) to 89 st (in both imitation tasks). Female speaker AR significantly changes her F0 median from 89 st (in both neutral reading and imitation from memory) to 97 st (in consecutive imitation), but no significant changes for the other speakers were found.

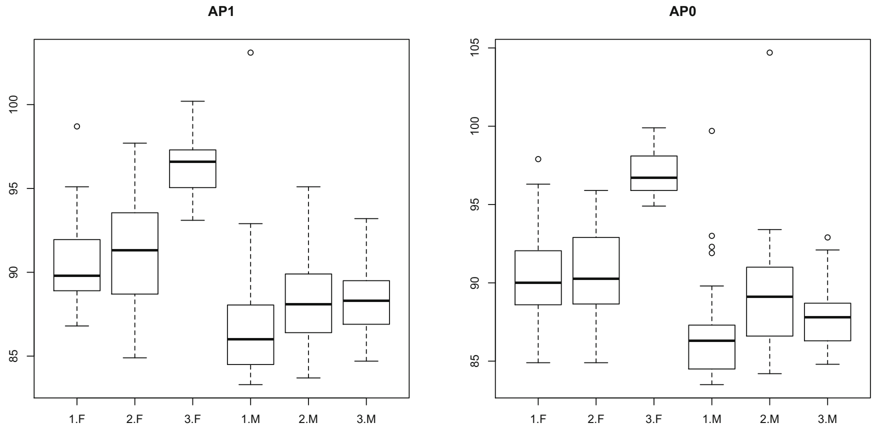


Fig. 3. F0 median in semitones with a 1 Hz reference for (1) neutral reading, (2) imitation from memory and (3) consecutive imitation for male (BH, LP) and female (AR, MC) French speakers, in accentual phrases labeled as initially stressed (AP1) or not (AP0).

F0 standard deviation and F0 range are significantly different only for female speaker AR, who changes F0 standard deviation from 2 st (in neutral reading) to 3 st (in both imitation tasks), and F0 range from 7 st (in neutral reading) to 10 st (in both imitation tasks): see Fig. 4. For F0 standard deviation, the overall SHR test is $H_{3,383} = 55.5$ for SPEAKER, $H_{2,383} = 17.2$ for STYLE and $H_{5,383} = 16.3$ for the interaction; for F0 range, the overall SHR test is $H_{3,383} = 65.4$ for SPEAKER, $H_{2,383} = 20.4$ for STYLE and $H_{5,383} = 19.6$ for the interaction. Speakers LP and MC did not significantly change their melodic parameters when imitating the broadcast news style, from memory or after having listened to the BFMTV news announcer.

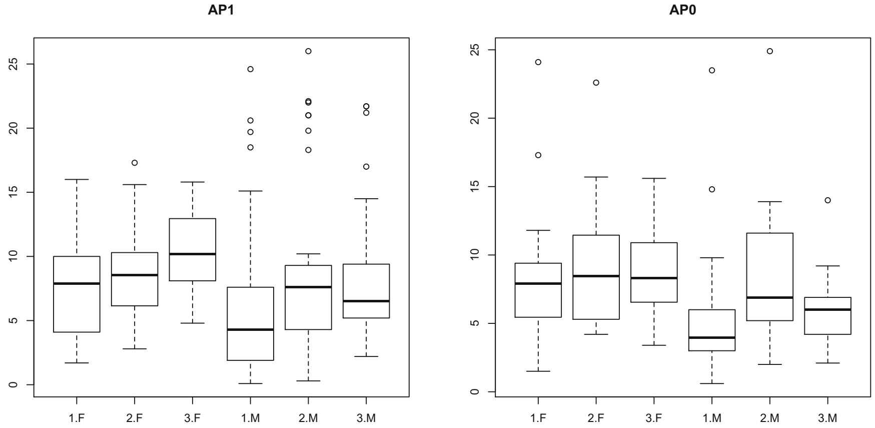


Fig. 4. F0 range in semitones in (1) neutral reading, (2) imitation from memory and (3) consecutive imitation for male (BH, LP) and female (AR, MC) French speakers in accentual phrases labeled as initially stressed(AP1) or not (AP0).

4 Discussion and Conclusions

The main findings of this study are the following: (1) an increase of initial stress proportion more expressive for French speakers when imitating from memory; (2) an increase of initial stress proportion in both BP and French in the case of consecutive imitation; (3) an increase of F0 median of at least 2 st, especially in the case of consecutive imitation; (4) an increase of F0 range of at least 2 st especially in BP and particularly in consecutive imitation; (5) an increase of F0 standard deviation in French; (6) a higher spectral emphasis in two imitations in the two languages, with overall higher values for female BP speakers; (7) inter-subject differences in the use of acoustic parameters.

These results suggest that initial prominence is an important feature of the broadcast news style in the languages under investigation. Also, changes in melodic parameters were observed in the imitation of this style, with a higher F0 and more extended pitch excursion. As for differences between the two languages, they seem to reside in a distinct behaviour facing consecutive imitation. Hypotheses (1) and (2) were confirmed, whereas hypotheses (3) and (4) were only partially confirmed. Regarding hypothesis (3), there is no or little convergence towards the target model in French speakers' consecutive imitation. Regarding hypothesis (4), differences were noticed between the two languages under study, which are inter-related to differences between subjects or genders.

The elasticity of intonation which this study has highlighted allows prosody to operate as a socioprofessional marker, indexical of a particular speaking style, which our imitators tried to reproduce. Their ability of doing so concerns both categorical (displacement of stress position) and non-categorical variables (F0 average and variation), which suggests a sensitivity to both phonetic detail and phonological representation. Our results support the idea that imitation is an interesting research paradigm which deserves to be developed. In fact, our speakers' behaviour confirms previous findings for broadcast news style.

Acknowledgements. We thank the FAPESP grant 2017/07546-0 for the second author and the CNPq grant 302657/2015-0 for the first author.

References

1. Cole, J., Shattuck-Hufnagel, S.: The phonology and phonetics of perceived prosody: What do listeners imitate? In: Proceedings of the Interspeech 2011, pp. 969–972, Florence, Italy (2011)
2. Zetterholm, E.: The same but different - three impersonators imitate the same target voices. In: 15th International Congress of Phonetic Sciences, Barcelona, pp. 2205–2208 (2003)
3. Delvaux, V., Soquet, A.: The influence of ambient speech on adult speech productions through unintentional imitation. Phonetica **64**(2–3), 145–173 (2007)
4. Buder, E.H., Eriksson, A.: Prosodic cycles and interpersonal synchrony in American English and Swedish. In: Fifth European Conference on Speech Communication and Technology, Rhodes, pp. 235–238 (1997)
5. Carton, F.: La prononciation. In: Antoine, G., Cerquiglini, S. (eds.) Histoire de la langue française 1945–2000, pp. 25–60. CNRS ditions, Paris (2000)
6. Evangelista, A.F., Almeida, T.A.R.: Assim fala a notícia: sotaques e regionalismos no telejornalismo paraibano. In: XVI Congresso de ciencias da comunicacao na regiao Nordeste, Joao Pessoa, pp. 1–15 (2014)
7. Silveira, R.C.P.: Uma pronúncia do português brasileiro. Cortez, São Paulo (2008)
8. Fónagy, I., Fónagy, J.: Prosodie professionnelle et changements prosodiques. Le français moderne **3**, 193–227 (1976)
9. Léon, P.: Précis de phonostylistique: Parole et expressivité. Fernand Nathan, Paris (1993)
10. Astésano, C.: Rythme et accentuation en français: Invariance et variabilité stylistique. L'Harmattan, Paris (2001)
11. Oakes, L.: Phonostylistique des annonceurs de la radio: Étude prosodique des textes radiophoniques. J. French Lang. Stud. **12**, 279–306 (2002)
12. Gendrot, C.: Aspects perceptifs, physiologiques et acoustiques de différentes catégories prosodiques en français. Ph.D. thesis, Université Sorbonne Nouvelle - Paris 3, Paris (2006)
13. Goldman, J.P., Auchlin, A., Simon, A.C.: Phonostylographe: un outil de description prosodique. comparaison du style radiophonique et lu. Nouveaux cahiers de linguistique française **28**, 219–237 (2007)
14. de Mareüil, B., Rilliard, A., Allauzen, A.: A diachronic study of initial stress and other prosodic features in the French news announcer style: corpus-based measurements and perceptual experiments. Lang. Speech **55**(2), 263–293 (2012)
15. Smith, C.: Review of handbook of the international phonetic association: a guide to the use of the international phonetic alphabet. Phonology **17**, 291–295 (2000)
16. Lopes, L.W.: Do texto ao contexto: a prosódia na construção da intencionalidade no relato de notícias. Tese de mestrado, Universidade Católica de Pernambuco, Recife, Brazil (2006)
17. Campos, L.C.P.: Radialista: análise acústica da variação entoacional na fala profissional e na fala coloquial. Master's thesis, University of Campinas, Campinas (2012)

18. Boersma, P., Weenink, D.: Praat: doing phonetics by computer, version 6.0.29 (2017). http://www.praat.org/
19. ’t Hart, J., Collier, R., Cohen, A.: A Perceptual Study of Intonation: An Experimental-phonetic Approach to Speech Melody. Cambridge University Press, Cambridge (1991)
20. Traunmuller, H., Eriksson, A.: Acoustic effects of variation in vocal effort by men, women and children. J. Acoust. Soc. Am. **107**(6), 3438–3451 (2000)

Automatic Detection of Prosodic Boundaries in Brazilian Portuguese Spontaneous Speech

Bárbara Teixeira[1](✉), Plínio Barbosa[2], and Tommaso Raso[1]

[1] Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
barbaraheloha@gmail.com
[2] University of Campinas, Campinas, São Paulo, Brazil

Abstract. This paper presents some models based on multiple phonetic-acoustic parameters for the automatic detection of prosodic boundaries in spontaneous speech. A sample with seven excerpts of monologic Brazilian Portuguese spontaneous speech was segmented into prosodic units by 14 trained annotators. The perceived prosodic boundaries were annotated as terminal or non-terminal prosodic boundaries. A Praat script was prepared in order to extract a set of acoustic parameters during the speech signal. Two statistical classifiers, namely *Random Forest* e *Linear Discriminant Analysis*, were used to generate models of subgroups of acoustic parameters that could work as predictors of prosodic boundaries in comparison with the human annotators. The initial evaluation of the classifiers showed that both present relative success in detecting boundaries. The LDA performed better in predicting boundaries and therefore its models were refined. The final model for terminal boundaries showed 80% of agreement with human annotators. As for non-terminal boundaries, three models were obtained. The sum of boundaries identified by the three models together corresponds to an agreement of 98% with the human annotators.

Keywords: Automatic detection · Prosodic boundaries · Spontaneous speech

1 Introduction

It is generally uncontroversial that speech is performed and perceived as small prosodic units, usually formed by a few words, marked by prosodic boundaries [1–4]. These units can be functionally analyzed according to different theoretical perspectives: syntactic [5, 6], pragmatic [7–9] and cognitive perspectives [10–12]. However, boundaries can be studied *per se*, independently of the theoretical perspective from which the units are observed [13].

A review of the literature points to the fact that the types of boundaries are associated with the perception of conclusion or continuation of the utterance [14–17]. However, so far, tools available for speech segmentation into units marked by prosodic boundaries are very uncommon.

One of such tools is the algorithm ANALOR [18] for French. This tool searches for pauses of at least 300 ms, syllable lengthenings, f0 variations, f0 *resets* and relevant prominences, and uses these features to segment speech into prosodic units.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 429–437, 2018.
https://doi.org/10.1007/978-3-319-99722-3_43

An important benefit of this tool is that it was planned to segment both read and spontaneous speech. Results show an agreement of 83.8% with human segmentation. However, the algorithm requires a previous syllabic segmentation.

The model proposed by Ni *et al.* in [19] detects terminal boundaries with high accuracy in Mandarin. For Mandarin, by using regression techniques with 30 acoustic measurements as predictors, the model achieves 85% of agreement with human annotation. For English, the model achieves 82% of agreement with 24 acoustic measurements as predictors. The final test for the model proposed by Kim in [20] shows a detection accuracy of 74% and 56% for terminal and non-terminal boundaries respectively.

With the exception of the model proposed by Kim in [20], these tools do not distinguish between the two main boundaries types, as it is not specified whether the detected boundaries correspond to terminal or non-terminal ones.

Differently from those models, in our work we investigate the acoustic features of prosodic boundaries in order to design a tool to perform an automatic (or at least semi-automatic) detection of different types of boundaries in monologic and spontaneous Brazilian Portuguese (BP) speech. Such a tool will be able to aid in the process of spoken corpora compilation by making the task of speech segmentation faster.

2 Methodology

In this section we present the data selected for the analysis, how they were treated and the statistical analyses.

2.1 Data

Our data consist of seven excerpts of monologic spontaneous speech from the three sections of the C-ORAL-BRASIL corpus [21, 22], namely the informal and the formal sections of the natural context and the media section. The excerpts have on average 191 words each. We chose to conduct this study using only the male monological speech because fundamental frequency differs a lot between men and women, and we wanted to exclude the gender variable.

2.2 Data Processing

Each excerpt was independently annotated by 14 trained members of the Empirical and Experimental Language Studies Laboratory (LEEL) of the Federal University of Minas Gerais. The annotators received the sound file and the orthographic transcription without any annotation. Their task was to annotate the two main types of boundaries following their perception using a simple slash symbol (/) to indicate a non-terminal boundary and a double slash one to indicate a terminal boundary.

All the annotators had some degree of experience in speech segmentation. They had been trained through a process that lasted approximately 4 months with the aim to prepare them for the segmentation of the C-ORAL-BRASIL corpora by raising awareness of the boundaries' perceptually relevant prosodic cues. For more details, see [23].

The agreement among the annotators was evaluated through the Fleiss kappa coefficient [24]. For the excerpts in the sample used in this work, the general interrater agreement was 0.80 for the annotation of terminal boundaries and 0.75 for non-terminal ones. The excerpts were annotated in 5 tiers using the Praat's TextGrid tool [25] as follows: (1) Segmentation in V-V units and transcription using a broad phonetic system in ASCII. (2) Annotation of the non-terminal prosodic boundaries, informing the number of annotators who detected them. (3) Annotation of the terminal prosodic boundaries, informing the number of annotators who detected them. (4) Annotation of the intervals referring to silent pauses. (5) Orthographic transcription of the utterances.

An extended version of the *ProsodyDescriptor* script [26] was developed in order to extract 111 acoustic-phonetic parameters from the speech signal. This extended version, named *BreakDescriptor* [27], extracts the acoustic-phonetic parameters for all the V-V units in a window centered in all the boundaries between phonological words, including therefore the positions perceived as prosodic boundaries by the annotators. The *BreakDescriptor* also extracts the 111 parameters for the positions perceived as non-boundaries and their respective windows.

The windows scanned by the *BreakDescriptor* include ten V-V syllables to the left and ten V-V syllables to the right of each analyzed V-V syllable (those at the boundary of phonological words). A position was used for the analysis of prosodic boundaries only if at least seven annotators perceived it as a boundary of the same type, that is, only if seven annotators perceived a non-terminal boundary or if at least seven annotators perceived a terminal one. The remaining positions between phonological words were treated by the script as non-boundary (Fig. 1).

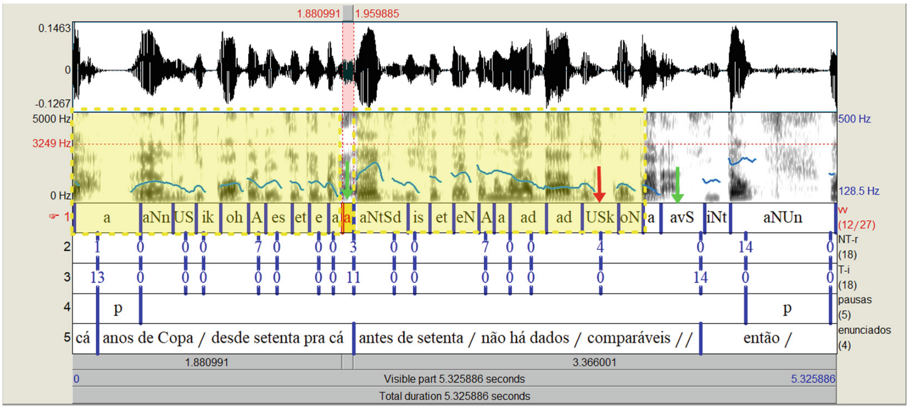


Fig. 1. Starting at the top: wave form, broad-band spectrogram, and the different tiers in a Praat TextGrid. The position of terminal boundary used here for the analysis is highlighted in yellow; it constitutes the central point of the analyzed window. Excerpt transcript: cá //anos de Copa /desde setenta pra cá //antes de setenta /não há dados comparáveis //então/ (Color figure online)

The physical properties extracted by the script include five groups of parameters: (a) Speech rate and rhythm measurements. (b) Normalized duration of the segments. (c) F0. (d) Intensity. (e) Silent pauses durations.

2.3 Statistical Analysis

The measurements automatically extracted were subjected to two methods for statistical classification, *Random Forest* (RF) and *Linear Discriminant Analysis* (LDA), in order to identify the combination of measurements that could better explain the perceptual segmentation of humans. In both the classifiers, the presence and the absence of boundary, terminal and non-terminal boundaries, were considered. We also tried to detect the boundary presence in two steps, first identifying the presence of boundary and then the type of boundary, but the performance of the classifier does not improve. The reason for this is the use of a more complex and ambiguous data matrix. The predictive power of both models was also taken into account. The training set consisted of a random selection of 70% of the V-V units in our data, whereas the test set consisted of the remaining 30% of the V-V units.

Three main steps for the statistical analysis were taken: (a) Initial evaluation of the classifiers. (b) Refinement of the LDA classifier. (c) Refinement of the model designed to detect non-terminal boundaries. All the statistical analyses were performed using R [28].

The initial evaluation aimed at estimating the best classifier for the task. Since LDA showed better a performance in identifying the two types of boundaries, its results were submitted to different refinement processes.

These processes were aimed at improving the performance of the classifier by looking for a more accurate prediction of the annotators' behavior. The 111 parameters, used as predictors, had their number reduced following two different heuristics. Firstly, we gradually eliminated the parameters ranked as less relevant in the hierarchy of the classification, following the weight assigned by the model. Secondly, we reintroduced or eliminated some parameters based on the findings available in the literature and not only on the weight attributed by the classifier.

For non-terminal boundaries a third step was needed, since the two first phases did not yield a satisfying result. In order to have a better prediction of non-terminal boundaries, we looked for models that could explain sub-groups of boundaries characterized by different configurations of parameters. After the best result obtained with the first model, we eliminated the boundaries that this model detected and created a new model for the remaining boundaries. We reached a very high result with three different models, each one explaining a subset of boundaries.

In order to better investigate the subgroups of non-terminal boundaries and to enhance the automatic classification, the measurements included in the final models were also subjected to an analysis of hierarchical cluster.

3 Results

3.1 Model for the Classification of Terminal Boundaries

The predictive power of the final model for the automatic detection of terminal boundaries reached 80% of agreement with the boundaries identified by the human annotators in the sample. Table 1 shows the relevant parameters to predict the presence of terminal boundaries.

Table 1. Acoustic-phonetic parameters of the model for recognition of terminal boundaries

Rank	Abbreviation weight	Parameters	Rank	Abbreviation weight	Measurements
1st	psdur 2.641	Pause presence after V-V unit	11th	df0meddloc 0.032	First derivative of F0 median: difference between 1st V-V unit on right window and last V-V- unit on left window
2nd	psp 1.948	Pause duration after V-V unit	12th	f0medd 0.029	Mean of F0 medians: difference between right and left windows
3rd	f0meddloc 0.329	First derivative of F0 median: difference between V-V at boundary and first V-V to its right	13th	zl10 0.028	Mean of smoothed z-score for 1st V-V unit on the left window
4th	df0medr1 0.264	Mean of F0 median first derivative on the left windows	14th	skf0d 0.025	Skewness of F0 medians: difference between right and left windows
5th	df0medl 0.257	First derivative of F0 median for 1st V-V unit on right window	15th	mzd 0.015	Mean of smoothed z-score: difference between right and left windows
6th	sddf0d 0.157	First derivative of F0 median: difference between right and left V-V unit	16th	skdf0d 0.011	Skewness of F0 first derivative medians: difference between right and left windows
7th	prd 0.101	Peak rate of smoothed z-score: difference between right and left windows	17th	SDzl 0.010	Standard deviation of smoothed z-score: difference between V-V units on left window
8th	sdf0l 0.091	Standard deviation of F0 medians on left window	18th	ard 0.003	Rate of non-salient V-V units per second: difference between right and left windows
9th	df0medl10 0.066	First derivative of F0 median for 1st V-V unit on left window	19th	zdloc 0.001	Mean of smoothed z-score: difference between 1st V-V unit on right window and V-V unit at window center
10th	f0rl 0.033	Peak rate of smoothed F0 peaks per second on the left windows	20th	emphl 0.001	Mean spectral emphasis for V-V unit at window center

3.2 Classification Models for Non-terminal Boundaries

In order to explain non-terminal boundaries, three different models were obtained. The total of boundaries identified by the three models corresponds to an agreement of 98% of the boundaries annotated by humans. Table 2 shows the acoustic-phonetic parameters included in each one of the three models.

