Recognizing Textual Entailment and Paraphrases in Portuguese

Gil Rocha(✉) and Henrique Lopes Cardoso

LIACC/DEI, Faculdade de Engenharia,
Universidade do Porto, rua Dr. Roberto Frias, 4200-465 Porto, Portugal
{gil.rocha,hlc}@fe.up.pt

Abstract. The aim of textual entailment and paraphrase recognition is to determine whether the meaning of a text fragment can be inferred (is entailed) from the meaning of another text fragment. In this paper, we address the task of automatically recognizing textual entailment (RTE) and paraphrases from text written in the Portuguese language employing supervised machine learning techniques. Firstly, we formulate the task as a multi-class classification problem. We conclude that semantic-based approaches are very promising to recognize textual entailment and that combining data from European and Brazilian Portuguese brings several challenges typical with cross-language learning. Then, we formulate the task as a binary classification problem and demonstrate the capability of the proposed classifier for RTE and paraphrases. The results reported in this work are promising, achieving 0.83 of accuracy on the test data.

1 Introduction

Recognizing Textual Entailment (RTE) [8] in natural language text is a task seeking to find entailment relations between text fragments. Given two text fragments, typically denoted as 'Text' (T) and 'Hypothesis' (H), RTE is the task of determining whether the meaning of the Hypothesis (H, *e.g.* "Joe Smith contributes to academia") is entailed (can be inferred) from the Text (T, *e.g.* "Joe Smith offers a generous gift to the university") [28]. In other words, a sentence T entails another sentence H if after reading and knowing that T is true, a human would infer that H must also be true.

We may think of textual entailment and paraphrasing in terms of logical entailment ($\models$) [4]. If the logical meaning representations of T and H are Φ_T and Φ_H respectively, then $\langle T, H \rangle$ corresponds to a textual entailment pair if and only if $(\Phi_T \wedge B) \models \Phi_H$, where B is a knowledge base containing postulates that correspond to knowledge that is typically assumed to be shared by humans (*i.e.* common sense reasoning and world knowledge). Similarly, if the logical meaning representations of text fragments T_1 and T_2 are Φ_1 and Φ_2 respectively, then T_1 is a paraphrase of T_2 if and only if $(\Phi_1 \wedge B) \models \Phi_2$ and $(\Phi_2 \wedge B) \models \Phi_1$.

It is well known that writers tend to avoid repetition of words (*e.g.* making use of different referring expressions) and omit implicit knowledge in order

E. Oliveira et al. (Eds.): EPIA 2017, LNAI 10423, pp. 868–879, 2017.
DOI: 10.1007/978-3-319-65340-2_70

to obtain a more fluent reading experience and capture a reader's attention. Writers often appeal to commonsense knowledge and inferring capabilities they assume the target reading audience to have, to convey information about the world. These assumptions turn out to pose very difficult challenges to computational systems aiming to automatically process and reason about information expressed in natural language texts. Furthermore, this phenomena is often associated with ambiguity presented in text written in natural language. Taking into account the characteristics of natural language text previously presented, the *NLP* community typically adopts a relaxed definition of textual entailment [4], so that T entails H if a human knowing that T is true would be expected to infer that H must also be true in a given context. A similar relaxed definition can be formulated for paraphrases.

RTE has been recently proposed as a general task that captures major semantic inference needs in several NLP applications [4,7], including question answering [22], information extraction [21], document summarization [19], machine translation [23] and argumentation mining [18,26].

Between 2004 and 2013, eight *RTE Challenges* [6] were organized aiming to provide concrete datasets that could be used by the research community to evaluate and compare different approaches. However, RTE from Portuguese text remains little explored. Recently, at the *PROPOR 2016* international conference, the ASSIN ("Avaliação de Similaridade Semântica e Inferência Textual") challenge was proposed [12]. This challenge introduced a corpus annotated for the semantic similarity and textual inference tasks from text written in Portuguese, providing the necessary resources for the development of NLP systems using machine learning (ML) techniques to address this challenging task.