Table 2. Acoustic-phonetic predictor parameters of the models for non-terminal boundaries

Model 1–9 parameters			Model 2–10 parameters		Model 3–8 parameters	
Rank	Abbrev. weight	Parameter measurements	Abbrev. weight	Parameter measurements	Abbrev. weight	Parameter measurements
1st	zl0 4.5	Mean of smoothed z-score of V-V unit at boundary point	srl 0.72	Rate of V-V units per second on the left window	prl 151.6	Peak rate of smoothed z-score on left window
2nd	zrl 4.4	Mean of smoothed z-score 1st right	sddf0l 0.63	Standard deviation of F0 median first derivative on left window	prd 150.6	Peak rate of smoothed z-score - difference between right and left window
3rd	zdloc 4.2	Rate of non-salient V-V units per second - difference between 1st right and left V-V units	sdf0l 0.47	Standard deviation of F0 medians on left window	prr 149.5	Peak rate of smoothed z-score on right window
4th	psp 2.6	Pause presence	ard(*) 0.45	Rate of non-salient V-V units per second - difference between right and left windows	sdf0r 0.5	Standard deviation of F0 medians on right window
5th	psdur 2.3	Pause duration	f0medl 0.37	Mean of F0 medians left window context	SDzl 0.3	Standard deviation of smoothed z-score on left window
6th	ard (*) 0.3	Rate of non-salient V-V units per second - difference between right and left windows	f0rd 0.21	Peak rate of smoothed F0 peaks per second difference of right and left windows	df0medr1 0.3	First derivative of F0 median for 1st V-V unit on the right
7th	srd 0.3	Rate of V-V units per second - difference between right and left windows	f0meddloc 0.10	F0 median - difference between last V-V unit on the left window and first unit on the right	df0medl10 0.2	First derivative of F0 median for 1st V-V unit on the left
8th	sdf0d 0.2	Standard deviation of F0 medians - difference between right and left windows	f0med0 0.09	F0 median of V-V unit at boundary point	df0meddloc 0.1	F0 median - difference between last V-V unit on the left window and first unit on the right
9th	zl10 0.2	Mean of smoothed z-score for 1st V-V unit on the left window	f0medr1 0.05	F0 median of V-V unit at 1st V-V unit on the right		
10th			emphl 0.01	Mean spectral emphasis on the left window		

The analysis of hierarchical cluster showed that the model can be divided in subgroups marked by different configurations of parameters and with different weights. Figures 2 and 3 below shows the results.

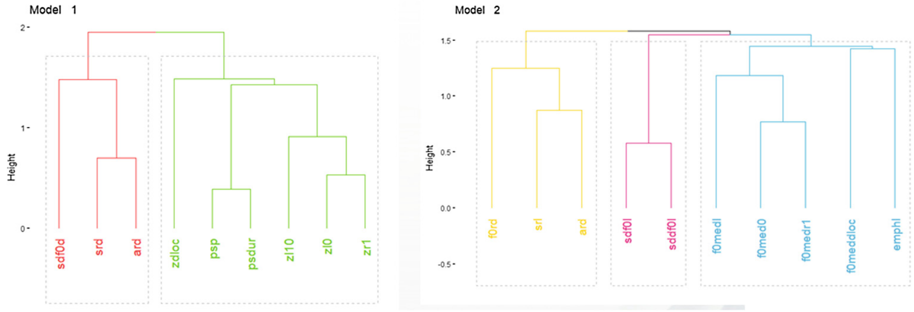


Fig. 2. Clusters of Models 1 and 2

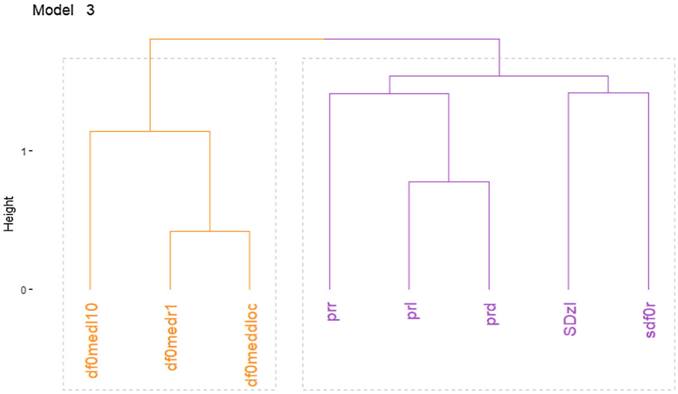


Fig. 3. Clusters of Model 3

4 Discussion

The method presented here uses phonetic transcription. In many cases we does not have access to the text related to the spontaneously spoken speech. This can be seen as a problem. However, automatic tools such as Automatic Aligners or Automatic Speech Recognition tools can minimize this problem.

The main differences between terminal and non-terminal boundaries with respect to their acoustic-phonetic parameters are related to the hierarchical relevance of the different measurements that allow the prediction of the two types of prosodic boundaries. Up to this point, the results of the research show that pauses and f0 measurements are very important for terminal boundaries and that this is not true for non-terminal ones. On the contrary, measurements of speech rate and of normalized and smoothed duration of the pre-boundary segments seem to be very important for non-terminal boundaries and less important for terminal ones. Measurements of intensity do not show much weight in both the boundary types.

With the exception of measurements of intensity, the model for terminal boundaries is consistent with the description of prototypical "conclusive" boundaries found in the literature. In relation to the intensity measurements, the difference found can be justified by the use of read speech or speech controlled in laboratory. Our model presents a clear hierarchy of acoustic parameters and also describes their relative importance.

At least in the database used in this study, signaling of utterance conclusion seems to be more typified, while signaling of boundaries of the non-terminal macrotype appears to be more stratified. We found three main groups of non-terminal boundaries mainly characterized by: (1) Measurements of pause and lengthening; (2) Measurements of speech rate, articulation rate and some f0 measurements; (3) Measurements of peak rate of smoothed z-score. The clusters corroborate the notion that prosodic boundaries are a complex and granular phenomenon, that is, the non-terminal category encompasses boundaries signaled by different sets of acoustic parameters, which probably correlate with different boundary sub-types.

The proposal presented in this paper is, in general, comparable with other proposals for automatic detection of prosodic boundaries available in the literature. However, this work presents some comparative advantages: to have spontaneous speech as target, to distinguish between two macro-types of boundaries (terminal and non-terminal ones) and to look for different configurations of non-terminal boundaries in order to better explain the variation among prosodic boundaries.

5 Future Research

Some important aspects to be looked at in future research include: a reduction of the analysis windows performed by the script; a better refinement of the model for terminal boundaries, which was temporarily put aside, given its good performance and the necessity of paying more attention to the models for non-terminal boundaries; a careful examination of the cases of disagreement between the automatic segmentation and the group of human annotators, seeking a better understanding of these disagreements; the application of the models to a larger dataset, which would be important especially to confirm the three models for non-terminal boundaries, since the third model had to be based on a very limited amount of data; the search for a functional explanation in syntactic and pragmatic terms for different types and sub-types of non-terminal boundaries.

References

1. Schubiger, M.: English Intonation: Its Form and Function. Niemeyer, Tübingen (1958)
2. Chafe, W.: The deployment of consciousness in the production of a narrative. In: Chafe, W. (ed.) The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production, pp. 9–50. Ablex, Norwood (1980). (Org.)
3. Schuetze-Coburn, S.: Prosody, syntax, and discourse pragmatics: assessing information flow in German conversation. Ph.D. University of California, Los Angeles (1994)

4. Ladd, R.: Intonational Phonology, 2nd edn. CUP, Cambridge (2008)
5. Cooper, W., Paccia Cooper, J.: Syntax and Speech. Harvard Universty Press, Cambridge (1980)
6. Selkirk, E. Comments on Intonational phrasing in English. In: Frota, S., Vigário, M., Freitas, M.J. (eds.) Prosodies, pp. 11–58. Mouton de Gruyter, Berlin (2005)
7. Halliday, M.A.K.: Speech and Situation. University College, London (1965)
8. Cresti, E.: Corpus di Italiano parlato, vol. 1. Accademia della Crusca, Firenze (2000)
9. Szczepek Reed, B.: Prosody, syntax and action formation: intonation phrases and action components. In: Bergmann, P. et al. (eds.), Prosody and Embodiment in Interactional Grammar, pp. 142–169. Mouton de Gruyter, Berlin (2012)
10. Chafe, W.: Discourse, Consciousness and Time: The Flow and Dsiplacement of Conscious Experience in Speaking and Writing. University of Chicago Press, Chicago (1994)
11. Croft, W.: Intonation Units and grammatical structure. Linguistics **33**(5), 839–882 (1995)
12. Bybee, J.: Language, Usage and Cognition. CUP, Cambridge (2010)
13. Barth-Weingarten, D.: Intonation Units Revised: Cesuras in Talk-in-Interaction. John Benjamins Publishing Company, Philadelphia (2016)
14. Pike, L.: The Intonation of American English. University of Michigan Press, Ann Arbor (1945)
15. Pierrehumbert, J. Phonetics and phonology of English intonation. Ph.D. Massachusetts Institute of Technology (1980)
16. Schegloff, E.: Reflections on studying prosody in talk-in-interaction. Lang. Speech **41**(3–4), 235–263 (1998)
17. Szczepek Reed, B.: Turn-final intonation in English. In: Couper-Kuhlen, E., Ford, C. (eds.), Sound Patterns in Interaction, pp. 97–117. Benjamins, Amsterdam (2004)
18. Avanzi, M., Lacheret-Dujour, A., Victorri, B.: A tool for semi-automatic annotation of french prosodic structure. In: ANALOR, pp. 119–122, Campinas, Brazil, (2008)
19. Ni, C.J., Zhang, A.Y., Liu, W.J., Xu, B.: Automatic prosodic break detection and feature analysis. J. Comput. Sci. Tchol. **27**, 1184–1196 (2012)
20. Kim, J.: Automatic detection of sentence boundaries, disfluencies, and conversational fillers in spontaneous speech. 103 f. Ph.D. University of Washington (2004)
21. Raso, T., Mello, H. (Org.): C-ORAL-BRASIL I: corpus de referência do português brasileiro falado informal, 1 edn. UFMG, Belo Horizonte (2012)
22. Raso, T., Mello, H. (Org.): C-ORAL-BRASIL II: corpus de referência do português brasileiro falado informal (forthcoming)
23. Mello, H.R. et al.: Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. In: Raso, T., Mello, H.R. (eds.) C-ORAL-Brasil I: Corpus de referência do português brasileiro falado informal, pp. 125–176. Editora UFMG, Belo Horizonte (2012)
24. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378–382 (1971)
25. Boersma, P., Weenink, D. Praat: doing phonetics by computer. 2015. Software http://www.praat.org/. Accessed 16 Jan 2015
26. Barbosa, P.: Semi-automatic and automatic tools for generating prosodic descriptors for prosody research. In: Bigi, B., Hirst, D. (eds.), Proceedings of the Tools and Resources for the Analysis of Speech Prosody, vol. 13, pp. 86–89. Aix-en-Provence: Laboratoire Parole et Language (2013). http://www.lpl-aix.fr/~trasp/Proceedings/19874-trasp2013.pdf
27. Barbosa, P.: BreakDescriptor (Versão 1.0) [Programa de computador] (2016). Available with the author
28. R Development Core Team: R a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2017)

Inner Speech in Portuguese: Acquisition Methods, Database and First Results

Carlos Ferreira[1,2,9], Alexandre Sayal[1,2], Bruno Direito[1], Marco Simões[2,3,4], Paula Martins[5,6,7], Catarina Oliveira[7,8], Miguel Castelo-Branco[2,3], and António Teixeira[7,8](✉)

[1] Institute of Nuclear Sciences Applied to Health, University of Coimbra, Coimbra, Portugal
cdferreira@ua.pt, c_dferreira@yahoo.com
[2] CIBIT Coimbra Institute for Biomedical Imaging and Translational Research, ICNAS, University of Coimbra, Coimbra, Portugal
[3] Faculty of Medicine, University of Coimbra, Coimbra, Portugal
[4] Center for Informatics and Systems, University of Coimbra, Coimbra, Portugal
[5] School of Health Sciences, University of Aveiro, Aveiro, Portugal
[6] Institute of Biomedicine, University of Aveiro, Aveiro, Portugal
[7] Institute of Electronics and Telematics Engineering of Aveiro (IEETA), Aveiro, Portugal
[8] Department of Electronics, Telecommunications and Informatics, University of Aveiro, Aveiro, Portugal
ajst@ua.pt
[9] Perspectum Diagnostics, Oxford, UK

Abstract. In this paper, we present a database developed for studying inner speech brain related areas using functional Magnetic Resonance Imaging (fMRI) in the context of the European Portuguese. First, we addressed the type of stimuli used in inner speech studies. In this sense, considering a preliminary study using a picture naming task, we defined a corpus. The corpus was designed based on cardinal vowels, syllable, disyllabic words and sentences with structure S(ubject)V(erb)O(bject) which were balanced in syllable number (six to ten). All the words used are common words from the Portuguese lexicon and possible ambiguities were excluded. Currently, the dataset includes data from twenty healthy participants native Portuguese speakers. Preliminary, exploratory analysis on the data allowed us to identify the most relevant areas part of the inner speech network, that include inferior frontal gyrus (including Broca's area), supplementary motor area and precentral gyrus. Ultimately, the better understanding of the inner speech mechanisms will pave way to the development of novel intervention strategies in linguistic disorders.

Keywords: Human speech production · Inner speech
Brain function · Resources
functional Magnetic Resonance Imaging (fMRI)

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 438–447, 2018.
https://doi.org/10.1007/978-3-319-99722-3_44

1 Introduction

The inner voice generated inside our brain is often called as inner speech [10]. Engaged in several processes of speech organization and language, inner speech is related with memory processing, reading, comprehension, consciousness, inner thought (self-reflection tasks) and prospective thought [2,5,7,10].

Neuroimaging studies have hypothesized an inner speech network involving different areas such as the left inferior frontal gyrus (including Broca's area), Wernicke's area, right temporal cortex, supplementary motor area (SMA), insula, right superior parietal lobule and right superior cerebellar cortex [2–4,6,10].

To better evaluate the particularities of inner speech in the context of each language, adapted protocols suited to the specific features of each language and target population are decisive. Very few studies have been proposed to assess inner speech for Portuguese language and, to best of our knowledge, no databases with functional MRI data is available to study inner speech in Portuguese.

For the Portuguese language, few databases exist to support speech production studies - being an example using MRI described in [13] - and to best of our knowledge no database of functional MRI for inner speech exists. In this sense, we believe that there is a need to create such a database using an adequate protocol, aligned with the state-of-the-art and taking in consideration the target language.

In this article, we will present our efforts to design a stimulation protocol using verbal stimuli and to develop a database, that will ultimately contribute to map inner speech network in the context of the european portuguese. First results, that can be obtained from database analysis, are also included.

Paper Structure: The paper is structured as follows: a brief introductory section presenting the concept of inner speech and the rationale behind the design and creation of the neuroimaging database; the "Protocol Definition" section presents the process that led to the definition of the final corpus; in the "A first inner speech database for Portuguese" section we present the corpus, the parameters for fMRI acquisition, the stimulation protocol, a characterization of the participants and the tools used for image processing and analyses. Section 4, "First Exploratory Results" presents some results and, finally, Sect. 5, presents the conclusions and some comments on future work.

2 Protocol Definition

A systematic review was conducted in order to evaluate the type of protocols currently used to map inner speech brain related areas. To this end, English language articles published prior to June 2017 were identified based on a query in PubMed with the following search key ("inner speech" or "silent speech" or "speech imagery") and (fMRI or "magnetic resonance" or MRI). The query returned seventy-one articles that were included in the analysis. The initial objective was to identify studies that used stimulation protocols based on verbal stimuli (e.g, words and sentences). The final inclusion criteria were articles written in

English, reporting task-related fMRI with BOLD (blood oxygen level dependent) activity (i.e. excluding resting-state studies) and included: healthy individuals; sentences and words as a stimulation paradigm; data from adults (> eighteen years old); the complete functional stimulation paradigm; more than one case (not case report). The query resulted on a set of three articles that were analyzed and supported the design of our stimulation protocol.

According to these studies [2,4,10], several paradigms can be used to assess inner speech brain related areas. From letter and object naming, to verb generation, reading, working memory task, counting or semantic fluency tasks, these are some of the stimuli that were used by the authors to map the inner speech brain network.

Based on a preliminary study using picture naming task approach and considering previous work by [1,9,11,14], several criteria were defined in order to design the stimulation protocol:

- All EP stressed vowels
- Two-syllable words stressed on the first syllable
- Most frequently used Portuguese words
- Subject–verb–object (SVO) sentences with six to ten syllables that include some words of the selected corpus.

The design of the stimulation protocol presented several challenges. For instance, the most recent linguistic tool developed for European Portuguese only counted frequency in a given text and did not have the results of frequency of the corpus. The unique database of frequency published for the European, Portuguese [8], is from 1987, and naturally differs from the one spoken actually. The inclusion of all stressed vowels is also challenging, essentially because of the size of the corpus. To address these challenges, we identified additional or alternative criteria, like the use of corner vowels instead of all stressed vowels and the use of common words instead of frequent words. As a result of these improvements the final criteria to select the corpus were defined as:

- Corner vowels ([a], [i] and [u]) to reduce the set and simplify the representation of intended vowel;
- Two-syllable words stressed on the first syllable;
- Common words in European Portuguese;
- subject–verb–object (SVO) sentences with six to ten syllables that include some words of the selected corpus.

The final stimulation protocol is illustrated in Fig. 1.

3 A First Inner Speech Database for Portuguese

The study consisted in the recording and analysis of fMRI data while native speakers of Portuguese performed inner speech tasks in response to verbal stimuli visually presented. All acquisitions were made at Institute of Nuclear Sciences Applied to Health in Coimbra.

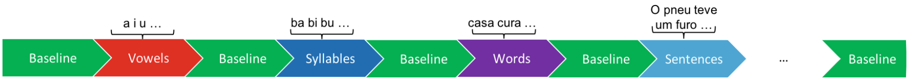


Fig. 1. Stimulation paradigm. Baseline - consisting in a fixation cross and the participants were instructed to focus on it; Vowels - vowels presentation block; Syllables - syllables presentation block; Words - words presentation block; Sentences - sentences presentation block. Each representation of the verbal stimuli was presented at the screen during 2 s.

3.1 Corpus

The corpus was essentially designed to evaluate inner speech brain network and uses Portuguese common words and sentences, complying with the criteria previously presented.

Fifteen syllables and words were chosen from a list of common words selected, for each cardinal vowel used. The list was selected based on common words of the Portuguese language and on words that include some of the syllables selected. Fifteen sentences were also created using words previously defined for the corpus. The list of elements included in the stimuli is available in Table 1.

3.2 Functional MRI Acquisition

The data were collected using a Siemens Magnetom Trio 3 T scanner (Erlangen, Germany) with a 12-channel head coil.

Anatomical images were acquired using a sagittal T1 3D MPRAGE sequence with the following parameters: TR = 2530 ms; TE = 3.42 ms; TI = 1100 ms; flip angle = 7°; 176 slices; matrix size 256 × 256; voxel size 1 × 1 × 1 mm.

After the anatomical scan, functional maps were obtained using axial gradient echo-planar imaging BOLD sequences parallel to the bi-commissural plane with the following parameters: TR = 2000 ms; TE = 30 ms; 38 slices; matrix size 70 × 70; voxel size 3 × 3 × 3 mm.

Visual stimuli were presented on a NordicNeuroLab (Bergen, Norway) LCD monitor, with a resolution of 1920 × 1080 pixels, refresh rate 60 Hz.

3.3 Speech Recording

As in [13], audio was recorded simultaneously with the fMRI acquisition inside the MR scanner to ensure that participants were performing the task silently. Audio was collected at a sampling rate of 16000 Hz, using a fiber optic microphone (Optoacoustics FOMRI III Dual Channel MRI microphone, Or Yehuda, Israel). The microphone was fixed on the head coil, with the protective popscreen placed directly against the speaker's mouth, according to the manufacturer's instructions.

Table 1. List of selected verbal stimuli for the inner speech database. Stimuli are separated in tables for Syllables, Words and Sentences. Tables for words and sentences present, in consecutive rows, the real stimuli and the English translation.