In this paper, we aim to explore different approaches to address the task of recognizing textual entailment and paraphrases from text written in the Portuguese language, using supervised ML algorithms.

This paper is structured as follows: Sect. 2 presents related work on recognizing textual entailment and paraphrases, focusing approaches based on text written in the Portuguese language. Section 3 introduces the corpus that was used in our experiments to validate the approach presented in this work. Section 4 describes the methods that were used to address the task of recognizing textual entailment and paraphrases using supervised machine learning algorithms. Section 5 presents the results obtained by the system described in this paper. Finally, Sect. 6 concludes and points to directions of future work.

2 Related Work

State-of-the-art systems for RTE and paraphrase in natural language text typically follow a supervised machine learning approach. These systems rely on heavily engineered NLP pipelines, extensive manual creation of features, several external resources (e.g. WordNet [10]) and specialized sub-components to address specific auxiliary sub-tasks [4,7,27], such as negation detection, semantic similarity and paraphrase detection [5,9,16]. Existing approaches differ mainly

on the initial assumptions and specific goals. In [4], the authors divided these systems in two main dimensions: (a) whether they focus on *paraphrasing* or *textual entailment* between text fragment pairs, and (b) whether they perform *recognition*, *generation* or *extraction* of paraphrases or textual entailment pairs. Since, in this paper, we focus on the recognition of paraphrase and textual entailment between each pair of sentences, the remainder of this section will focus on related work for this specific task. The main input given to a paraphrase or textual entailment recognizer is a pair of sentences, possibly in a particular context. The desired output is a (probabilistic) judgment, indicating whether or not the text fragments are paraphrases or a textual entailment pair.

For English text several challenges have been proposed, namely the RTE Challenges [6], SICK [20] and STS at SemEval [1].

The ASSIN challenge [12] follows similar guidelines and introduces the first corpus containing entailment and semantic similarity annotations between pairs of sentences in two Portuguese variants, European and Brazilian, suitable for the exploration of supervised machine learning techniques to address these tasks. To the best of our knowledge, the best ML approaches for RTE and paraphrases in Portuguese texts are presented in the ASSIN challenge. In [15], Hartmann followed the supervised machine learning paradigm with an approach based on the cosine similarity of the vectorial representation of each sentence. These sentence representations were obtained from the sum of the vectors representing each word in a sentence using two language models: *TF-IDF* and *word2vec*. Then, Hartmann computes cosine similarity metrics for each pair of sentences from the two representations (*TF-IDF* and *word2vec*) and uses them as features that are given to a linear classifier.

Fialho *et al.* [11] extracted several metrics for each pair of sentences, namely edit distance, words overlap, *BLEU* [24] and *ROUGE* [17], amongst others. They reported several experiments considering different preprocessing steps in the NLP pipeline, namely: original sentences, removing stop-words, lower-case words and clusters of words. A feature set containing more than 90 features to represent each pair of sentences was used as input for a *SVM* classifier. Fialho *et al.* also reported experiments merging the original ASSIN corpus with annotated data from the SICK corpus translated from English to Portuguese. They added 9191 examples from the *SICK* corpus to the 6000 examples from the ASSIN training set in one of their experiments. The results reported on the augmented version of the training data were worst than the results reported on the original training data. The authors associate these results to translation errors that were probably made during the process. In addition, they trained their model in one of the Portuguese variants of the ASSIN corpus and evaluated the performance of the model in the other Portuguese variant. Reported results following this experimental setup were worst when compared with the model trained and tested in the same variant, but were better than the results obtained in the augmented version of the original dataset (with the *SICK* data). They obtained the best results for recognizing textual entailment in the ASSIN challenge: 0.843 of accuracy and 0.66 of macro F1-score.