Syllables

pa	ta	ka	ba	da	ga	ma	na	fa	sa	cha	va	ja	la	ra
pi	ti	ki	bi	di	gui	mi	ni	fi	si	chi	vi	ji	li	ri
pu	tu	ku	bu	du	gu	mu	nu	fu	su	chu	vu	ju	lu	ru

Words

casa	quilo	cura
(house)	(kilo)	(cure)
dado	dica	duche
(given)	(cue)	(shower)
faca	figo	fula
(knife)	(fig)	(furious)
fato	fita	furo
(fact)	(tape)	(hole)
lado	liso	lupa
(side)	(smooth)	(magnifying glass)
lama	lixo	luta
(mud)	(garbage)	(fight)
mala	missa	mula
(bag)	(mass)	(mule)
mapa	mito	muro
(map)	(myth)	(wall)
pano	pinho	pula
(cloth)	(pine)	([she] jumps)
pato	pipa	puro
(duck)	(kite)	(pure)
rato	rica	rumo
(mouse)	(rich)	(bearing)
ramo	ripa	russo
(branch)	(slat)	(russian)
saco	silo	sujo
(bag)	(silo)	(dirt)
sala	sino	sumo
(room)	(bell)	(juice)
chave	china	chuva
(key)	(China)	(rain)

Sentences

A avó bebeu o sumo.
(Grandma drank the juice.)
O barco perdeu o rumo.
The boat lost its way.
A casa é amarela.
(The house is yellow.)
A casa é de madeira.
(The house is wooden.)
A criança come o figo.
(The child eats the fig.)
A faca corta a maçã.
(The knife cuts the apple.)
O lixo cheira muito mal.
(The garbage smells really bad.)
A pipa está cheia.
(The kite is full).
O pneu teve um furo.
(The tyre had a puncture.)
O rapaz toma duche.
(The boy takes a shower.)
O rato come queijo.
(The mouse eats cheese.)
A sala é pequena.
(The room is small.)
A senhora é rica.
(The lady is rich.)
O sino está muito alto.
(The bell is too high.)
O xarope cura a tosse.
(The syrup cures the cough.)

A computer running OptiMRI software (version 3.1), located in the adjacent MRI control room, recorded the dual-channel microphone outputs, the filtered speech processed by DSP and up to 3 TTL pulses, all of them synchronized with high accuracy (FOMRI III user manual).

3.4 Stimulation Protocol

The experimental protocol comprised one anatomical and three functional runs. The functional runs consisted of an inner speech production task, based on visually presented instructions. Each run included four task conditions and were based on the following verbal stimulus conditions: vowels, syllables, words and sentences. Each trial comprised rest blocks of 12 s interleaved with four task blocks of 30 s, one of each type. Each run consisted of three repetitions of each condition.

During the rest blocks, participants were instructed to focus at a central fixation cross, while during the task blocks, the participants' instruction was to silently name the vowel, syllable, word or sentence presented at each 2 s. Given the list of verbal stimuli, each word and syllable was repeated two times, each sentence six times and each vowel thirty times, across the three runs.

3.5 Speakers

Twenty-two healthy volunteers native Portuguese speakers (mean age 28.7 years old; eleven males) were enrolled in this study.

All participants had normal or corrected to normal vision, and no history of neurological disorders. The Edinburgh handedness test was applied to the participants to ensure they were all right handed (mean 90% right) and Portuguese was their native language.

The study was approved by the Ethics Commission of the Faculty of Medicine of the University of Coimbra and was conducted in accordance with the declaration of Helsinki. All subjects provided written informed consent to participate in the study.

3.6 Post-processing

Preprocessing and analysis were conducted using BrainVoyager 20.6 (Brain Innovation, Maastricht, Netherlands). Preprocessing of single-subject fMRI data included slice-time correction, realignment to the first image to compensate for head motion and temporal high-pass filtering to remove low-frequency drifts. The anatomical images were co-registered to the functional volumes and all images were normalized to Talairach coordinate space [12].

After preprocessing, in the first-level analysis of the functional data, general linear model (GLM) analysis was used for each run. Predictors were modeled as a boxcar function with the length of each condition, convolved with the canonical hemodynamic response function (HRF). Six motion parameters (three translational and three rotational) and predictors based on spikes (outliers in the BOLD time course) were also included into the GLM as covariates.

4 First Exploratory Results

An exploratory analysis was performed with the dataset available in order to investigate the areas activated by inner speech.

To this end, we grouped individual data and report group level results. For this analysis, at the group level, we applied 3D spatial smoothing with a Gaussian filter of 6 mm to the data and performed a random effects GLM (RFX-GLM) analysis to map the most important brain regions involved in inner speech. We used the contrast "task" > "baseline", correcting for multiple comparisons with False Discovery Rate (FDR) correction (considering a maximum false discovery rate of 5%).

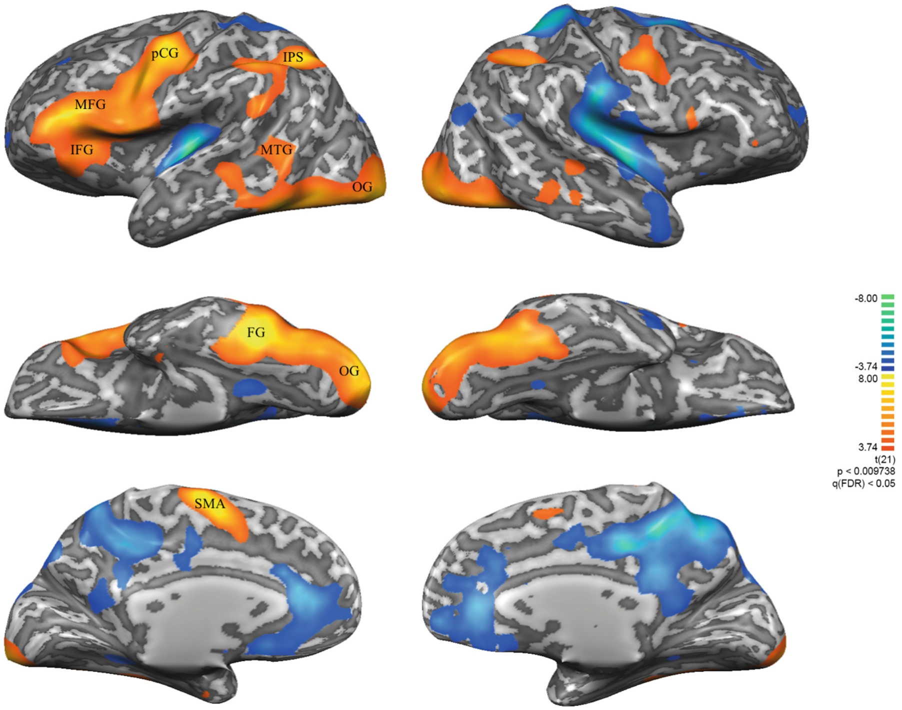


Fig. 2. RFX-GLM group activation map for the inner speech runs (q(FDR)< 0.05), showing areas with higher activation during the four task conditions than during baseline (lateral view, bottom view and medial view). Several regions are marked: (IFG) Inferior Frontal Gyrus including Broca's area; (MFG) Middle Frontal Gyrus; (pCG) preCentral Gyrus; (IPS) Intraparietal Sulcus; (MTG) Middle Temporal Gyrus including Wernicke's area; (FG) Fusiform Gyrus; (OG) Occipital Gyrus; (SMA) Supplementary Motor Area.

The results of the group analysis, as presented in Figs. 2 and 3, allow the identification of the most important brain areas related to inner speech:

- Inferior Frontal Gyrus (IFG), including Broca's area;
- Middle Frontal Gyrus (MFG);
- Middle Temporal Gyrus (MTG), including Wernicke's area;

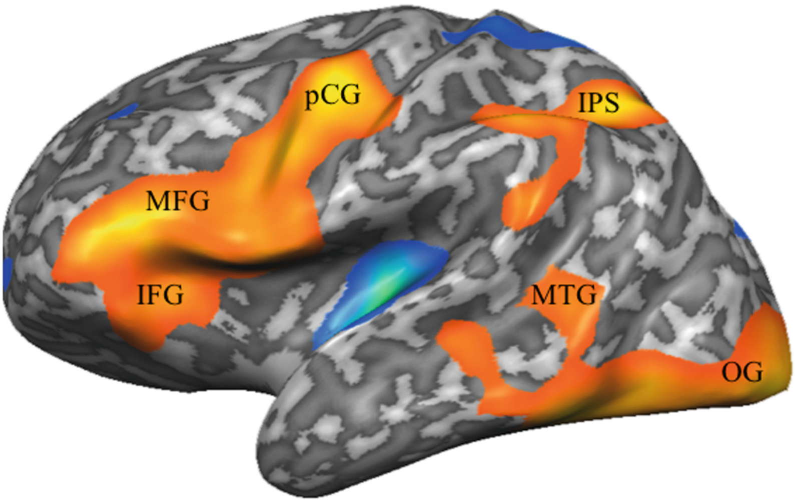


Fig. 3. Left lateral sagittal view of group activation map of Fig. 2 in more detail, showing areas with higher activation during the four task conditions than during baseline. Several regions are marked: (IFG) Inferior Frontal Gyrus, including Broca's area; (MFG) Middle Frontal Gyrus; (pCG) preCentral Gyrus; (MTG) Middle Temporal Gyrus, including Wernicke's area; (IPS) Intraparietal Sulcus; (OG) Occipital Gyrus.

- preCentral Gyrus (pCG);
- Fusiform Gyrus (FG);
- Supplementary Motor Area (SMA);
- Intraparietal Sulcus (IPS);
- Occipital areas.

These results are in accordance with the previous studies for English language [2–4, 6, 10].

Additionally, preliminary audio analysis confirms that the participants were performing the task silently as there were no speech recordings in the collected audio. This strengthens the hypothesis that these results are task related and excludes motor execution components associated with speech production. Ultimately, this study represents the first effort to map an inner speech European Portuguese brain network, and supports the notion that most regions overlap across languages.

5 Conclusion

We were able to develop a novel verbal stimuli database for the European Portuguese language which allowed us to map several brain areas related with inner speech production. To the best of our knowledge, an updated, adapted stimuli as such was lacking for the Portuguese language and this stands as the first database established in this direction.

Preliminary results in a healthy population show the feasibility of the paradigm, highlighting the most relevant brain areas associated with inner speech tasks in an fMRI context. Concomitant audio recordings, ensured that overt speech was absent during the task.

5.1 Future Work

Using the database, and in conjunction with overt speech fMRI data using the same corpus, several studies will be performed, such as: evaluation of the parametric effect of task difficulty (from vowels to sentences), assessment of possible differential activation in task-related brain areas relative to task difficulty, comparison between the activation maps of inner and overt speech tasks and functional connectivity analysis to evaluate the connections between the areas recruited during an inner speech task. A better understanding of the inner speech mechanisms will ultimately be decisive to develop intervention strategies in linguistic disorders.

Acknowledgements. Research partially supported by IEETA Research Unit, funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UID/CEC/00127/2013.

Research also partially supported by CNC.IBILI Research Unit, funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UID/NEU/04539/2013, COMPETE POCI-01-0145-FEDER-007440 and by the Grant BIGDATIMAGE, CENTRO-01-0145-FEDER-000016 financed by Portugal 2020 Centro 2020 FEDER COMPETE.

References

1. Berken, J.A., et al.: Neural activation in speech production and reading aloud in native and non-native languages. Neuroimage **112**, 208–217 (2015)
2. Geva, S., Jones, P.S., Crinion, J.T., Price, C.J., Baron, J.C., Warburton, E.A.: The neural correlates of inner speech defined by voxel-based lesion-symptom mapping. Brain **134**(10), 3071–3082 (2011)
3. Jones, S.R., Fernyhough, C.: Neural correlates of inner speech and auditory verbal hallucinations: a critical review and theoretical integration. Clin. Psychol. Rev. **27**(2), 140–154 (2007)
4. Marvel, C.L., Desmond, J.E.: From storage to manipulation: how the neural correlates of verbal working memory reflect varying demands on inner speech. Brain Lang. **120**(1), 42–51 (2012)
5. Morin, A., Hamper, B.: Self-reflection and the inner voice: activation of the left inferior frontal gyrus during perceptual and conceptual self-referential thinking. Open Neuroimaging J. **6**, 78–89 (2012)
6. Morin, A., Michaud, J.: Self-awareness and the left inferior frontal gyrus: inner speech use during self-related processing. Brain Res. Bull. **74**(6), 387–396 (2007)
7. Morin, A., Uttl, B., Hamper, B.: Self-reported frequency, content, and functions of inner speech. Procedia Soc. Behav. Sci. **30**, 1714–1718 (2011)

8. Bacelar do Nascimento, M.F., Garcia Marques, M.L., Segura da Cruz, M.L.: Português Fundamental - Métodos e Documentos, Tomo 1, Inquérito de Frequência. INIC, CLUL, Lisboa, 1 edn. (1987)
9. Partovi, S., et al.: Effects of covert and overt paradigms in clinical language fMRI. Acad. Radiol. **19**(5), 518–525 (2012)
10. Perrone-Bertolotti, M., Rapin, L., Lachaux, J.P., Baciu, M., Loevenbruck, H.: What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. Behav. Brain Res. **261**, 220–239 (2014)
11. Raichle, M.E.: A brief history of human brain mapping. Trends Neurosci. **32**(2), 118–126 (2009)
12. Talairach, J., Tournoux, P.: Co-planar Stereotaxic Atlas of the Human Brain. Thieme, New York (1988)
13. Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., Shosted, R.: Real-time MRI for Portuguese. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (eds.) PROPOR 2012. LNCS (LNAI), vol. 7243, pp. 306–317. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28885-2_35
14. Tremblay, P., Small, S.L.: From language comprehension to action understanding and back again. Cereb. Cortex **21**(5), 1166–1177 (2011). https://doi.org/10.1093/cercor/bhq189. https://doi.org/10.1093/cercor/bhq189

CNN-Based Phonetic Segmentation Refinement with a Cross-Speaker Setup

Luis Gustavo D. Cuozzo, Diego Augusto Silva(✉), Mario Uliani Neto,
Flávio Olmos Simões, and Edson Jose Nagle

CPqD, Campinas, SP, Brazil
{lcuozzo,diegoa,uliani,simoes,nagle}@cpqd.com.br

Abstract. This work proposes a method to improve the performance of automatic phonetic alignment of speech data. The method uses a deep convolutional neural network (CNN) trained on a combination of acoustic features extracted from labeled data to fine tune the position of each boundary within a fixed-size window around the original boundary position. The proposed method is robust to speaker identity, which means that a system trained with enough labeled data can be used to fine tune alignment on any speech file, regardless of speaker identity. With an absolute gain between 20% and 33% in cross speaker scenario, our results demonstrate the applicability of deep learning for this task.

Keywords: Phonetic segmentation refinement · Forced alignment
Deep neural networks

1 Introduction

The construction of a text-to-speech (TTS) system requires a collection of speech data, together with an indication of all phoneme boundaries present in the data. The quality of the synthesized speech depends on an accurate indication of phoneme boundaries throughout the dataset. Manual boundary positioning can be performed with the aid of tools such as WaveSurfer [13], Praat [5] and Elan [15]. This is a time-consuming task, demanding as much as 130 times the duration of the audio [8], which makes it impractical for typical TTS datasets consisting of several hours of speech.

Employing multiple human annotators can reduce the time required for labeling, at the expense of introducing inconsistencies in boundary annotation, which can be harmful to synthesized speech quality as well.

Automatic approaches are employed to alleviate the above mentioned issues. Early approaches consisted of syllable segmentation with dynamic time warping (DTW) [3]. Such alignment tools, known as "forced aligners", have been reported to produce a consistent and reproducible alignment in a relatively shorter time compared to the manual approach. Automatic alignment tools include Prosodylab-aligner [6] and Montreal Forced Align (MFA) [10] supported

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 448–456, 2018.
https://doi.org/10.1007/978-3-319-99722-3_45

by Kaldi toolkit [11], both employing Hidden Markov Models (HMM) segmentation; as well as Support Vector Machines (SVM) based approaches [9] and DNN-based approaches [1,4].

Our work proposes a post-processing tool that fine adjusts in the position of each boundary within the limits of a fixed-size analysis window around the original boundary position. Additionally, the proposed procedure has the advantage that it can be employed in conjunction with traditional automatic forced-alignment tools.

Automatic segmentation of speech [7] is also useful for the building automatic speech recognition (ASR) systems and a detailed review of phonetic segmentation for the purpose of ASR is beyond the aim of this article.

This paper is organized as follows: Sect. 2 describes the proposed system; Sect. 3 describes the speech data used in the experiments, consisting of Brazilian Portuguese training and testing data, with manual alignments used as ground truth; Sect. 4 details the experimental setup used to assess the proposed system. Section 5 exhibits achieved results and discusses them. Finally, Sect. 6 presents conclusions and future extensions.

2 Proposed Refinement Scheme

2.1 Front-End Processing

We generated an appropriate combo vector sequence, with 16 features each vector, to create the input to the neural network. First, 12 Mel-generalized Cepstral Coefficients (MGCC) [14] including energy are computed on speech files sampled at 16 kHz, 16 bits per sample with $\alpha = 0.42$ and $\gamma = -0.5$ using the SPK toolkit [2]. The analysis frame is 25 ms long, with a 2 ms frame shift. Then, the MGCC features are appended with other 4 features: 2 based on Euclidean distance computed between MGCC (without energy) vectors for 10 and 20 ms spacing, centered on the frame whose features are being computed; whereas other 2 ΔEnergy features are computed between energy coefficient for 10 and 20 ms spacing centered on the same frame, completing the feature vector.

2.2 Deep Neural Network

The neural network architecture of our phonetic refinement approach is depicted in Fig. 1. The topology of the network is a simplified version of VGG16-net [12], with a several convolutional, pooling layers followed by two fully connected.

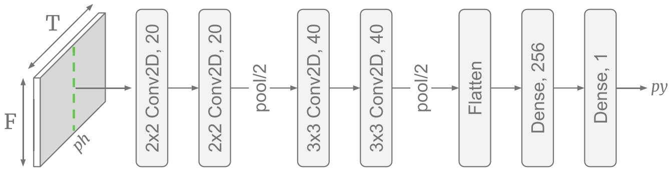


Fig. 1. Neural network topology

The input of the network is a matrix Sh representing a context window centered around the boundary ph, and the output is an index ranging from -25 to 25, representing the optimal position of the phoneme boundary within the context window.

To better illustrate the input of the neural network, temporal (upper image), spectrogram (middle image) and phone (bottom image) representations are shown in Fig. 2. The vertical axis of the spectrogram image corresponds to the spectrum, while T is the context size in number of time bins.

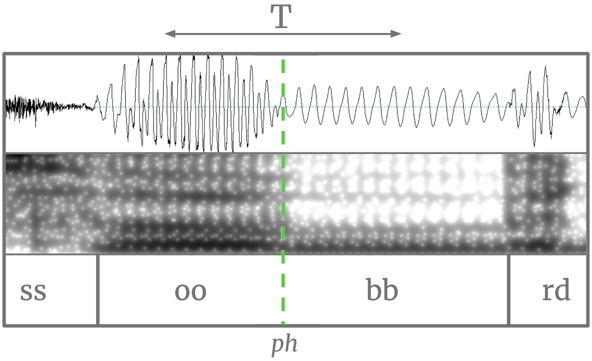


Fig. 2. Neural network context size representation. Illustrated with a part of the word "**sobr**inho" (nephew, in English)

In our case, the spectrum is represented by our combo feature vector whose dimension is set to $F = 16$ and the context size of the input is $T = 50$.

$$dim(Sh) = (F, T) \tag{1}$$

During training, 50 different input matrices are generated for each boundary. In each of them, the manually annotated position ph occupies a different position along all 50 possible time bins. During inference, only one input matrix is used for each boundary to be updated, in which the MFA boundary is positioned on the central bin.

The CNN was trained with the input matrices described above and no pre-training technique was applied. Twenty training epochs were executed with the RMSprop optimizer, and the mean-square error (MSE) was used as the loss function. For a given input matrix we define the error as the distance, in the acoustic feature space, between the central position of the input context and the ground true boundary. The following values were used as the optimizer parameters: learning rate = 0.001, $\rho = 0.9$ and ϵ = 1e-08. The training partition was divided in the following manner: 80% for training and 20% for validation.

3 Databases

We applied the proposed methodology using clean dictated sentences designed for speech synthesis purposes. The speech samples are part of proprietary datasets collected by CPqD. The material consist of three separate datasets (two female speakers and one male speaker). The development set consists of samples from Female 1, while evaluation was performed with samples from all speakers, including the one used for training. Train and test samples from Female 1 do not overlap.

The sentences present in all sets were built to be phonetically balanced, covering, as much as possible, all phonetic and prosodic contexts that can occur in Brazilian Portuguese. The phonetic transcriptions were generated by a proprietary grapheme to phoneme tool and the manual alignment, when available, was created with the Praat tool. Table 1 summarizes the main characteristics of the datasets.

Table 1. Speakers database duration (#sentences), DNA - Does not apply

Dataset	Train	Test
Female 1	450 min (10277)	33 min (563)
Female 2	DNA	21 min (438)
Male	DNA	33 min (736)

4 Experiments

Shortly, the proposed scheme consists of a refinement CNN operating through a context window centered around phonetic boundary outputs predicted by the MFA automatic aligner. Both MFA and CNN models are trained in a supervised manner, using the training partition of Female 1 dataset. We implemented two experimental pipelines which are described below:

1. Same speaker
   (a) Align Female 1 test partition (baseline),
   (b) Refine boundaries predicted by aligner using CNN models.
2. Cross speaker
   (a) Align Female 2 and Male sets using MFA model (baseline),
   (a) Refine boundaries predicted by aligner using CNN models.

4.1 Baseline: Training and Aligning with MFA

Our baseline system uses the GMM-based MFA forced alignment tool [10]. Automatic phonetic transcriptions of all sentences in the datasets were generated by means of a Brazilian Portuguese grapheme to phoneme conversion module that is part of CPqD commercial text-to-speech system.

A two-letter representation of phonetic symbols were employed in order to represent 42 Brazilian Portuguese phonemes, which were grouped into 9 classes by similarity: four unvoiced plosives {pp, tt, ts, kk}, four voiced plosives {bb, dd, dz, gg}, three unvoiced fricatives {ff, ss, sh}, four voiced fricatives {vv, zz, zh, rx}, five liquids {rf, rd, rr, ll, lh}, three nasal consonants {mm, nn, nh}, seven tonic vowels {ii, ee, eh, aa, oh, oo, uu}, five non-tonic vowels {ic, ij, ac, uc, uw} and seven nasal vowels {in, ij, en, an, on, un, wn}.

The MFA model was trained with one Brazilian Portuguese speaker (Female 1 train partition) by means of the MFA binary tool `mfa_train_and_align`. The forced alignment task was then applied to Female 1 test partition (same speaker pipeline) as well as to Female 2 and Male datasets (cross speaker pipeline), using `mfa_align` tool.

4.2 Proposition

During training, a model is created for each pair of phonetic classes, which are described in Sect. 4.1. A total of 63 models were created. Some pairs were not trained because the results obtained with MFA for these pairs are already good, so no refinement is necessary. The 63 trained models cover 85% of phoneme boundary occurrences in our datasets.

At inference time, a context window $X = (Sx, px)$ is used as the CNN input, with the central position px predicted by the MFA automatic aligner. The output of the refinement is of type py, corresponding to the index point position of phoneme boundary as illustrated in Fig. 3.

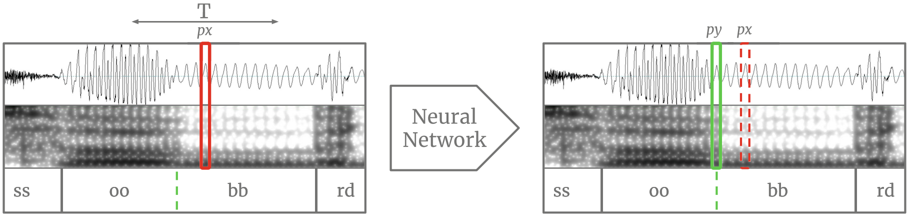


Fig. 3. Inference schematic. Illustrated with a part of the word "**sobr**inho" (nephew, in Portuguese)

The CNN inference is executed for all boundaries that can be mapped to one of the 63 trained models. After inference, each boundary is adjusted to its optimal point within the corresponding context window, and the metric used for accuracy of the refinement procedure is the difference (in ms) from the ground-truth annotation.

Figure 3 also shows an example of phoneme boundary, whose left phone {oo} is associated to the tonic vowel class and whose right phone {bb} belongs to the voiced plosive class. All evaluations were performed on both same and cross-speaker scenarios specified above.

5 Results

We present our results based on the same and cross-speaker scenarios and compare two systems: MFA and MFA + CNN with different tolerance levels. Table 2 shows the error rates (in percentage) of phoneme boundaries that were not correctly positioned in each dataset when compared to ground-truth for both MFA-only and MFA + CNN systems, with tolerance values of 5, 10, 25 and 50 ms (for a given tolerance, we consider that the boundary position is correct when its absolute difference from the ground truth position is lower than the tolerance value).

Table 2. Error rate for different tolerances (percentage below a cutoff)

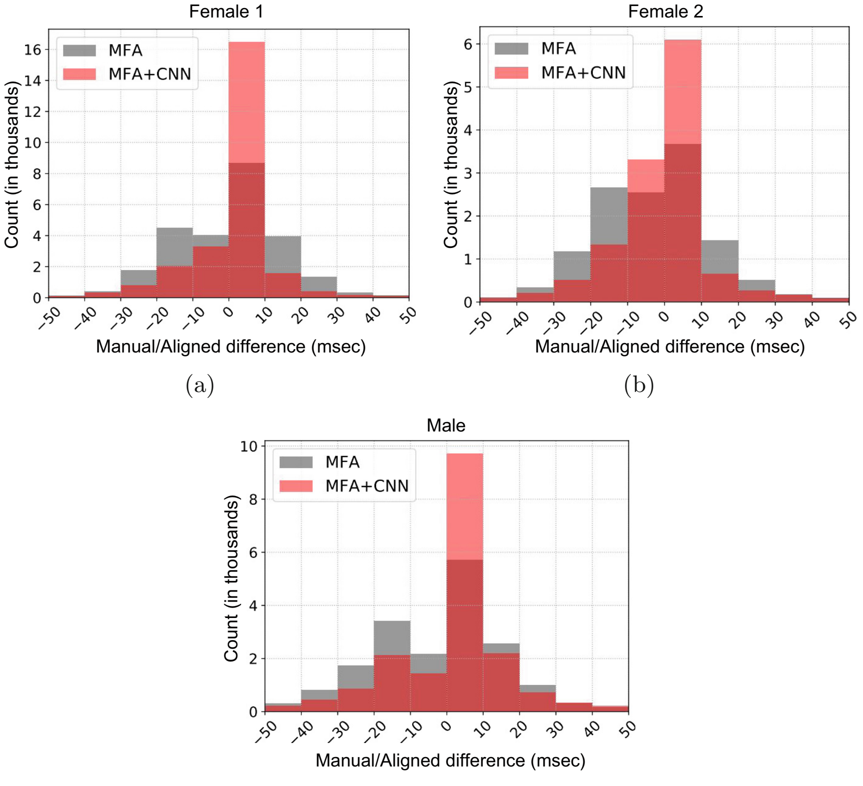
		Tolerance (ms)			
		<5	<10	<25	<50
Female 1	MFA	74.37	50.34	8.12	0.97
	MFA + CNN	41.48	19.98	3.95	0.81
Female 2	MFA	71.10	52.58	11.42	1.74
	MFA + CNN	44.91	27.70	7.89	1.65
Male	MFA	76.43	58.51	15.57	2.77
	MFA + CNN	55.84	41.00	11.12	2.54

For the Female 1 test set (Fig. 4(a)), adding the CNN refinement step reduced the percentage of incorrect boundaries from 74.37% to 41.48% for a tolerance value of 5 ms, from 50.34% to 19.98% for 10 ms, from 8.12% to 3.95% for 25 ms and from 0.97% to 0.81% for 50 ms. In this case, the speaker identity of train and test sets are the same.

In the case of Female 2 test set (Fig. 4(b)) (same sex but different speaker from training set), the percentages of incorrect boundaries are slightly higher, as expected, but still lower than those obtained with MFA-only: the error rates reduced from 71.10% to 44.91% for a 5 ms tolerance, from 52.58% to 27.70% for 10 ms, from 11.42% to 7.89% for 25 ms and from 1.74% to 1.65% for 50 ms.

Finally, for the Male test set (Fig. 4(c)), which is highly mismatched when compared to the training set, the increase in the correctly positioned boundaries is still significative: the error rates reduced from 76.43% to 55.84% for a 5 ms tolerance value, from 58.51% to 41.00% for 10 ms, from 15.57% to 11.12% for 25 ms and from 2.77% to 2.54% for 50 ms.

Figure 4 shows the distribution of boundary position errors for all systems and databases, in the form of histograms. Errors produced for the MFA + CNN system are lower for all tolerance values, which explains the narrower histograms (in red) when compared with the MFA-only gray histograms.



(a)

(b)

(c)

Fig. 4. Histograms of difference between force-aligned phone boundaries using MFA, MFA+CNN aligner and gold-standard annotations. (Color figure online)

6 Conclusions and Future Developments

This work presents a DNN-based approach to refine the position of phonetic boundaries generated by automatic forced-alignment tools. The well known Montreal Forced Aligner was used as a benchmark, and an experimental setup was built to measure in what extent our approach could contribute to increase the correctness of boundary locations.

The experimental results showed that the proposed system can lower significantly the alignment error rates. Moreover, our approach proved to be robust to speaker identity, which means that a system pre-trained on samples from a manually annotated single speaker dataset can be used to improve alignment performance on samples uttered by different speakers. This is particularly useful when creating new voices for a text-to-speech system, for which it allows fast and accurate alignment of newly recorded samples without the need to manually annotate data to adapt the aligner for a new speaker.

All datasets employed in our experiments are currently being used to build real commercial text-to-speech systems. The techniques described in this work have been incorporated to our process of creating new synthetic voices, since increasing the proportion of correct boundaries, specially for tolerance values between 5 and 10 ms, leads to a significant improvement in the perceived quality of synthetic speech.

Future work could examine the new conditions of the network training and new neural network topologies, which could also increase accuracy. Other possibilities involve adding phonetic and prosody features in a front-end processing, including an attempt to group all phonemes in a unique model, and apply speaker adaptation techniques in the alignment and refinement pipelines.

References

1. Forced alignment and goodness of pronunciation (GOP) with DNN support. https://github.com/tbright17/kaldi-dnn-ali-gop. Accessed 30 Mar 2018
2. Speech signal processing toolkit (SPTK), version: SPTK-3.11.tar.gz. http://sp-tk.sourceforge.net/
3. Adell, J., Bonafonte, A., Gómez, J.A., Castro, M.J.: Comparative study of automatic phone segmentation methods for TTS. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), vol. 1, p. I-309. IEEE (2005)
4. Baby, A., Prakash, J.J., Vignesh, R., Murthy, H.A.: Deep learning techniques in tandem with signal processing cues for phonetic segmentation for text to speech synthesis in Indian languages. Proceedings of Interspeech 2017, pp. 3817–3821 (2017)
5. Boersma, P.: Praat: doing phonetics by computer (2006). http://www.praat.org/
6. Gorman, K., Howell, J., Wagner, M.: Prosodylab-aligner: a tool for forced alignment of laboratory speech. Can. Acoust. **39**(3), 192–193 (2011)
7. van Hemert, J.P.: Automatic segmentation of speech. IEEE Trans. Signal Process. **39**(4), 1008–1012 (1991)
8. Kawai, H., Toda, T.: An evaluation of automatic phone segmentation for concatenative speech synthesis. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), vol. 1, p. I-677. IEEE (2004)
9. Lo, H.Y., Wang, H.M.: Phonetic boundary refinement using support vector machine. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 4, p. IV-933. IEEE (2007)
10. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: trainable text-speech alignment using Kaldi. In: Proceedings of interspeech (2017)
11. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, No. EPFL-CONF-192584. IEEE Signal Processing Society (2011)
12. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs], September 2014

13. Sjölander, K., Beskow, J.: Wavesurfer-an open source speech tool. In: Sixth International Conference on Spoken Language Processing (2000)
14. Tokuda, K., Kobayashi, T., Masuko, T., Imai, S.: Mel-generalized cepstral analysis. In: Proceedings of the ICSLP-94, pp. 1043–1046 (1994)
15. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: ELAN: a professional framework for multimodality research. In: 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 1556–1559 (2006)

Syntax & Parsing

Tagsets and Datasets: Some Experiments Based on Portuguese Language

Cláudia Freitas[1], Luiza F. Trugo[1], Fabricio Chalub[2], Guilherme Paulino-Passos[2,3], and Alexandre Rademaker[2,4](✉)

[1] PUC-Rio, Rio de Janeiro, Brazil
claudiafreitas@puc-rio.br, luizafrizzo@gmail.com
[2] IBM Research, Rio de Janeiro, Brazil
{fchalub,gpaulino,alexrad}@br.ibm.com
[3] UFRJ/COPPE/PESC, Rio de Janeiro, Brazil
[4] FGV/EMAp, Rio de Janeiro, Brazil

Abstract. We report the results of two experiments aimed at investigating the impact of linguistic variation on *PoS* tagging. In both cases, we depart from the conversion of the corpus MacMorpho [1], which was re-annotated according to the Universal Dependencies *PoS* tagset. Throughout the conversion process, we faced some linguistic challenges related to the past participle forms. As a result, we created two corpora (MacMoprho-UD and MacMorpho-UD+PCP). We used these three corpora (MacMorpho; MacMoprho-UD and MacMorpho-UD+PCP) to assess the impact on *PoS* learning in different scenarios.

Keywords: Corpus annotation · Tagset alignment · Past participles

1 Introduction

Part-of-speech (*PoS*) tagging is one of the earliest steps of NLP. However, in spite of its linguistic nature, linguistic studies comparing the impact of different tagsets on the performance of NLP systems are scarce – from the early work of [11], which examines the grammatical constructions which cause statistical n-grams taggers to falter more frequently, and the Eagles document produced in 1996, this seems to be an unattractive subject. On the other hand, comparison of PoS-taggers doesn't suffer from the same problem. A possible reason for this imbalance may lie in the belief that the distribution of words along certain categories are based on objective and stable properties associated with words, and it is up to machines or programmers to develop the best classification strategy. Additionally, a requirement for this kind of study is the existence of comparable materials – the same corpus annotated with different tagsets – a requirement that seems wasteful considering the amount of work involved.

For the Portuguese language, the disparity between studies focusing on taggers and on tagsets is the same. [13] assess the ambiguity in *PoS* tagging, considering a morphologic parser. Twenty years later, [7] is the only work we know

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 459–469, 2018.
https://doi.org/10.1007/978-3-319-99722-3_46

which focuses on tagsets, and it is conducted from the computer science perspective only. As to PoS tagger evaluation, on the other hand, the scenario is far richer, as seen in [9].

In this paper, we present a study on tagsets, originated from the conversion of the corpus MacMorpho [1], which was re-annotated, at the *PoS* level, according to the Universal Dependencies (UD) tagset [16]. Throughout the process, we faced some linguistic challenges, especially with past participle forms. As a result, we created a second version of the corpus, in which we kept the UD tagset but added the `PCP` tag - specific for past participles – used in the original MacMorpho corpus/tagset. We then conducted two experiments: the first one aimed to verify the impact of tagsets – original MacMorpho; MacMorpho-UD; and MacMorpho-UD+`PCP` – on system performance; in the second experiment, taking advantage of the converted Mac-Morpho-UD corpus, we assessed the impact of size and quality in training: we used both UD-Portuguese-Bosque [18] and MacMorpho-UD in the training phase.

2 Corpus Conversion

MacMorpho [1] is a Brazilian Portuguese newswire corpus developed within the scope of the Lacio-Web project [2]. It contains 1.1 million words, annotated at the *PoS* level and manually revised. The tagset uses 23 labels. Since 2003, MacMorpho underwent two revisions, eliminating noise and making changes in the tagset, thus generating MacMorpho version 2 [7] and version 3 [6]. In the first review (version 2), repeated sentences and sentences with missing words were eliminated. There were also tokenization changes – contractions were considered a single token. In the second revision (version 3), more repeated sentences or sentences with missing words were withdrawn, and some PoS tags were simplified, aiming at achieving a coarser tagset: “auxiliary verb”, “relative connective pronoun” and “relative connective adverb” were annotated with the more general labels “verb”, “connective pronoun” and “connective adverb”, respectively. For the present venture, we used the original version of MacMorpho (version 1), annotated with the original tagset. The choice was motivated by practical issues regarding the conversion: the Universal Dependencies scheme undo contractions and contains the tag “auxiliary verb” (`AUX`). Nevertheless, MacMorpho tagset is more granular than the UD tagset, with differences that are not easily circumvented with automatic alignment, thus making the alignment task a source of interesting linguistic challenges, the main one being *past participle* forms (for a detailed version of the conversion, see [22]).

2.1 The Target: Universal Dependencies Tagset

Focusing on multilingual NLP, Universal Dependencies (UD) [16] is a framework for cross-linguistic grammatical annotation that aims at developing a language-independent annotation scheme, flexible for specific extensions of a given language.

The initial UD *PoS* tasgset was proposed by [17], and consisted of 12 part-of-speech categories. The current UD tagset contains 17 labels, plus specific features that can be used to achieve a more fine-grained classification.

The UD 2.0 release contains two corpora for the Portuguese language, and one of these is the UD-Portuguese-Bosque [18]. This corpus, however, is relatively small, with 9,370 sentences and 244,675 words. Although missing (to date) syntactic dependencies, MacMorpho corpus has also been revised, and it is widely used in Portuguese *PoS* training. Besides making more material available in a context that looks promising for cross-lingual NLP, we could also investigate the impact of different tagsets in NLP.

2.2 The Alignment

The alignment of different tagsets is not a purely mechanical task of conversion. The same label can be used for different purposes, and the fact that two corpora are annotated with the same tagset does not guarantee that they are aligned[1]. For example, MacMorpho's tagset has the label `NUM`, used for numerals. The UD tagset has exactly the same tag, standing for the same category. However, while in the UD scheme the orientation is that cardinal numbers should be marked with `NUM`, in the MacMorpho corpus, if the numeral is functioning as the head of a NP, it should be tagged as `NOUN`. The MacMorpho label `PCP` has no direct equivalent in the UD scheme, and along the conversion process, we must choose between VERB or ADJ. In the following section, we specifically address the participles.

Past Participles. In modern Portuguese grammars, there is no doubt that participle forms integrate the class of verbs. However, when we look at the history of grammatical thought, we came across some interesting facts. In the *Téchné grammatiké*, which conveys some seminal ideas of what we mean by grammar, they are an independent class [3]. For the Stoic philosophers (301 B.C.), participles were considered *verbal names*, verbs with cases, among other terms that show their hybrid nature. When parts of speech were translated from Greek into Latin, the participle (*participium*) was named precisely for "participating" in two classes at the same time: nouns and verbs [14]. There is a large number of occasions in which participles present syntactic properties of both adjectives and verbs, making a clear-cut identification nearly impossible (See [22] for a comparison of contemporary Portuguese grammars regarding past participle forms). Sentence 1 bellow illustrate this point:

(1) Refiro-me mesmo àqueles programas de interesse mais geral, como as telenovelas, já que nem essas entram às horas *anunciadas*. (I'm referring to those programs of more general interest, such as soap operas, since neither do they start at the announced/expected hours.)

[1] In this paper, we use the term *alignment* in a broad sense, meaning being equated.

This is not a specific feature of the Portuguese language, as indicated by [10]. In order to validate the linguistic decisions underlying the PCP conversion, we conducted a linguistic experiment aimed at verifying the classification of past participle forms by professionals with solid linguistic background. Not surprisingly, the results showed a huge divergence in classifications. In fact, there were sentences in which half of the volunteers used the VERB label and the other half used the ADJ label. Details of the experiment are presented in [22]. Taking into account the difficulty in deciding how to distribute past participles into VERB or ADJ, we decided to create a second corpus, with a hybrid tagset: UD labels plus PCP[2]. It is worth noting that the PCP label was added to the MacMorpho tagset precisely to avoid the endless discussion among (human) annotators about whether past participles should be annotated as verbs or adjectives.

3 Setting Up the Scene

Corpus and Tagset Conversion. To perform the tagset conversion, we created a set of general rules and a set of specific rules, designed to account for individual cases. When there were directly equivalent tags in the tagsets, the task was simple: we wrote a rule to convert all occurrences of a tag (e.g. PREP) to its correspondent in the other tagset (in this case, ADP). When there was no direct equivalence, the procedure was significantly more laborious. To convert PCP, for example, we first took a sample of 200 cases where the label appeared and read the sentences looking for patterns that could become general conversion rules.

A library in Common Lisp was developed in order to apply the rules. This library is freely available in https://github.com/own-pt/cl-tag-rewriting and it is already incorporated in the CL-CONLLU library [15]. The library produces not only the output data but also some detailed report of the rules applied to each sentence (log files). We have also developed auxiliary functions in the library to analyze the log files producing some statistics about the rules applications that helped us, for instance, to identify superfluous rules.

Since the rules are stored independently of the corpus, it is trivial to recreate the corpus. So, even though we used MacMorpho v1 for conversion, converting the material based on v2 (where sentences were deleted) or v3 is automatic (although in the latter case some minor work on merging labels will be required). In this way, we also make it feasible to study the impact on predictive models of eliminating noise from data, a point that has not yet received the relevance it deserves [19].

In order to perform the two experiments described in the following section, we run the Maximum Entropy model (Generalized Iterative Scaling method), provided by the OpenNLP suite. For both experiments, we merged *train* and *development* partitions.

[2] The UD+PCP corpus, as well as the hybrid tagset, was created with the purpose of serving as a basis for linguistic and computational experiments; and it is not our intention to integrate it into the UD consortium.

4 Experiments

The result of this conversion process is the genesis of three corpora: MacMorpho-UD (MacMorpho corpus annotated with the UD tagset), MacMorpho UD+PCP (MacMorpho corpus annotated with the UD tagset plus the PCP label) and the original MacMorpho.

4.1 First Experiment

We used the Maximum Entropy (MaxEnt) model provided by the OpenNLP suite. For each dataset (UD, UD+PCP and original), we trained in the train+dev partitions and evaluated in the test partition, as provided in the MacMorpho website. Table 1 presents the results according to each tagset.