In [3], Alves *et al.* explored two different approaches for RTE and paraphrases: a supervised ML approach ("Reciclagem" system) and a heuristic-based approach ("ASAPP" system). The "Reciclagem" system is based on lexical and semantic knowledge that calculates the similarity and relations of two sentences without any kind of supervised machine learning methods. This system was used as a baseline for the "ASAPP" system and to evaluate the quality of different lexical and semantic resources for Portuguese. The "ASAPP" system follows the supervised ML approach and adds to "Reciclagem" features based on the syntactic and structural information extracted from the pair of sentences, such as: number of tokens, overlapping words, synonyms, hyperonyms, meronyms, antonyms and number of words with negative connotation, type of named entities, amongst others. In their experiments, the authors explored different strategies to divide the training data, to combine results from different classifiers and several feature selection techniques. They reported 0.731 of accuracy and 0.43 of macro F1-score on the European-Portuguese test data.

3 Data

A corpus with sentence pairs labeled with the type of relation (*Entailment*, *Paraphrase* or *None*) is an important requirement in order to address the task of recognizing textual entailment and paraphrases using supervised ML techniques. The ASSIN corpus [12] is, to the best of our knowledge, the first corpus annotated with pairs of sentences written in Portuguese that is suitable for this task. The corpus contains pairs of sentences extracted from news articles written in European-Portuguese (EP) and Brazilian-Portuguese (BP), obtained from *Google News* Portugal and Brazil, respectively.

The ASSIN challenge [12] included two tasks, both using the ASSIN corpus: (a) semantic similarity and (b) textual entailment and paraphrase recognition. We will focus on the latter: the "entailment" label is the attribute that will be used as target label for the proposed task.

In total, the ASSIN corpus contains 10.000 pairs, half in each of the Portuguese variants. The distribution of $\langle T, H \rangle$ pairs between each "entailment" label and between texts written in BP and EP is shown in Table 1. It is important to notice that the corpus is unbalanced in relation to the "entailment" and "paraphrase" labels. This can bring some issues that should be taken into account.

Table 1. Distribution of labels in ASSIN corpus.

	BP		EP	
Label/partition	*Train*	*Test*	*Train*	*Test*
None	2331	1553	2046	1386
Entailment	529	341	729	481
Paraphrase	140	106	225	133

The inter-annotator agreement metrics related to this corpus are the following: *Fleiss's* $\mathcal{K}$ of 0.61 and Concordance of 0.8. The *Fleiss's* $\mathcal{K}$ value is relatively low, demonstrating the subjectivity associated with the annotation process [12]. However, these values are not very different from the values reported in other corpora used for the same task: for instance, in the RTE Challenges the values ranged from 0.6 in the first RTE Challenge to 0.75 or more in the following challenges [6,12].

Table 2 shows one example of the content and annotations available in the ASSIN corpus for each of the labels.

Table 2. Annotated examples from the ASSIN corpus (extracted from [12]).

Label	Pair of Sentences
None	As apostas podem ser feitas até as 19h (de Brasília). (T)
	As apostas podem ser feitas em qualquer lotérica do país. (H)
Entailment	Como não houve acordo, a reunião será retomada nesta terça, a partir das 10h. (T)
	As partes voltam a se reunir nesta terça, às 10h. (H)
Paraphrase	Vou convocar um congresso extraordinário para me substituir enquanto presidente. (T)
	Vou organizar um congresso extraordinário para se realizar a minha substituição como presidente. (H)

4 Methods

We here describe the approach we followed to address the task of entailment and paraphrase recognition from natural language Portuguese text. We formulate the problem following two different settings. First, as a multi-class classification problem, in which we aim to classify each $\langle T, H \rangle$ with one of the labels *Entailment* (if $T \models H$), *Paraphrase* (if $T \models H$ and $H \models T$, *i.e.*, if T is paraphrase of H), or *None* (if T and H are not related with one of the previous labels). Second, as a binary classification problem, aiming to distinct each $\langle T, H \rangle$ with one of the labels *Entailment* or *None*. We employed supervised ML techniques given a set of annotated data, the ASSIN corpus.

To transform each sentence into the corresponding set of tokens and to obtain for each token the corresponding lemma and part-of-speech information (including syntactic function, person, number, tense, amongst others) we used the *CitiusTagger* [13] NLP tool. This tool includes a named entity recognizer trained in natural language text written in Portuguese.