Table 1. Learning results considering each corpus/tagset

Dataset	MaxEnt accuracy
MacMorpho-UD+PCP	0.9624
MacMorpho-UD	0.9607
MacMorpho	0.9594

The UD+PCP scenario obtained the best results, with a slightly higher performance than the UD tagset. In general, the results point to the success of less granular tagsets, but they also indicate that the criterion of granularity is not absolute: the slight advantage of UD+PCP on UD suggests that, in certain cases, the creation of a class can be a facilitator in learning, bringing consistency. On the other hand, we know that the creation of the PCP label, within the scope of the Lacio-Web project, was due precisely to the lack of agreement in the annotation of past participle forms. Thus, the results suggest that, regarding consistency in learning PoS, the creation of the PCP tag seems to have been an appropriate decision.

In order to verify whether the difference between UD and UD+PCP was due to the ambiguity of the past participle forms, we did an error analysis, starting from the confusion matrix of each scenario. Tables 3, 4 and 5 present the confusion matrix of UD+PCP and UD, respectively (predicted labels are the lines; golden labels are the columns).

As to the UD+PCP scenario, the pairs AUX-VERB (18%), PRON-DET (7%) and ADJ-NOUN (5%) are responsible for most of the confusion. The first error type/confusion (AUX-VERB) comes as no surprise. Auxiliary verbs are also verbs, and the definition of which verbal constructions should be considered auxiliaries may vary not only between languages, but also between grammarians in the same language (compare, for example, the analyses provided by [4,5,20]). It is not a coincidence that one of the changes that took place from version 2 to version 3 in the MacMorpho corpus was the elimination of the distinction between auxiliaries and verbs. The confusion between DET and PRON corresponds to ambiguous words such as *o*, *a*, *os*, *as*, *todos*, *todas*, that can be either PRON or DET, depending on the context.

Finally, the confusion between ADJ and NOUN is an old acquaintance of all students of Linguistics – in this case, we speak of a certain "fluctuation" between nouns and adjectives, which is mainly due to the common practice of naming something from its qualifications.

It is also interesting to note the confusion between PCP and NOUN (almost 4%). There is a vast amount of nouns in Portuguese which result from a lexicalization of participle forms, such as *resultado*. The analysis of a sample of this confusion suggests that this was the case.

However, to elucidate the difference between the tagsets, the most interesting confusion is the one between ADJ-VERB and ADJ-PCP on one hand, and VERB-NOUN and VERB-PCP on the other. None of the four cases stands out with the tagset UD+PCP.

Considering the confusion matrix in the MacMorpho-UD scenario, as expected, the main points of confusion observed in the UD+PCP tagset remained. However, interestingly enough, the confusion matrix also showed a growth in confusion between ADJ, VERB and NOUN. Table 2 compares the classes VERB, NOUN and ADJ in the scenarios with and without the PCP label. The results indicate that the best performance in the tagset UD+PCP is actually the result of the addition of this label, which plays the role of a disambiguator, artificially constructing a consensus where there is none.

Table 2. Comparison of confusion between the scenarios with and without PCP.

Golden PoS	Predicted PoS	Confusion	
		With PCP	Without PCP
VERB	NOUN	188	261
NOUN	VERB	182	271
ADJ	VERB	**53**	**295**
VERB	ADJ	**32**	**287**

4.2 Second Experiment

In a second experiment, we used the MacMorpho-UD corpus to verify the impact of both training size and quality on learning. We created the following test scenarios (in all of them we run the Maximum Entropy model used in Experiment 1):

(A) Training with MacMorpho-UD; evaluation with Bosque-UD test;
(B) Training with the Bosque-UD train; evaluation with Bosque-UD test;
(C) Training with MacMorpho-UD; evaluation with Bosque-UD (complete Bosque-UD).

Table 3. Confusion matrix for Experiment 1, dataset MacMorpho-UD.

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	8869	18	61	0	0	22	0	463	12	0	3	86	0	2	0	287	0
ADP	19	32040	129	6	13	101	0	88	4	0	45	294	0	136	0	40	1
ADV	43	42	5204	1	64	86	4	39	0	0	11	31	0	50	0	13	1
AUX	3	0	0	2737	1	1	0	12	0	0	0	13	0	0	0	298	0
CCONJ	1	1	19	0	4742	0	4	3	0	0	1	90	0	6	0	0	0
DET	21	284	72	1	4	30099	1	25	109	0	249	154	0	13	1	10	0
INTJ	0	0	0	0	1	1	7	0	1	0	0	1	0	0	0	1	0
NOUN	510	98	86	6	0	7	3	39759	41	0	17	654	8	14	0	261	0
NUM	10	2	0	0	0	8	0	86	4671	0	3	50	10	0	1	0	0
PART	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRON	5	2	11	0	2	68	0	15	1	0	3050	28	0	148	0	4	0
PROPN	86	54	16	3	9	7	1	706	34	0	8	20058	1	2	1	30	3
PUNCT	26	4	0	0	2	7	0	27	3	0	6	45	29513	2	1	4	0
SCONJ	2	29	39	0	17	16	0	4	2	0	81	18	0	4588	0	7	0
SYM	0	2	0	0	0	0	0	1	0	0	0	1	1	0	813	0	0
VERB	295	6	10	615	7	10	0	271	6	0	1	80	0	3	0	19228	1
X	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	31

Table 4. Confusion matrix for Experiment 1, dataset MacMorpho-UD+PCP.

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PCP	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	8123	17	59	0	0	20	0	430	14	0	31	2	86	0	2	0	32	0
ADP	21	32048	131	7	13	103	0	87	5	0	1	45	300	0	134	0	37	1
ADV	39	45	5207	1	65	87	4	39	0	0	1	12	32	0	51	0	13	1
AUX	4	0	0	2709	1	1	0	12	0	0	1	0	14	0	0	0	276	0
CCONJ	1	1	19	0	4741	0	4	3	0	0	0	1	92	0	6	0	0	0
DET	19	271	69	1	4	30100	1	24	109	0	7	249	155	0	12	1	8	0
INTJ	0	0	0	0	1	1	7	1	1	0	0	0	1	0	0	0	1	0
NOUN	497	95	85	7	0	7	3	39803	39	0	160	17	659	7	15	0	188	0
NUM	11	2	0	0	0	9	0	85	4671	0	0	3	52	10	0	1	0	0
PART	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PCP	59	2	3	3	0	4	0	83	0	0	3927	0	6	0	0	0	11	0
PRON	3	2	11	0	2	67	0	15	1	0	0	3052	28	0	143	0	2	0
PROPN	83	56	16	3	9	7	1	707	34	0	11	8	20041	1	2	1	32	3
PUNCT	25	4	0	0	2	7	0	26	3	0	3	6	46	29514	4	0	3	0
SCONJ	3	30	39	0	18	16	0	3	2	0	2	79	19	0	4594	0	6	0
SYM	0	2	0	0	0	0	0	1	0	0	0	0	1	1	0	814	0	0
VERB	53	7	7	601	6	4	0	182	5	0	17	1	69	0	1	0	16400	1
X	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	31

Table 5. Confusion matrix for Experiment 1, dataset MacMorpho.

	ADJ	ADV	ADV-KS	ADV-KS-REL	ART	CUR	IN	KC	KS	N	NIL	NPROP	NUM	PCP	PDEN	PREP	PRO-KS	PRO-KS-REL	PROADJ	PROPESS	PROSUB	V	VAUX
ADJ	0	0	8134	56	0	0	4	0	0	0	1	431	0	97	0	29	2	19	16	0	0	0	4
ADV	0	0	40	4376	1	6	1	0	4	71	38	42	0	28	0	1	59	55	87	2	0	1	13
ADV-KS	0	0	0	0	22	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
ADV-KS-REL	0	0	0	3	3	120	0	0	0	0	11	0	0	0	0	0	0	0	0	0	1	0	0
ART	0	0	8	26	1	0	26793	1	0	0	4	27	0	129	100	11	3	251	4	0	0	82	66
CUR	0	0	0	0	0	0	0	484	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
IN	0	0	0	0	0	0	0	0	7	1	0	1	0	2	0	0	0	0	0	0	0	0	0
KC	0	0	4	21	0	0	0	0	4	4769	12	7	0	90	1	0	7	2	0	1	0	2	3
KS	0	0	1	33	20	20	0	0	0	10	1962	6	0	10	1	1	14	23	7	16	105	15	5
N	1	0	531	102	0	0	2	0	3	0	10	41533	1	714	104	152	3	72	4	3	2	1	12
NIL	0	0	0	0	0	0	0	0	0	0	0	2	141	2	0	0	0	0	0	0	0	0	0
NPROP	2	0	82	12	0	0	6	0	1	11	2	729	11	19885	4	11	0	47	1	0	0	1	6
NUM	0	0	3	1	0	0	5	0	0	0	0	103	0	21	3159	0	0	0	0	0	0	0	2
PCP	0	0	60	4	0	0	6	0	0	0	0	75	0	4	0	3938	0	1	0	0	0	0	2
PDEN	0	0	0	43	0	0	0	0	0	10	3	3	0	2	0	0	1062	10	2	0	0	0	2
PREP	0	0	18	148	31	10	98	0	0	20	95	90	0	313	2	1	36	31780	9	0	0	11	35
PRO-KS	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	202	28	0	36
PRO-KS-REL	0	0	0	0	0	5	3	0	0	0	152	0	0	1	0	0	1	2	5	69	1834	0	17
PROADJ	0	0	17	45	0	0	0	0	1	1	3	8	0	10	0	2	5	1	3265	7	3	1	75
PROPESS	0	0	1	0	0	0	3	0	0	1	54	4	0	17	0	0	1	0	1	0	2	2105	2
PROSUB	0	0	2	10	0	0	22	0	0	1	4	10	1	8	0	1	1	3	38	76	8	1	959
V	0	0	44	13	0	0	1	0	0	3	0	174	1	64	0	12	14	6	5	3	0	0	2
VAUX	0	0	3	0	0	0	0	0	0	2	0	8	0	12	0	1	1	0	0	1	0	0	0

This gives us the following three issues for analysis, and the results are in Table 6: 1. The impact of training size on learning, through the comparison between scenario B and experiment 1 (corpus MacMorpho UD). 2. The impact of variation on training and test material, by comparing scenarios A and B. We note that both corpora (MacMorpho and Bosque) are composed by newspaper texts, therefore we do not expect much variation. 3. The variation in the size of the test material, by comparing scenarios A and C.

Surprisingly (for us), the results are mostly affected by the quality of the training and test materials. The surprise comes from the fact that both corpora (MacMorpho and Bosque) are journalistic texts, with overlapping materials.

Table 6. Results for scenario

Scenario	MaxEnt accuracy
A	0.7762
B	0.9504
C	0.7647

The fact that the performance of the model with the MacMorpho-UD corpus was good in experiment 1 (0.9607) and close to that of scenario B (0.9504) discards the hypothesis that the weaker performance in scenarios A and C results from inconsistencies in the annotation of the MacMorpho-UD. In other words: in both experiments 1 and scenario B, we trained and tested with the same corpus (MM-UD and Bosque-UD, respectively) and, in both cases, the results are close and equally good. We could expect internal inconsistencies in MM-UD and Bosque-UD to lead to poor performance, but that was not the case. Furthermore, since the difference between scenario B and scenarios A and C is the variation between training and test material, the difference in results suggests one of the two possibilities: (1) an alignment inconsistency between Bosque-UD and MacMorpho-UD; or (2) the interference of quality of training and test materials – in scenarios A and C the corpus used for training was different from the corpus used for testing. To investigate the alignment hypothesis, we turned again to the analysis of the confusion matrices, looking for patterns that could indicate inconsistencies between the datasets.

The main confusions are distributed in 3 groups: AUX-VERB; PRON-SCONJ and ADJ-NOUN. The confusion between ADJ-NOUN and AUX-VERB repeats the results observed in experiment 1; the only difference is in the new PRON-SCONJ confusion, which might suggest an inconsistency in alignment. Analysis of a sample of divergent cases, however, rejects this explanation. Almost all cases refer to the form *que* (*that*), which is ambiguous between relative pronoun and subordinate conjunction. The (few) remaining errors refer to *quem* (*who*) or *quantos* (*how many*). In none of the cases we noticed systematic errors from the golden corpus that suggested alignment problems.

Regarding the impact of training size on learning, and if we compare only experiment 1 (dataset MacMorho-UD) with scenario B – and assuming that both datasets are equally consistent – we observe a 1% improvement in learning when there is more training material. It is worth remembering that [8], in the context of cross-lingual parsing, indicate that the size of the training corpus ceases to be relevant from a certain amount of data.

Finally, the slightly lower results of scenario C when compared to scenario A suggest that, with more room for evaluation, performance will decline.

5 Side Effect: Optimization of Corpus Revision

Throughout the process of analyzing the confusion matrix, we were faced with confusion data that was not derived from system errors, but from the golden corpus instead (or errors from both system and golden corpus). Besides pointing out to an unfair penalization of systems, such errors served as a strategy for an optimized revision of the MacMorpho corpus annotation. All datasets used in the present study (and made available in https://github.com/own-pt/macmorpho-ud) have already incorporated such revisions.

6 Concluding Remarks

We reported here the results of two experiments aimed at investigating the impact of (i) variation on tagsets, and (ii) the size and quality of dataset training in learning.

The first experiment was empirically and theoretically motivated. The empirical motivation comes from the conversion of the *PoS* annotation of the MacMorpho corpus, which was re-annotated with the Universal Dependencies tagset. It also comes from the fact that there is a lack of an environment that enabled testing with tagsets. From a theoretical point of view, the study is linguistically motivated and takes as its starting point the well-known discussion about past participle forms. As to the second experiment, the results suggest that variation in size may not be as significant as variation in quality.

Error analysis made us realize that much of the systematic confusion reproduces the human divergence in linguistic analysis. The PCP label, the focus of the study, rightly highlights a boundary zone between two classes. At this point, it is worth remembering that the development of post-taggers arises as an engineering response to the problem of explosion due to the ambiguity of classes [21]. But the challenge remains when what is at stake is not ambiguity but "fuzziness". This seems to be the case with past participles. This point relates to what [12] indicates as non-categorical representations, taking as an example the V-ing forms of English, which may be ambiguous between nouns and verbs in the gerund. Such cases would favor the idea of non-discrete classifications in the language, with which we agree. When the linguistic annotation method forces us to use clear categorizations, such as traditional parts of speech classes, it also shows us how limited this practice can be.

Also in regard to tagsets, an important point is to conduct the evaluation in subsequent tasks, such as syntactic dependencies. The idea is to investigate the extent to which it is important to disambiguate the confusions detected for the subsequent NLP tasks.

As additional contributions of this study, we made available to the community two corpora, plus a revised version of MacMorpho v.1, and we encourage the community to repeat our experiments. We also provide the conversion and alignment rules for easily reproduction of the experiments with versions 2 and 3 of MacMorpho. Finally, we indicate that the strategy for error analysis based on confusion matrix seems to be a good way to optimize the linguistic revision. This is a hypothesis we are investigating and in the near future we hope to develop a suite for testing and reviewing tagsets.

References

1. Aluísio, S., Pelizzoni, J., Marchi, A.R., de Oliveira, L., Manenti, R., Marquiafável, V.: An account of the challenge of tagging a reference corpus for Brazilian Portuguese. In: Mamede, N.J., Trancoso, I., Baptista, J., das Graças Volpe Nunes, M. (eds.) PROPOR 2003. LNCS (LNAI), vol. 2721, pp. 110–117. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45011-4_17
2. Aluísio, S.M., Pinheiro, G.M., Finger, M., das Graças V. Nunes, M., Tagnin, S.E.O.: The Lacio-Web project: overview and issues in Brazilian Portuguese corpora creation. In: Proceedings of Corpus Linguistics. UCREL Technical Papers (2003)
3. Auroux, S.: La révolution technologique de la grammatisation: introduction à l'histoire des sciences du langage. Philosophie et langage, Mardaga (1994)
4. Bechara, E.: Moderna Gramática Portuguesa. Nova Fronteira (2012)
5. Cunha, C., Cintra, L.: Nova gramática do português contemporâneo. Obras de referência, Lexikon (2008)
6. Fonseca, E.R., G Rosa, J.L., Aluísio, S.M.: Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. J. Braz. Comput. Soc. **21**(1), 2 (2015). https://doi.org/10.1186/s13173-014-0020-x
7. Fonseca, E.R., Rosa, J.L.G.: Mac-Morpho revisited: towards robust part-of-speech tagging. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (2013)
8. García, M., Gamallo, P.: A rule-based system for cross-lingual parsing of romance languages with universal dependencies. In: CoNLL Shared Task (2017)
9. García, M., Gamallo, P., Gayo, I., Cruz, M.A.P.: PoS-tagging the Web in Portuguese. National varieties, text typologies and spelling systems. Procesamiento del Lenguaje Nat. **53**, 95–101 (2014)
10. Kilgarriff, A., Kosem, I.: Corpus tools for lexicographers. In: Granger, S., Paquot, M. (eds.) Electronic Lexicography, Chap. 3. Oxford University Press (2012)
11. Macklovitch, E.: Where the tagger falters. In: Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation, pp. 113–126 (1992)
12. Manning, C.D.: Computational linguistics and deep learning. Comput. Linguist. **41**(4), 701–707 (2015). https://doi.org/10.1162/COLI_a_00239
13. Medeiros, J.C., Marques, R., Santos, D.: Português Quantitativo. In: Actas do 1° Encontro de Processamento da Língua Portuguesa (escrita e falada) EPLP 1993, pp. 33–38, 25–26 de Fevereiro 1993
14. de Moura Neves, M.: A vertente grega da gramática tradicional uma visão do pensamento grego sobre a linguagem. UNESP (2005)
15. Muniz, H., Chalub, F., Rademaker, A.: Cl-conllu: dependências universais em common lisp. In: V Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana (TILic). Uberlândia, MG, Brazil (2017). https://sites.google.com/view/tilic2017/

16. Nivre, J., et al.: Universal dependencies v1: a multilingual treebank collection. In: Calzolari, N., et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France, May 2016
17. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Chair, N.C.C., et al. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA), Istanbul, Turkey, May 2012
18. Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., de Paiva, V.: Universal dependencies for Portuguese. In: Proceedings of the International Conference on Dependency Linguistics. Pisa, Italy, September 2017
19. Redman, T.C.: If your data is bad, your machine learning tools are useless. Harvard Business Review, 02 April 2018. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless
20. da Rocha Limax, C.H.: Gramática Normativa Da Língua Portuguesa. José Olympio (2010)
21. Santos, D.: POS tagging: clarificaão histórico-terminológica, 29 de Junho - 3 de Julho 2009. http://www.linguateca.pt/Diana/download/SantosEdV2009PoS.pdf
22. Trugo, L.F.: Classes de palavras - da Grécia Antiga ao Google: Um estudo motivado pela conversão de tagsets. Master's thesis, PUC-Rio, August 2016

Dependency Graphs and TEITOK: Exploiting Dependency Parsing

Maarten Janssen(✉)

CELGA-ILTEC, Coimbra, Portugal
maartenjanssen@uc.pt

Abstract. This article describe a set of modules and functions added to the TEITOK corpus environment that turn TEITOK into a full environment for working with dependency parsed corpora, allow parsing document, correcting parsing errors, visualizing parse results, and searching the corpus with a modified version of the CQL query language that can exploit dependency relations.

Keywords: Dependency grammar · TEI/XML · Corpus

1 Introduction

Corpora with dependency relations have become popular in recent years, which is in part due to the fact that for many purposes, and especially for computational purposes, a dependency tree is easier to use than a constituent tree: a dependency tree directly links related words, whereas finding related words in a constituent tree is much less straightforward. And of late, dependency parsers have become significantly more reliable and lighter to use, meaning it is becoming increasingly easy to create a constituent-parsed corpus.

But there are not too many corpus tools that allow you to easily handle dependency relations in corpora. The TEITOK corpus framework [1] has recently incorporated a number of modules and functions dedicated to dependency trees that help to make dealing with dependencies in corpora easier, by providing an online interface to a parser, a visualization module for viewing dependency parsed sentences, an editing mode to correct dependency parses, and a query language to search through the parsed corpus. These modules turn TEITOK into a convenient framework for working with dependency relations, and so in an environment that supports a lot of other types of annotations as well.

This article first gives a short explanation of how dependency relations are stored in the TEITOK file system, then explains the visualization and editing mode, and finally explains the search options provided to search through the corpus. TEITOK is an open-source, browser-based software package, made available as a git repository. More information can be found on the project website: http://www.teitok.org.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 470–478, 2018.
https://doi.org/10.1007/978-3-319-99722-3_47

2 TEITOK Format

A corpus in TEITOK consists of a collection of XML files, in a TEI/XML compliant format. The input consists of files in any TEI compliant format, including formats such as XCES, to which linguistic annotations are added. The use of TEI/XML means a lot of different types of information can be embedded in the file, such as a rich set of metadata, typesetting information, headers, paragraphs, etc. And the wide-spread use of TEI/XML means the data are compatible with many existing corpus projects.

TEI documents in TEITOK are tokenized using in-line mark-up. Rather than the default TEI tag ⟨w⟩, TEITOK uses a non-standard tag ⟨tok⟩ for tokens, to highlight two differences: firstly, TEI distinguishes between words and punctuation marks, while TEITOK treats both as tokens, in line with standard practice in computational linguistics. And secondly, there is a limited set of features allowed for ⟨w⟩, whereas in TEITOK any linguistic annotation can be added to a ⟨tok⟩.

One of the things that can be modeled over tokens are dependency relations. The attributes used for this are based on the CoNLL-U format, and it is easy to convert CoNLL-U files into TEITOK format and vice-versa. To demonstrate this, a small two-word example in provided in Fig. 1, for which the TEITOK equivalent is given in Fig. 2.

Like the CoNNL-U format, the TEITOK format can deal with ranges, i.e. words that belong together, typically multi-word expressions and contractions. In TEITOK, the orthographic word always corresponds to the ⟨tok⟩: for multiword expressions, multiple tokens can be grouped together in an ⟨mtok⟩, whereas

```
# newdoc id = test.txt
# newpar
# sent_id = 1
# text = Good day
1   Good  good ADJ  JJ  Degree=Pos  2  amod  _  _
2   day   day  NOUN NN  Number=Sing 0  root  _  _
```

Fig. 1. Example in CoNLL-U format

```
<TEI>
<teiHeader/>
<text id="test.txt">
<p id="p-1"><s id="s-1">
  <tok id="w-1" lemma="good" upos="ADJ" xpos="JJ"
    feats="Degree=Pos" head="w-2" deprel="amod">Good</tok>
  <tok id="w-2" lemma="day" upos="NOUN" xpos="NN"
    feats="Number=Sing" deprel="root">day</tok>
</s></p>
</text></TEI>
```

Fig. 2. Example in TEITOK format

for contraction, a single token can contain various ⟨dtok⟩. This makes it possible to generate a different verticalized version of the corpus depending on the needs.

The TEITOK distribution comes with a script that automatically creates a CoNLL version of each sentence in a TEITOK/XML file. However, this is not the most typical use in TEITOK since it does not keep any typesetting information; it is more common to create a TEI file by any means, and then tokenize it inside the system and run the parser within the system from the tokenized version. This will export a verticalized version of the corpus, keeping the ID of each token, run the web service parser of UDPIPE [2], and then import the parser data back into the XML file.

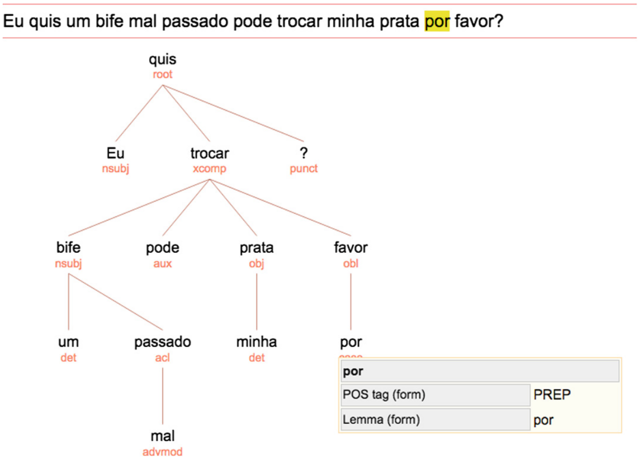
Dependency relations are just one of the things that can be modeled over ⟨tok⟩ nodes: tokens in TEITOK can contains many other types of information as well, such as audio alignment, facsimile alignment, normalizations, error annotation, etc. Contrary to what happens in a tabular design such as the one used in CoNLL, the use of XML makes that all these data can peacefully coexist within a single file, without breaking existing applications.

3 Dependency Trees

Once a file is provided with dependency relations, the TEITOK interface can draw a dependency graph, in either of the two formats typically used for dependency graphs: a sentences with arches above the words, or as a tree. The tree representation is used as the default. In both views, the dependency relation on the link will be displayed with the connecting line; below the node in the case of the tree view, and above the arch in the graph view. Both visualizations are drawn as an SVG image, which can be downloaded either as an SVG, or rasterized to a PNG image to easily include illustrations in an article or website.

One of the problems of the tree view is that it is not easy to reconstruct the original sentence from the tree. To this end TEITOK displays the full sentence above the tree. The sentence is shown in XML format, in which all typesetting information present in the sentence is displayed. And the tree and the sentence view are linked by the ID of the token: when hovering the mouse over a node in the tree, the word in the sentence that corresponds to it will get highlighted, making it easy to see to which word in the sentence each node in the tree corresponds. Furthermore, the other attributes on the token, such as the lemma, the POS tag, the normalized orthography, etc. will be displayed below the node.

A screenshot of this interface is given in Fig. 3, using an example sentence from the COPLE2 learner corpus [3], which was tagged for this example using the UD2.0 language model for Portuguese in UDPIPE. In that figure, you see what happens if the mouse is positioned over the leaf *por* (for) in the right-lower corner of the tree. Firstly, it will display the other relevant features on that token, in this case the lemma and the POS (which in this project does not follow the UD tagset, but rather follows the EAGLES standard). And secondly, it highlights the word "por" in the sentence, making it easy to see this node corresponds to the penultimate word in the sentence.



Eu quis um bife mal passado pode trocar minha prata por favor?

por	
POS tag (form)	PREP
Lemma (form)	por

Fig. 3. A screenshot of the TEITOK dependency tree view

3.1 Editing Mode

Contrary to most other corpus tools, but also for editing corpora. Despite the quality of the UDPIPE output, there are always errors, especially in a learner corpus as in Fig. 3. In TEITOK, editing is done directly in the interface, although editing can of course only be done when logged in as a corpus administrator. This is a different philosophy from corpus tools such as CQPWeb [4], which assumes the corpus to be finalized, or ANNIS [5], which typically relies on external tools for editing.

Editing in TEITOK is typically done manually via the graphical user interface, and not on the basis of rules, as for instance in the case of DepEdit [6]. The base edit mode in TEITOK is to click on any word in the corpus, which will pop-up an HTML form allowing you to change any of the attributes on that token. Which attributes those are is defined in the settings file. This makes it quick and easy to correct for instance the POS tag or lemma for any token.

Dependency relations could be edited in this way, but it would be quite cumbersome. Therefore, the dependency tree module comes with an easy-to-use online editing interface, which works similar to the editing in UD Annotatrix [7]: in order to link a node to a different head, the only thing you need to do is click on the node you want to move, which will show the node in purple to indicate it has been selected. And then you just click on the node you want to reattach it to, after which the tree will be redrawn. And in order to change the edge label, you just click on the label, which will pop-up a list of all the edge labels defined for the corpus, by default using the universal dependencies [8]. And clicking on the new label will replace the label in the tree.

As a web-based interface, changing the tree will not automatically change the XML file - it merely changes the tree on your local computer. In order to save the changes, you need to click the save button, which will submit the new tree to the server, where all the modifications will be made in the corresponding XML file by changing the relevant *head* and *deprel* attributes. Since changes are not made immediately, it is easy to undo the changes by just reloading the page. But saving will store the changes, making it is easy to correct any mistakes you encounter in the automatically generated dependency trees by hand.

4 Dependency Searches

Searching in TEITOK is not done directly on the XML files, but rather the corpus is first exported as a corpus in the Corpus WorkBench (CWB) [9], and all searches are done in that indexed corpus using the CWB Corpus Query Language (CQL). Since version 2, TEITOK comes with its own variant of CQL, which is used in several parts of the system as a replacement for CQP. This C++ application, called TT-CQP, is meant to be used in tandem with TT-CWB-ENCODE, which is an application that reads XML files in the TEITOK/XML format, and produces an indexed corpus in the CWB corpus format. Apart from the standard CWB files, it also writes a couple of additional files that can be used for searches in TT-CQP.

TT-CQP has some additional features such as the option to produce search results in JSON or XML format, and to provide statistical data, for instance providing word-per-million data when grouping results by a structural attribute. Also, instead of producing search results, TT-CQP can give the XML fragment (from the original XML files used to compile the corpus) corresponding to the CQL search results.

But from the perspective of dependency grammar, the most relevant aspect of TT-CQP is the option to use dependent positions, meant primarily to point to the head of a token: for the attribute *head*, TT-CWB-ENCODE stores not only the value of the attribute (which is the ID of the head token), but also the corresponding corpus position. And these corpus positions can then be used in queries to check for positional attributes on the head of a token. To make this easier, names for tokens used in the query are permanent in TT-CQP, meaning you can refer back to them in tabulation, sort, or group commands, as exemplified in (1), where the use of the variable *a* in the global conditions is equal to the use it has in CQP, but the use in the sort option is specific to TT-CQP.

```
(1) a:[word="house"] :: a.upos="NN"; sort a.upos;
```

When we define *head* as a dependent attribute in TEITOK, *head* will function as a reserved name for the head of the first position in a search result (i.e. the head of match). To refer to the head of any other part of the query, you can use a named position between round brackets: `head(a)`. So as a variant on (1) we can use the query in (2) to search for occurrences of *house* in our corpus that have a verbal head, and then sort the resulting matches by the lemma of that head.

```
(2) a:[word="house" & deprel="obj"] :: head(a).upos="VERB";
    sort head(a).lemma;
```

In standard CQL, everything you search for will become part of the range match..matchend. For dependent positions like *head(a)* this is not the case: they are stored separately from the main result match. This makes it possible to use long-distance dependencies without unnecessarily enlarging the resulting match. The head works just like any other position name in CQL, and can be used in subsequent `cat`, `sort`, or `group` commands, and head[1] will refer to the first word to the right of the head, while head(target)[-5]..head(target)[-1] refers to the five words to the left of the head of the target (which in TT-CQP can be either set by using `@[]` as in standard CQL, or by `target:[]` with exactly the same effect).

With dependent position, it becomes possible to search for a wide range of phenomena typically searched for using graph-based search languages like TIGER-search [10]. For example, the query in (2) can be easily converted into the TIGER-search format, where the corresponding query (without the sorting) would be the query in (3).

```
(3) #b:[upos="VERB"] >obj #a:[word="house"]
```

TT-CQP does not (yet) have all the features graph-based languages such as TIGER provide, including the option to have something corresponding to the `>*` command in TIGER. But it does allow for a rich combination of search features. One example explicitly using dependency relations is the query in (4), which in agreement languages searches for Noun-Adjective pairs that do not agree in gender (assuming gender is a pattribute in the corpus, otherwise a slightly more complex expression is needed). It does this independently of the order of the words, or whether the adjective is used predicatively or as a modifier.

```
(4) a:[upos="NOUN"] :: head(a).upos = "ADJ"
        & head(a).gender != a.gender;
```

Agreement errors like that should be rare in typical corpora, but in a learner corpus like COPLE2 they are quite common. A more complex real-life use of dependency relation is given in the next section.

4.1 Word Sketches

TEITOK comes with a module that exploits the dependency relations to create something similar to the word sketches [11] from the SketchEngine [12]. A wordsketch is an overview of the most typical arguments of a word, and initially meant for lexicographic studies: to see what we typically *eat*, or what we typically say about a *house*.

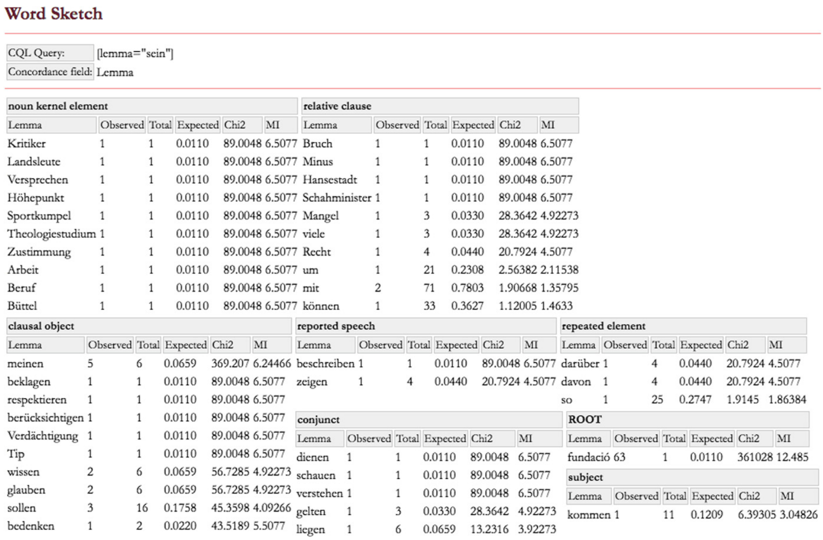
TEITOK uses a simple reinterpretation of this idea to achieve the same effect by using dependency relation: a wordsketch in TEITOK is a set of ordered lists of the strongest collocates of a word, where collocates are heads or daughters (rather than defined by proximity), and the collocates are grouped by their deprel relation. So get the most typical *obj* daughters of *eat*, or the most typical *nmod* daughters of *house*.

To get the relevant data out of the corpus, the wordsketch module uses two simple commands, and have the output produced as JSON. The two commands to get the word sketch for the search-word *casa* organized by lemma are given in (5) and (6).

```
(5) Matches = [lemma="casa"];
       group Matches head.lemma, match.deprel;
```

```
(6) Matches = a:[] :: head.lemma="casa";
       group Matches a.lemma, a.deprel;
```

The command in (4) simply looks for occurrences of the word *casa* (by lemma), and then groups the result by the lemma of the head of the matching results, together with the dependency relation on the match. The command in (5) instead looks for occurrences of tokens that have a head with the lemma *casa*, and groups the result by the dependency relation and the lemma of those occurrences.



Word Sketch

CQL Query:	[lemma="sein"]
Concordance field:	Lemma

noun kernel element

Lemma	Observed	Total	Expected	Chi2	MI
Kritiker	1	1	0.0110	89.0048	6.5077
Landsleute	1	1	0.0110	89.0048	6.5077
Versprechen	1	1	0.0110	89.0048	6.5077
Höhepunkt	1	1	0.0110	89.0048	6.5077
Sportkumpel	1	1	0.0110	89.0048	6.5077
Theologiestudium	1	1	0.0110	89.0048	6.5077
Zustimmung	1	1	0.0110	89.0048	6.5077
Arbeit	1	1	0.0110	89.0048	6.5077
Beruf	1	1	0.0110	89.0048	6.5077
Büttel	1	1	0.0110	89.0048	6.5077

relative clause

Lemma	Observed	Total	Expected	Chi2	MI
Bruch	1	1	0.0110	89.0048	6.5077
Minus	1	1	0.0110	89.0048	6.5077
Hansestadt	1	1	0.0110	89.0048	6.5077
Schahminister	1	1	0.0110	89.0048	6.5077
Mangel	1	3	0.0330	28.3642	4.92273
viele	1	3	0.0330	28.3642	4.92273
Recht	1	4	0.0440	20.7924	4.5077
um	1	21	0.2308	2.56382	2.11538
mit	2	71	0.7803	1.90668	1.35795
können	1	33	0.3627	1.12005	1.4633

clausal object

Lemma	Observed	Total	Expected	Chi2	MI
meinen	5	6	0.0659	369.207	6.24466
beklagen	1	1	0.0110	89.0048	6.5077
respektieren	1	1	0.0110	89.0048	6.5077
berücksichtigen	1	1	0.0110	89.0048	6.5077
Verdächtigung	1	1	0.0110	89.0048	6.5077
Tip	1	1	0.0110	89.0048	6.5077
wissen	2	6	0.0659	56.7285	4.92273
glauben	2	6	0.0659	56.7285	4.92273
sollen	3	16	0.1758	45.3598	4.09266
bedenken	1	2	0.0220	43.5189	5.5077

reported speech

Lemma	Observed	Total	Expected	Chi2	MI
beschreiben	1	1	0.0110	89.0048	6.5077
zeigen	1	4	0.0440	20.7924	4.5077

conjunct

Lemma	Observed	Total	Expected	Chi2	MI
dienen	1	1	0.0110	89.0048	6.5077
schauen	1	1	0.0110	89.0048	6.5077
verstehen	1	1	0.0110	89.0048	6.5077
gelten	1	3	0.0330	28.3642	4.92273
liegen	1	6	0.0659	13.2316	3.92273

repeated element

Lemma	Observed	Total	Expected	Chi2	MI
darüber	1	4	0.0440	20.7924	4.5077
davon	1	4	0.0440	20.7924	4.5077
so	1	25	0.2747	1.9145	1.86384

ROOT

Lemma	Observed	Total	Expected	Chi2	MI
fundació	63	1	0.0110	361028	12.485

subject

Lemma	Observed	Total	Expected	Chi2	MI
kommen	1	11	0.1209	6.39305	3.04826

Fig. 4. A screenshot of the TEITOK word sketch for *sein*

The results of those two commands provides counts for pair of lemma and deprel. For each resulting lemma, the system then looks up the total frequency of that lemma, and uses that in combination with the frequency of the search term (i.e. the word *casa*) and the overall corpus size to calculate both Chi-square and mutual information scores. It then displays the results grouped by *deprel*, and ordered by either Chi-square or MI. The result is illustrated in Fig. 4, showing the word sketch for the word *sein* (*to be* in German) on the CoNLL 2009 corpus for German. Due to the small size of that corpus the word sketch is not very revealing, but it should nevertheless give a good idea about the functionality.

5 Conclusion

TEITOK is an online corpus platform that makes it easy to work with dependency annotated corpora: it not only has an interface to parse texts with the UDPIPE dependency parser and view the results online, but it also allows the easy editing of the resulting dependency trees, and allows searching through the corpus using the dependency relations. And it does all this in a framework that is not uniquely dedicated to dependency parsed corpora, but can also handle a wide array of other linguistic data, such as typographic annotation, time-aligned sound transcription, manuscript-driven document transcription, and stand-off error annotation. As such, TEITOK makes it easy to incorporate dependency parsing into a pipeline for a large number of different types of corpora, making it possible to add syntactic information to corpora that otherwise would not have been provided with such information. And to explore various different annotations on the same documents at the same time within a single.

TEITOK is open-source and can be installed locally on a server. All the files of the corpus are stored as simple XML files, which makes it easy to manipulate the XML files not only inside the TEITOK system, but also for the command line where necessary. And the fact that the files are stored in the TEI/XML format means there is a wide range of available tools to handle the files. Hopefully, this makes TEITOK into a tool that can give a boost to the use of dependency relations in a wide range of smaller corpora.

References

1. Janssen, M.: TEITOK: text-faithful annotated corpora. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, pp. 4037–4043 (2016)
2. Straka, M., Straková, J.: Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, pp. 88–99. Association for Computational Linguistics, August 2017
3. Mendes, A., Antunes, S., Janssen, M., Gonçalves, A.: The COPLE2 corpus: a learner corpus for Portuguese. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France. European Language Resources Association (ELRA), May 2016
4. Hardie, A.: CQPweb - combining power, flexibility and usability in a corpus analysis tool. Int. J. Corpus Linguist. **17**(3), 380–409 (2012)
5. Krause, T., Zeldes, A.: ANNIS3: a new architecture for generic corpus query and visualization. Digit. Scholarsh. Humanit. **31**(1), 118–139 (2016)
6. Zeldes, A., Schroeder, C.T.: An NLP pipeline for coptic. In: Proceedings of the 1st SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) (2016)
7. Tyers, F.M., Sheyanova, M., Washington, J.N.: UD annotatrix: an annotation tool for universal dependencies. In: Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pp. 10–17 (2018)

8. Nivre, J., et al.: Universal dependencies v1: a multilingual treebank collection. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, 23–28 May 2016
9. Evert, S., Hardie, A.: Twenty-first century corpus workbench: updating a query architecture for the new millennium. In: Corpus Linguistics 2011 (2011)
10. Knig, E., Lezius, W.: The TIGER language - a description language for syntax graphs - formal definition, May 2003
11. Kilgarriff, A., Tugwell, D.: Sketching words. In: Corréard, M.H. (ed.) Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins. EURALEX, pp. 125–137 (2002)
12. Kilgarriff, A., et al.: The sketch engine: ten years on. Lexicography **1**, 7–36 (2014)

PassPort: A Dependency Parsing Model for Portuguese

Leonardo Zilio(✉), Rodrigo Wilkens, and Cédrick Fairon

Centre de traitement automatique du langage – CENTAL, Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium
{leonardo.zilio,rodrigo.wilkens,cedrick.fairon}@uclouvain.be

Abstract. Parsers are essential tools for several NLP applications. Here we introduce PassPort, a model for the dependency parsing of Portuguese trained with the Stanford Parser. For developing PassPort, we observed which approach performed best in several setups using different existing parsing algorithms and combinations of linguistic information. PassPort achieved an UAS of 87.55 and a LAS of 85.21 in the Universal Dependencies corpus. We also evaluated the model's performance in relation to another model and different corpora containing three genres. For that, we annotated random sentences from these corpora using PassPort and the PALAVRAS parsing system. We then carried out a manual evaluation and comparison of both models. They achieved very similar results for dependency parsing, with a LAS of 85.02 for PassPort against 84.36 for PALAVRAS. In addition, the results from the analysis showed us that better performance in the part-of-speech tagging could improve our LAS.

Keywords: Dependency parsing · Parsing performance
Universal Dependencies · Parsing for Portuguese

1 Introduction

The processing of Portuguese has evolved much in the past years. We saw new corpora being created, and new tools emerge that came to cover the lack of resources that we formerly had in different areas of language processing. Still, there is some ground to cover, and one of the tools required for processing a natural language, the dependency parsing, has not fared that well compared, for instance, to the state of the art for English (e.g. neural parsers, such as [5,24]).

At the same time, the introduction of the Universal Dependencies (UD) [14], a project that developed freely available, dependency-annotated corpora for multiple languages, presents new corpora for Portuguese, and, coinciding with that, other studies present a series of new state-of-the-art parsing algorithms with a relatively simple training interface.

Supported by the Walloon Region (Projects BEWARE 1510637 and 1610378) and Altissia International.

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 479–489, 2018.
https://doi.org/10.1007/978-3-319-99722-3_48

In this paper, we will be focusing on dependency parsing for the Portuguese language, but we do not aim at conceiving a new parsing algorithm. We took our inspiration from the work of Silva et al. [18] for developing a battery of tests, this time having dependency parsing as main focus and using the Universal Dependency (UD) corpus for Portuguese. Our objective here is thus to test several setups and evaluate their performances with different algorithms. Among the tested algorithms, we selected the one with best performance and compared it with a widely used parsing system for Portuguese. To achieve that, we first directly compared the results of different parsing algorithms in the context of the UD for Portuguese, and, later, we compared the performances across different dependency formalisms. Our hypothesis is that the recent development in dependency parsing task allows for training a model for Portuguese using a black-box approach that outperforms a parser that was deeply customized for a specific language.[1]

This paper is organized as follows: in Sect. 2, we present existing parsing systems and briefly describe their algorithms; Sect. 3 then describes the Universal Dependency corpus for Portuguese that we use as basis for developing our model; in Sect. 4, we present the methodology and results for different models that were trained; in Sect. 5, we compare the best model with the PALAVRAS parsing system by means of a manual evaluation of dependency parsing accuracy; then, in Sect. 6, we make some considerations about the tag sets employed by the different formalisms; lastly, we present our final remarks in Sect. 7.

2 Related Work

Since we are interested in dependency parsing, this section will revolve around the state of the art of dependency parsing. We especially focus on the results for Portuguese of the CONLL-X shared task on Multilingual Dependency Parsing [4]. First, we briefly present parsing algorithms, focusing on those that were used for training a model for Portuguese. We then explore existing dependency parsers for Portuguese.

The approaches presented in CONLL-X may be organized in two categories [9]: graph-based (e.g., the MaltParser [12]) and transition-based (e.g., the MST Parser [8,10] and the Stanford Parser [5]). In terms of algorithms for choosing dependency pairs, the MST Parser uses an online, large-margin learning algorithm [7], MaltParser employs Support Vector Machine, and the Stanford Parser takes advantage of neural network learning [5]. By comparing those three parsing algorithms, the results of Chen and Manning [5] for Chinese and English point to a better performance of the Stanford Parser, followed by the MST Parser. The *CONLL-X 2006* [4] used the Bosque corpus [1] as basis for the Portuguese language, and the LAS of the systems were all above 70. The best results were 87.60 (MaltParser [13]), followed by 86.8 (MST Parser [10]).

[1] The parser model, along with the material that was used in this paper can be found in https://cental.uclouvain.be/resources/smalla_smille/passport/.

Apart from the CONLL shared task, among the existing systems that cover dependency parsing for Portuguese, probably the most well known is the PALAVRAS Parsing System [3]. This system provides full parsing stack, while also annotating semantic information and several other features that can be applied to both the Brazilian and the European variants. The system is based on a Constraint Grammar and reports a performance of 96.9 in terms of LAS in a five-thousand-word sample [3].

Another system that provides dependency parsing for Portuguese is the LX-DepParser[2], which was trained using the MST Parser [8,10] on the CINTIL corpus [2] and reports an unlabeled attachment score (UAS) of 94.42 and a labeled attachment score (LAS) of 91.23.

Finally, Gamallo [6] presented the DepPattern, a dependency parsing system that uses a rule-based finite-state parsing strategy [6,15,16]. Its algorithm minimizes the complexity of rules by using a technique driven by the "single-head" constraint of Dependency Grammar. It was compared with MaltParser using Bosque (version 8). MaltParser achieved an UAS of 88.2 and DepPattern, 84.1.

3 Resources

For training the parser models, we used the Portuguese Universal Dependency (PT-UD) corpus [17][3]. The PT-UD corpus has 227,792 tokens and 9,368 sentences. It was automatically converted from the Bosque corpus [1], which was originally annotated with the PALAVRAS parser [3], and then revised. This corpus contains samples from Brazilian and European Portuguese, and is available in three separate sets: training, test and development.

For testing different setups of dependency parsing for Portuguese, we used different linguistic information and three off-the-shelf parsing systems, which were already introduced in Sect. 2: Stanford Parser 3.8.0 [5], MST Parser 0.5.0 [8], and MaltParser 1.9.1 [12].

4 Dependency Parsing

In this section, we use the resources presented so far in a series of experiments. First, we describe how we organized the setups for the experiments and then we compare the systems among themselves. In the comparison subsection, we first test how much each individual feature contributes to dependency parsing, and then we apply different combinations of these features to train and compare the performance of existing parsing algorithms for Portuguese.

[2] lxcenter.di.fc.ul.pt/services/pt/LXServicesParserDepPT.html.

[3] By the time of the execution of the experiments in this paper, the available PT-UD corpus was in its version 2.1.

4.1 Setup Organization

The first step was to establish different setups that could be used to test the different linguistic information that was available in the corpus. There are four main categories of information available in the PT-UD corpus: surface form, lemma form, short part of speech (short POS), and long part of speech (long POS). The difference between short and long POS reflects the richness of the Portuguese morphology, so that the short POS presents only the word class, while the long POS displays more detailed morphosyntactic information on top of the word class (e.g., person, number, tense). The short POS can normally be automatically derived from the long POS, but there are some ambiguous cases in the corpus[4].

Before going further into the setups, it is important to highlight that we cleaned the long POS field in the corpus, so that all tags that were between angular brackets in the long POS information were deleted, since these represent various types of information that are not always morphosyntactic[5].

From the three systems that were employed for training, all use extensively, per default, the surface and long POS information from the training file, and the Stanford Parser and the MST Parser have an influence of the lemma information[6]. To assure that the parser would receive only the information that we wanted, all information that was not relevant was set to "_" (i.e., underline) in the training, test and development sets. Since the Stanford Parser also uses embedding information during training, we used a model with 300 dimensions[7] that was trained on the brWaC corpus [20–22] using word2vec [11].

4.2 System Comparison

At first, we wanted to observe which of the four main linguistic features contributed the most for the dependency parser accuracy. As such, we tested four setups that contained only one feature (surface, lemma, short POS, or long POS), aiming to evaluate, as a secondary hypothesis, if the addition of morphology has an impact on the dependency parsing task (long versus short POS). Results have shown that the Stanford Parser model was superior in all four individual features, and they ranked from long POS (LAS = 82.74) to short POS (LAS = 79.82), to lemma (LAS = 77.54), and, finally, to surface (LAS = 74.28).

We then followed up with various setups using two features. This time, as we can see in Table 1, it was made clear that, on the morphosyntactic aspect, the long POS is superior to short POS in all setups; however, on the lexical side, the differences in the setups with lemma and surface were not significant

[4] For instance, the tag *DET* in the short POS appears as *DET* or *ART* in the long POS, while the tag *DET* in the long POS appears as *DET* or *PRON* in the short POS.

[5] This modified version of the corpus is available along with the parser model at the PassPort website https://cental.uclouvain.be/resources/smalla_smille/passport/.

[6] We detected some fluctuation in the scores during preliminary testing.

[7] Zeman et al. [23] argue that larger dimensions may yield better results for parsing.

(95% confidence)[8]. We can also see that the Stanford Parser outranks the other two in performance, achieving consistently better scores.

Table 1. Setups using two features as basis (UAS: unlabeled attachment score; LA: label accuracy; LAS: labeled attachment score)

System	Score	Lemma + POS_{short}	Lemma + POS_{long}	POS_{short} + POS_{long}	Surface + POS_{short}	Surface + POS_{long}
Stanford	UAS	85.92	**87.17**	86.28	86.32	**86.90**
	LA	90.80	**92.42**	91.70	90.98	**91.98**
	LAS	83.01	**84.88**	83.53	83.20	**84.29**
MST	UAS	84.57	85.45	85.00	85.19	85.60
	LA	88.54	89.64	89.18	88.96	89.61
	LAS	80.22	81.61	80.85	80.89	81.67
Malt	UAS	84.96	85.29	84.73	84.47	85.09
	LA	88.43	89.39	89.51	88.25	88.95
	LAS	81.59	82.73	81.83	81.15	82.42

Lastly, since the Stanford Parser and the MST Parser do present some fluctuations in the score when lemma information is added to the mix, we created two further setups for these two parsers, both using surface and lemma, but one using only short POS and the other, only long POS. The results have shown that there was no significant difference (with 95% confidence) in any of the measures (UAS, LA, and LAS).

By looking at these results, we can conclude that, in terms of dependency parsing, it is possible to choose one type of lexical information (either surface or lemma) and one morphosyntactic information and it is enough to have good results, but the richer the morphosyntactic information, the better (long POS proved to be significantly better than short POS)[9]. It is also clear that the Stanford Parser yielded the best results for the task, outperforming the other two in all setups that were trained.

After testing this battery of setups, we focused on improving the parser output quality and, for that, we trained a new embeddings model. Up until now, we have been using a model with 300 dimensions, but Chen and Manning [5] suggest using a model of 50 dimensions. So we trained a new embeddings model, by applying word2vec [11] on the raw-text brWaC corpus [20–22], and the results did improve significantly (95% confidence). In Table 2, we present our two previous best setups trained using the new embeddings model, and, in fact, the use of less dimensions proved to be better.

[8] The best system was run five times with randomized train and test sets.

[9] Using the most recent PT-UD corpus (version 2.2) in similar setups, we also had a better performance using long POS information over short POS.

Table 2. Stanford Parser: Two best models using embeddings of 50 dimensions (UAS: unlabeled attachment score; LA: label accuracy; LAS: labeled attachment score.)

System	Score	Lemma + POS_{long}	Surface + POS_{long}
Stanford	UAS	87.48	87.55
	LA	92.12	92.41
	LAS	85.00	85.21

Since the UD presents two corpora for Portuguese (one with only Brazilian Portuguese and the one that we used with both European and Brazilian variants), we also tested the performance of the Stanford Parser on the Brazilian UD corpus (BR-UD)[10]. The BR-UD corpus features only surface and short POS, so we used only these features, and the LAS of the model was 87.30. This corpus yields a better score, but it also has fewer information, and it is dedicated to only one variant of the Portuguese language. For the remainder of this paper, we will refer to our best model that uses surface and long POS from the PT-UD (with LAS of 85.21) as *PassPort* (Parsing System for Portuguese). PassPort is the model that we compare with PALAVRAS in the next section.

5 Parsing: Manual Evaluation

After comparing several parsing models, we wanted to compare the results of PassPort with those from one of the most well-known and customized parsers for Portuguese: the PALAVRAS parsing system [3]. Since both parsers employ different tag sets and formalisms, a direct evaluation of both systems using a single gold standard is not possible. To bridge these two different tag sets and organization of dependency parsing, we designed a manual evaluation using as basis a single corpus of 90 randomly selected sentences from three different genres[11].

The selected genres were literature[12], newspaper articles (from the Diário Gaúcho corpus[13]) and subtitles (from the Portuguese corpus of subtitles compiled by [19]). Thirty sentences were randomly extracted from each of these corpora and all of them were then parsed using PassPort and PALAVRAS. The genres present very different sentence sizes, so here we present the evaluated token account for the three samples: 471 tokens for newspaper, 182 tokens for subtitles, and 642 tokens for literature.

[10] Available at: https://github.com/UniversalDependencies/UD_Portuguese-GSD/tree/master.

[11] Although there are 30 sentences selected from each genre, in the results, it is possible to observe that both parsing systems (PassPort and PALAVRAS) use their own sentence splitters, so that the final sentence numbers are different (for instance, PALAVRAS splits sentences when there is a colon).

[12] Selected romances from www.dominiopublico.gov.br.

[13] This corpus was compiled in the scope of the project PorPopular (www.ufrgs.br/textecc/porlexbras/porpopular/index.php).

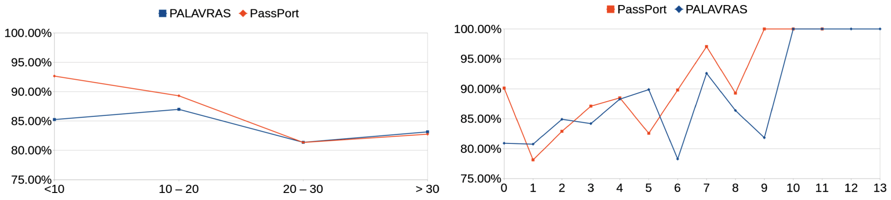
The annotation of both parsers was manually evaluated by one linguist in terms of accuracy (UAS, LA, and LAS), respecting the individual assumptions of each parser (tags, tag order, attachment patterns etc.). The results of the evaluation are shown in Table 3. In the table the results are shown in terms of evaluated tokens[14] and full sentences (sentences in which all tokens were correct for the given measure). The results show that both parsers are very similar in the tested corpus: in terms of tokens, PALAVRAS gets better dependency parsing in the newspaper subcorpus, but PassPort has superior dependency parsing for subtitles and literature and also in the full corpus; in terms of full sentences, PALAVRAS has better results for literature, but PassPort fares better in the full corpus and individually for newspaper articles and subtitles. The differences, however, are small for both sides, and both systems perform very similarly in terms of LAS. In terms of part of speech, PassPort is worse, achieving 94.59% of accuracy against PALAVRAS' 97.53% in the full corpus.

Table 3. Accuracy evaluation of PassPort and the PALAVRAS parsing system (UAS: unlabeled attachment score; LA: label accuracy; LAS: labeled attachment score)

		Newspaper		Subtitles		Literature		Total	
		PassPort	PAL	PassPort	PAL	PassPort	PAL	PassPort	PAL
Tokens	UAS	88.75	**89.56**	**96.70**	90.75	**89.41**	89.36	**90.19**	89.63
	LA	88.32	**91.42**	**92.86**	89.02	**88.79**	87.23	**89.19**	88.97
	LAS	84.93	**87.70**	**92.86**	86.71	**82.87**	81.34	**85.02**	84.36
Sentences	UAS	**43.33**	40.00	**90.32**	68.75	30.00	**43.90**	**54.95**	50.49
	LA	**36.67**	30.00	**70.97**	65.63	26.67	**34.15**	**45.05**	42.72
	LAS	**30.00**	26.67	**70.97**	62.50	16.67	**29.27**	**39.56**	38.83

Following the work of McDonald and Nivre [9], we further investigated the parsing results of the manually evaluated corpus. We start by looking at the labeled attachment score (LAS) in function of the length of the sentences. After dividing the sentences in ranges of evaluated tokens (10, 20, 30+ tokens), we analyzed their mean LAS. The results are shown in Fig. 1a. As we can see, PassPort performed better at lower sentence lengths and was a bit worse in longer sentences (more than 30 words); however, a t-test ($p < 0.05$) reveals that these results are not significantly different. We also evaluated how the deepness of the dependency (i.e., the distance of the token in relation to the root) affects the LAS. The results in Fig. 1b indicate that both parsers perform well even in deeper dependencies.

[14] We did not evaluate punctuation tokens, since PALAVRAS does not provide dependency label for them and, in both parsing models, they are simply attached to the root or the closest dependency to the root.



(a) LAS versus sentence length (b) LAS versus deepness

Fig. 1. Analysis of sentence length and deepness in relation to LAS

6 Discussion

As we could see in Sect. 5, PassPort performs well and is on par with PALAVRAS. Even so, there are some considerations to be made in terms of the dependency tags for both parsers.

Regarding the Universal Dependencies (UD), which were used in PassPort, at least in the PT-UD corpus that was used for training, the tag *obl* is not very informative, since it applies both to adjuncts and to indirect objects introduced by preposition (dative pronouns are tagged as *iobj*)[15]. The UD also present no tag for predicative relations, since the copula verbs are always attached to the predicative (which receives a *root* or a clausal tag). This is much more richly done by PALAVRAS, which presents different tags both for predicatives and for distinguishing indirect objects and adjuncts (but the one for adjuncts doesn't have a good label accuracy – LA – in our corpus: 77.9).

In the case of the tags presented by the PALAVRAS parsing system, the two most frequent tags in our evaluation corpus are *@N* and *@P*[16]. Both of these tags, have a LA higher than 95.4, but they do not describe a dependency relation, they only indicate that the token is attached to a token with a certain part of speech (noun or preposition, respectively). As such, these labels are redundant in the annotation. This is also true for some less frequent tags, such as *@A*, which indicates attachment to an adjective. These cases are better represented in the UD, which presents a label for the relations, and not only the attachment. In addition, PALAVRAS does not consider parataxis, which could pose a problem for annotating oral texts and more freely written language.

7 Final Remarks

In this paper, we trained a new dependency parsing model for Portuguese based on the Universal Dependencies. We used the PT-UD corpus and trained several

[15] This is not in line with the UD guidelines (universaldependencies.org/u/dep/iobj.html), which indicate that the indirect objects should be marked as *obj* (if they are the sole object of the verb) or as *iobj* (if there is another *obj* in the clause). According to the guidelines, *obl* should only be used for adjuncts, but that is not the case in the PT-UD corpus.

[16] The tags present also a < or > symbol, which indicates the attachment direction.

different parsing models based on different lexical and morphological information before selecting the best setup. During the testing phase, we compared three parsing systems (MST, MaltParser, and Stanford Parser) in terms of their performance. Stanford Parser presented the best results in all setups.

After the testing phase, we used our best setup and trained a new parsing model, which we called PassPort. Aiming at observing how PassPort compare to another dependency parser for Portuguese, we compiled a corpus of sentences from different genres, and we then used this common corpus to manually evaluate the accuracy of PassPort against the PALAVRAS parsing system. This evaluation showed that both parsers performed very similarly in terms of the standard parsing scores (unlabeled attachment score, label accuracy, and labeled attachment score). We then ran some further analysis to evaluate the performance of the labeled attachment score in relation to sentence length and deepness of the dependency (distance to the root), and we saw that, here too, both models perform very similarly.

Regarding our hypothesis that the recent development in the dependency parsing task allows for training a model for Portuguese using a black-box approach that outperforms a highly customized parser, we could see that PassPort competes toe to toe with PALAVRAS, having a slight edge on the scores[17].

Overall, PassPort had a performance that is compatible to the state of the art in Portuguese and also in other languages (according to the results of Chen and Manning [5] for English and Chinese using the Stanford Parser). This performance could perhaps be improved if we had delved deeper into the tuning of the parser model, and possibly also if we had dedicated the same attention to the part-of-speech tagging as we dedicated to the dependency parsing model. This remains, however, as a future development of PassPort.

References

1. Afonso, S., Bick, E., Santos, D., Haber, R.: Floresta sintá (c) tica: um “treebank” para o português. quot. In: Gonçalves, A., Correia, C.N., (eds.) Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001), Lisboa 2–4 de Outubro de 2001, Lisboa Portugal: APL (2001)
2. António, B., Castro, S., Silva, J., Costa, F.: Cintil depbank handbook: design options for the representation of grammatical dependencies. Department of Informatics, University of Lisbon, Technical reports nb. di-fcul-tr-11-03, pp. 86–89 (2011)
3. Bick, E.: The Parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus Universitetsforlag (2000)
4. Buchholz, S., Marsi, E.: CoNLL-X shared task on multilingual dependency parsing. In: Proceedings of the Tenth Conference on Computational Natural Language Learning, pp. 149–164. Association for Computational Linguistics (2006)
5. Chen, D., Manning, C.: A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 740–750 (2014)

[17] The model, training datasets and evaluation files will be made available with the final version.

6. Gamallo, P.: Dependency parsing with compression rules. In: Proceedings of the 14th International Conference on Parsing Technologies, pp. 107–117 (2015)
7. McDonald, R., Crammer, K., Pereira, F.: Online large-margin training of dependency parsers. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 91–98. Association for Computational Linguistics (2005)
8. McDonald, R., Lerman, K., Pereira, F.: Multilingual dependency analysis with a two-stage discriminative parser. In: Proceedings of the Tenth Conference on Computational Natural Language Learning, pp. 216–220. Association for Computational Linguistics (2006)
9. McDonald, R., Nivre, J.: Analyzing and integrating dependency parsers. Comput. Linguist. **37**(1), 197–230 (2011)
10. McDonald, R., Pereira, F.: Online learning of approximate dependency parsing algorithms. In: 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
12. Nivre, J., Hall, J., Nilsson, J.: MaltParser: a data-driven parser-generator for dependency parsing. In: International Conference on Language Resources and Evaluation, vol. 6, pp. 2216–2219 (2006)
13. Nivre, J., Hall, J., Nilsson, J., Eryiğit, G., Marinov, S.: Labeled pseudo-projective dependency parsing with support vector machines. In: Proceedings of the Tenth Conference on Computational Natural Language Learning, pp. 221–225. Association for Computational Linguistics (2006)
14. Nivre, J., et al.: Universal dependencies v1: a multilingual treebank collection. In: International Conference on Language Resources and Evaluation (2016)
15. Otero, P.G., González, I.: DepPattern: a multilingual dependency parser. In: International Conference on Computational Processing of the Portuguese Language (PROPOR 2012), Coimbra, Portugal, pp. 659–670. Citeseer (2012)
16. Otero, P.G., López, I.G.: A grammatical formalism based on patterns of part of speech tags. Int. J. Corpus Linguist. **16**(1), 45–71 (2011)
17. Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., de Paiva, V.: Universal dependencies for Portuguese. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling), Pisa, Italy, pp. 197–206, September 2017. http://aclweb.org/anthology/W17-6523
18. Silva, J., Branco, A., Castro, S., Reis, R.: Out-of-the-box robust parsing of Portuguese. In: Pardo, T.A.S., Branco, A., Klautau, A., Vieira, R., de Lima, V.L.S. (eds.) PROPOR 2010. LNCS (LNAI), vol. 6001, pp. 75–85. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12320-7_10
19. Tiedemann, J.: Finding alternative translations in a large corpus of movie subtitle. In: International Conference on Language Resources and Evaluation (2016)
20. Filho, J.A.W., Wilkens, R., Zilio, L., Idiart, M., Villavicencio, A.: Crawling by readability level. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 306–318. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_31
21. Wagner Filho, J., Wilkens, R., Idiart, M., Villavicencio, A.: The brWaC corpus: a new open resource to aid in the processing of Brazilian Portuguese. In: 11th edition of the Language Resources and Evaluation Conference (LREC) (2018)
22. Wagner Filho, J.A., Wilkens, R., Villavicencio, A.: Automatic construction of large readability corpora. In: Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), p. 164 (2016)

23. Zeman, D., et al.: CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 1–19 (2017)
24. Zhou, H., Zhang, Y., Huang, S., Chen, J.: A neural probabilistic structured-prediction model for transition-based dependency parsing. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 1: Long Papers, pp. 1213–1222 (2015)

Effective Sequence Labeling with Hybrid Neural-CRF Models

Pablo da Costa[1](✉) and Gustavo H. Paetzold[2](✉)

[1] Accenture Brazil, São Paulo, Brazil
pablo.da.costa@accenture.com
[2] University of Sheffield, Sheffield, UK
g.h.paetzold@sheffield.ac.uk

Abstract. Sequence tagging models can take many forms, each featuring strong points and limitations. In this contribution, we introduce a hybrid model for sequence tagging that combines recurrent neural networks with conditional random fields. It avoids feature engineering and addresses rare and out-of-vocabulary words by complementing typical word embeddings with compositional character-to-word representations. Using shared parameters across multiple tasks, we are able to achieve performance scores that are either superior or comparable to current state-of-the-art models.

Keywords: Neural networks · Conditional random fields
Sequence labeling · Portuguese

1 Introduction

Sequence tagging is a fundamental problem in natural language processing and can take many forms, such as part-of-speech (POS) tagging and named entity recognition (NER). The goal of sequence tagging is, given a sequence of inputs, to predict a certain type of label for each one of them. Recent contributions focus mostly on neural sequence tagging models, and the great majority of them address exclusively languages like English and Spanish [16,17].

This paper explores the power of hybrid neural models in sequence tagging for the Portuguese language. We tackle three main challenges: addressing rare and out-of-vocabulary words properly without feature engineering, combining neural networks with other types of models in reliable fashion, and creating a multi-purpose model architecture that can achieve state-of-the-art results in many tasks with shared parameters. Multi-purpose architectures explore the fact that different sequence tagging tasks can exploit similar language-specific regularities. For example, models of Portuguese POS tagging and Portuguese NER might benefit from using similar representations for words and other representations.

Currently, there are many examples of effective multi-purpose architectures of neural networks in sequence tagging tasks, like multilingual POS tagging [14,16], NER [14,16] and chunking [16]. Collobert et al. propose a neural network

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 490–498, 2018.
https://doi.org/10.1007/978-3-319-99722-3_49

architecture that is used in four natural language processing tasks: POS tagging, named entity recognition, semantic role labeling and chunking. They train word representations on large amounts of unannotated texts from Wikipedia, then tune the pre-trained word representations for each individual task. More recently, other contributions achieved interesting results using new deep learning techniques, like adversarial training [17], bi-directional recurrent models [14], log-linear models with deep learning [16], and character-level embedding representations [6].

In this work, we combine powerful elements from these successful architectures in order to develop a hybrid multi-purpose model that can achieve state-of-art results for Portuguese sequence tagging tasks without using feature engineering or pipelines. Our model handles multi-purpose sequence tagging tasks by simply sharing the network architecture and model parameters across them. In order to address the challenge of handling rare and out-of-vocabulary words, we combine both typical word representations with compositional character-to-word embeddings. We also maximize the performance of our model by combining consolidated recurrent neural networks with conditional random fields (CRF), creating a hybrid architecture.

We describe our model and experiments in what follows.

2 Model

Our sequence labeling model is inspired by the architecture proposed by Plank et al. Much like any other sequence labeler, our model takes as input a sentence, and produces as output a label for each word in the sentence. The label set depends, of course, on the task being addressed. Its architecture, which is illustrated in Fig. 1, is composed of four main components:

- a bidirectional LSTM that produces a character-to-word compositional representation of each word;
- a word embedding layer that produces a typical word-level representation of each word;
- a bidirectional LSTM that combines these two representations and produces a final encoding for each word; and
- a CRF layer that predicts labels based on the final word representations.

2.1 Bi-LSTMs for Sequence Representation

Long short-term memory (LSTM) Hochreiter and Schmidhuber [7] is a type of recurrent neural network that is able to learn interdependencies from sequential data. LSTMs address the vanishing gradient problem by using a series of switches that learn how to balance the influence of encoded and new information during the training process. In other words, the gates control how the network's memory of old information is backpropagated in the hidden states of the model.

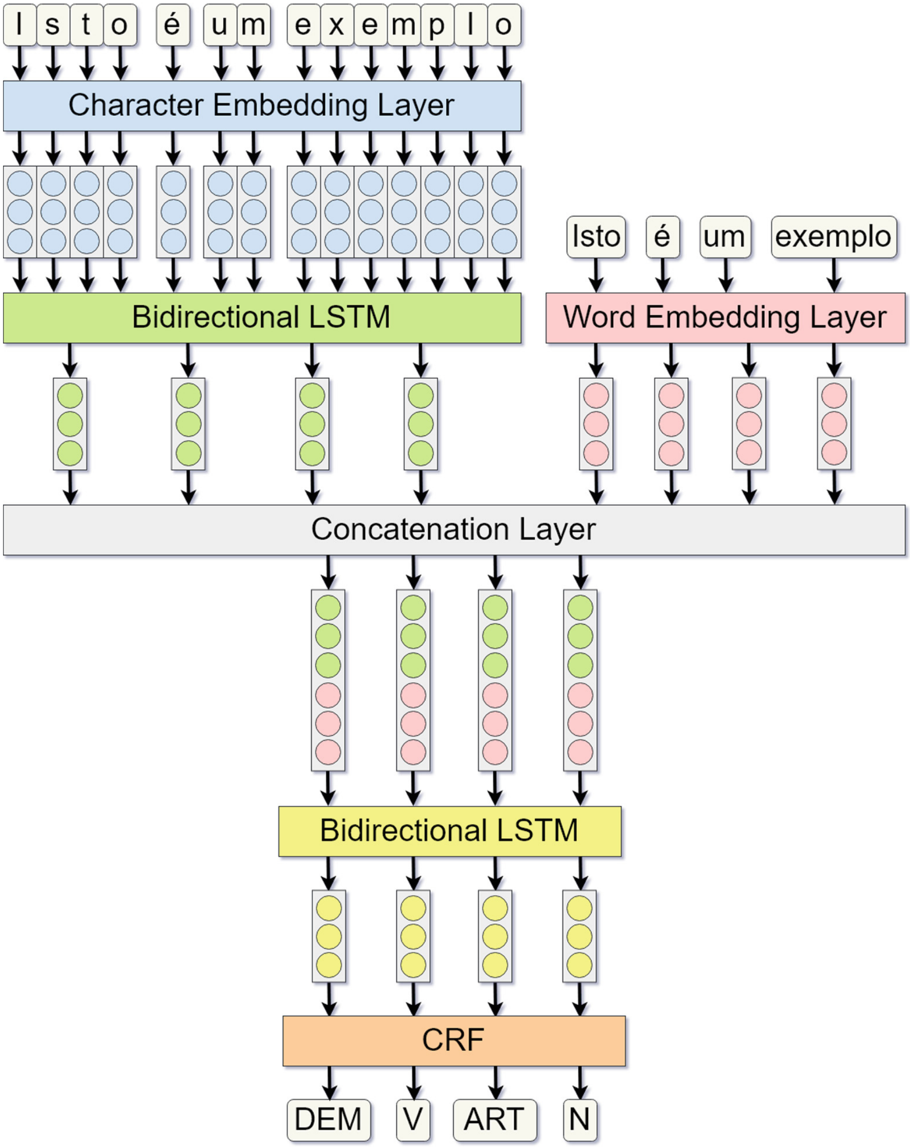


Fig. 1. Hybrid neural-CRF model architecture

Bi-directional LSTMs (Bi-LSTMs) today are used in many NLP tasks to scan and learn both left-to-right and right-to-left dependencies, which can capture complementary types of information from the input. The hidden representations produced by both LSTMs (left and right) can be linearly combined (θ) to form a final representation $h_t = h_t^{\leftarrow}\ \theta\ h_t^{\rightarrow}$.

Bi-LSTMs have been successfully applied in many NLP tasks, such as dependency parsing Kiperwasser and Goldberg [8,9]; name entity recognition Yang et al. [16]; chunking Yang et al. [16], and dozens of others.

We use a Bi-LSTMs for two different purposes: to transform each sequence of encoded characters into a single compositional word representation, and to combine the sequences of typical and compositional word embeddings into a sequence of inputs for our CRF model.

2.2 Character-to-Word Embeddings for Reliability

In many NLP tasks, sub-word information has been shown to achieve great improvements in terms of word representation, specially for models that transform embedding representations of characters into representations of words. This is mainly because of two reasons:

- **Character-to-word embeddings address the out-of-vocabulary problem:** When only a typical word embeddings model is used, every word that is not present in the initial vocabulary over which the model was trained will be represented by a generic "missing word" vector. This can compromise the quality of a prediction, since the model has no access to any form of information about the missing word in question, regardless of how obvious its meaning is, or how similar it is to other words in the vocabulary. These embeddings are specially useful for numbers, since regardless of how large a training corpus is, it will definitely not feature occurrences of each and every conceivable number.
- **Character-to-word embeddings can create a more reliable representation for under-represented non-missing words:** Since typical word embedding models require for large amounts of text, it is very common for one to combine corpora from different, often unreliable sources in order to create enough volume. Because of that, one can often find under-represented misspellings, abbreviations or even rare morphological variants of certain words in these corpora. Because they do not appear very frequently in the corpus, the reliability of their representations tends to be compromised. Character-to-word embeddings address this problem by exploiting the information inherent to character sequences, allowing for under-represented words to "borrow" information from similar, more frequently occurring variants.

We complement our character-to-word embeddings with typical word embeddings, which have also been demonstrated to produce powerful semantic representations for frequently occurring words [11]. By exclusively using automatically learned embeddings as input for prediction, we impart resource efficiency to the model, since it requires no engineered features from external resources to operate.

2.3 Conditional Random Fields for Prediction

After we use our Bi-LSTMs to produce a hidden representation of each word in the sentence, we are able to predict their labels. A simple way of doing so would be by applying a linear transformation to the hidden representations that yields one value for each possible label in our tagset, then applying a softmax function over these values so that we get a probability distribution over the labels. We, however, take a different approach.

Instead of said transformation, we input the sequence of hidden representations to a CRF model Lafferty et al. [10], which, as demonstrated by [17], can be an effective approach to learning interdependencies between the labels in a sequence.

We first use the CRF to produce a probability[1] distribution over the tagset. With a probability distribution at hand, we employ the Viterbi algorithm to decode the optimal output sequence of labels for the sentence.

3 Experimental Setup

We address two sequence labeling tasks for the Portuguese language: named-entity recognition and POS tagging. For named-entity recognition, we use the HAREM corpus presented by [3]. This corpus is annotated with ten named entity categories: Person (PESSOA), Organization (ORGANIZACAO), Location (LOCAL), Value (VALOR), Date (TEMPO), Abstraction (ABSTRACCAO), Title (OBRA), Event (ACONTECIMENTO), Thing (COISA) and Other (OUTRO). In our experiments, we use the original HAREM corpus as the training set and the MiniHAREM corpus as the test set. This is the same setup used by dos Santos and Guimarães in their evaluation. Additionally, we sperated a development-set for tuning using 5% of the training set. We chose the HAREM corpus because it is the most widely used for this task. Table 1 describes some statistics of our setup for named-entity recognition and PoS tagging.

For POS tagging we used the universal dependencies corpus v1.2 [13] the canonical data splits and a tag-set containing 17 different POS tags. We chose this corpus because it is very well described and structured in comparison to others, and it is also the most frequently used corpus in recent contributions [13,15].

Table 1. Table defining the used corpus.

Corpus	Train		Test	
	Sentences	Tokens	Sentences	Tokens
HAREM I	4,749	93,125	3,393	62,914
UD 1.2	20,200	242,702	2,244	54,777

[1] The log probability learned from the CRF'layer is backpropagated via cross-entropy.

We configure our hybrid model in the following way:

- **Character-to-word embedding size:** 600
- **Typical Word embedding size:** 300
- **LSTM hidden layer size:** 300
- **Character LSTM hidden layer size:** 300
- **Number of LSTM layers (character and sentence level):** 2
- **Dropout proportion:** 0.5
- **Learning method:** Adam
- **Learning rate:** 0.001
- **Maximum number of epochs:** 300 (using early stop with 5 epochs).
- **Mini-Batch size:** 20

In order to highlight the role of the most important components of our neural-CRF model[2], we trained three different variants of it:

- **Full:** The model illustrated in Fig. 1, with all of its components.
- **Word+CRF:** A model that uses only typical word embeddings as input, passes them through a bidirectional LSTM, then predicts labels through a CRF model.
- **Word-NoCRF:** Identical to Word+CRF, except instead of a CRF model, it applies a soft-max function over the output produced by the bidirectional LSTM.

The evaluation metrics we use are:

- **Accuracy:**
$$\frac{total\_correct\_predictions}{total\_predictions} \times 100 \tag{1}$$
- **Weighted average of the Precision of each class:**
$$\frac{correct\_class\_predictions}{total\_class\_predictions} \tag{2}$$
- **Weighted average of the Recall of each class:**
$$\frac{correct\_class\_predictions}{total\_class\_labels} \tag{3}$$
- **Weighted average of the F-score of each class:**
$$2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

4 Results

In this section, we discuss the results obtained with the aforementioned models.

[2] For pre-trained word embeddings we used the ones in https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md.

4.1 Named Entity Recognition

Table 2 showcases the results obtained for the Portuguese named-entity recognition task. As it can be noticed, the performance of our complete model (Full) surpasses the former state-of-the-art approach of dos Santos and Guimarães [5] by 3.73% at F-score. We can also observe a substantial increase in Recall, by adding extra contextual features from the modeling of character level information and by it's linear combination with the pre-trained word embeddings.

Also, as we can see in the Table 2, the use of long range contextual information from the CRF layer was crucial to improve our model's performance.

Table 2. Table comparing our model results at HAREM corpus against state of art results.

Model	Accuracy	F-score	Precision	Recall
Full	94.01	69.14	68.95	69.34
Word+CRF	91.67	60.56	66.83	55.36
Word–NoCRF	90.21	51.61	52.85	50.43
dos Santos and Guimarães [5] - CNN	-	65.41	67.16	63.74

4.2 POS Tagging

The results obtained for POS tagging are featured in Table 3. As we can see, the performance of our approach considerably increases as we move from the simpler variants (Word+CRF and Word-NoCRF) to the complete version (Full). This suggests that our character-to-word embeddings not only address the problem of out-of-vocabulary and under-represented words, but it also helps the CRF layer in performing more reliable predictions.

Table 3. Table comparision of our model results at universal dependencies corpus version 1.2 against state of art results.

Model	Accuracy	F-score	Precision	Recall
Full	97.87	97.58	97.52	97.64
Word+CRF	94.70	94.06	94.11	94.01
Word-NoCRF	94.70	94.10	94.20	94.00
Yasunaga et al. [17] - AD	98.07	-	-	-
Yasunaga et al. [17] - NOAD	97.94	-	-	-
Plank et al. [14] - BILSTM	97.90	-	-	-
Plank et al. [14] - TNT	96.27	-	-	-
Plank et al. [14] - CRF	96.32	-	-	-
Berend [2]	95.50	-	-	-
Nguyen et al. [12]	97.50	-	-	-

In addition to that, our model offers another advantage over some of the other approaches compared. The approach of [12] achieves its high Accuracy by employing multilingual word embeddings, which can be hard to induce and require resources that go beyond just raw text [1]. Our model, however, requires only for raw text in Portuguese in order for the embeddings to be trained.

Finally, in comparison to other models from literature, our Full model performed very well overall, achieving Accuracy scores only 0.20% short of the former state-of-the-art approach of dos Santos and Guimarães [17]. We hypothesize that this small difference may be due simply to the use of different types of word embedding models.

5 Conclusions

In this paper, we presented a multi-purpose hybrid sequence labeling model for Portuguese that combines recurrent neural networks and conditional random fields. As input, it uses both typical word embeddings and character-to-word compositional embeddings, so that it can enrich the representation of both in and out-of-vocabulary words. We applied our approach to two tasks: named entity recognition and POS tagging.

Both experiments revealed that our model offers performance scores that are either superior or comparable to former state-of-the-art approaches. We also found that incorporating both character-to-word embeddings and a CRF layer into the model yields considerably better results than using just typical word embeddings and recurrent neural networks.

Overall, we can conclude that our model is a powerful solution to sequence labeling tasks for Portuguese: it offers state-of-the-art performance, requires no engineered features, and effectively handles rare and out-of-vocabulary words.

In future work, we plan to investigate other character-to-word compositional models for embeddings, and intend to apply our model to a wider variety of tasks and languages. Our models and code were implemented using Tensorflow 1.8 and can be founded here: https://bitbucket.org/pablocosta/deepnlptoolkit.git.

References

1. Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., Smith, N.A.: Massively multilingual word embeddings. arXiv preprint arXiv:1602.01925 (2016)
2. Berend, G.: Sparse coding of neural word embeddings for multilingual sequence labeling. arXiv preprint arXiv:1612.07130 (2016)
3. Cardoso, N., Santos, D.: Directivas para a identificação e classificação semântica na colecção dourada do harem (2007)
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**, 2493–2537 (2011)
5. dos Santos, C.N., Guimarães, V.: Boosting named entity recognition with neural character embeddings. CoRR, abs/1505.05008 (2015). http://arxiv.org/abs/1505.05008

6. Dos Santos, C.N., Zadrozny, B.: Learning character-level representations for part-of-speech tagging. In: Proceedings of the 31st International Conference on International Conference on Machine Learning, ICML 2014, vol. 32, pp. II-1818–II-1826. JMLR.org (2014). http://dl.acm.org/citation.cfm?id=3044805.3045095
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Kiperwasser, E., Goldberg, Y.: Easy-first dependency parsing with hierarchical tree LSTMs. CoRR, abs/1603.00375 (2016). http://arxiv.org/abs/1603.00375
9. Kiperwasser, E., Goldberg, Y.: Simple and accurate dependency parsing using bidirectional LSTM feature representations. CoRR, abs/1603.04351 (2016). http://arxiv.org/abs/1603.04351
10. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001, San Francisco, CA, USA, pp. 282–289. Morgan Kaufmann Publishers Inc. (2001). http://dl.acm.org/citation.cfm?id=645530.655813. ISBN 1-55860-778-1
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. **26**, 3111–3119 (2013)
12. Nguyen, D.Q., Dras, M., Johnson, M.: A novel neural network model for joint POS tagging and graph-based dependency parsing. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 134–142 (2017). http://www.aclweb.org/anthology/K17-3014
13. Nivre, J., et al.: Universal dependencies 1.2 (2015)
14. Plank, B., Søgaard, A., Goldberg, Y.: Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. arXiv preprint arXiv:1604.05529 (2016)
15. Tsarfaty, R., Seddah, D., Kübler, S., Nivre, J.: Parsing morphologically rich languages: introduction to the special issue. Comput. Linguist. **39**(1), 15–22 (2013)
16. Yang, Z., Salakhutdinov, R., Cohen, W.W.: Multi-task cross-lingual sequence tagging from scratch. CoRR, abs/1603.06270 (2016). http://arxiv.org/abs/1603.06270
17. Yasunaga, M., Kasai, J., Radev, D.: Robust multilingual part-of-speech tagging via adversarial training. arXiv preprint arXiv:1711.04903 (2017)

Author Index