Several experiments were made using different NLP techniques to process the sentences received as input: removing stop-words, removing auxiliary words (*i.e.* words relevant for the discourse structure but not domain specific, such as:

prepositions, determiners, conjunctions, interjections, numbers and some adverbial groups) and lemmatization. Transforming each token in the corresponding lemma is a promising approach because it will make explicit that some of the words are repeated in both sentences even if small variations of these words are used in each sentence (*e.g.* different verb tenses). After this step, each sentence contained in T and H from the pair $\langle T, H \rangle$ under analysis were represented in a structured format (set of tokens) and annotated with some additional information regarding the content of the text (*e.g.* part-of-speech tags).

In order to apply ML algorithms we need to represent the training instances by a set of numerical features. Since in this problem we receive a pair of sentences as input and we aim to automatically classify the relation between them as output, the feature set should be designed taking special attention to the properties that characterize such relation. To represent each pair $\langle T, H \rangle$ we employed a set of features (listed in Table 3) at the lexical, syntactic and semantic level. The first four lexical features listed in Table 3 aim to capture the overlap of information expressed in T in relation to H and vice-versa. Feature *T_Bigger_H* tries to capture the intuition that in a relation of *Entailment*, sentence H is usually smaller than sentence T. Regarding syntactic features, changes in verb tense are typically not expected to occur in *Paraphrase* relations, but rewriting the same sentence using alternation between passive and active voice is the most common case of paraphrase relations. Semantic features were employed for tokens in one of the sentences that do not occur in the other, after removing named entities (to avoid overlap with lexical features). The first three features capture semantic relations between each pair of tokens using knowledge extracted from a Portuguese wordnet. The last two features explore the word embeddings model and aim to capture different ways of measuring semantic relations between H and T, after projecting each sentence in the embedding space.

Table 3. Feature set

Type	Feature	Description
Lexical	Overlap_T	% of (unique) tokens in T that exist in H
	Overlap_H	% of (unique) tokens in H that exist in T
	NE_T	% of (unique) named entities in T that exist in H
	NE_H	% of (unique) named entities in H that exist in T
	T_Bigger_H	If $\|T\| > \|H\|$ returns 1. Returns 0, otherwise
Syntactic	Tense	If T and H are written in the same grammatical tense
	Voice	If T and H are written in the same grammatical voice
Semantic	Synonym	% of tokens in T synonyms of tokens in H. And vice-versa
	Hyperonym	% of tokens in T hyperonyms of tokens in H. And vice-versa
	Meronym	% of tokens in T meronyms of tokens in H. And vice-versa
	Cos_Sim	cosine similarity between $\vec{e}(T)$ and $\vec{e}(H)$
	Entail_Versor	entailment versor ($\hat{d}$) in the word embeddings space

Knowledge about the words of a language and their semantic relations with other words can be exploited with large-scale lexical databases. To enrich the feature set shown in Table 3 with semantic knowledge, we explored external semantic resources. By exploiting these resources we aim to enable the system to deal better with the diversity and ambiguity of natural language text. Similarly to WordNet [10] for the English language, CONTO.PT [14] is a fuzzy wordnet for Portuguese, which groups words into sets of cognitive synonyms (called *synsets*), each expressing a distinct concept. In addition, synsets are interlinked by means of conceptual and semantic relations (*e.g.* "hyperonym" and "part-of"). Synsets included in CONTO.PT were automatically extracted from several linguistic resources. All the relations represented in CONTO.PT (i.e. relations between words and synsets, as well as relations between synsets) include degrees of membership. Two tokens (obtained after tokenization and lemmatization) are considered synonyms if they occur in the same synset. One token T_i is considered hyperonym of T_j if there exists a hyperonym relation ("hyperonym_of") between the synset of T_i and the synset of T_j. Similarly, T_i is considered meronym of T_j if there exists a meronym relation ("part_of" or "member_of") between the synset of T_i and the synset of T_j.

Finally, we exploit a distributed representation of words (word embeddings) to compute the last two features described in Table 3. These distributions map a word from a dictionary to a feature vector in high-dimensional space, without human intervention, from observing the usage of words on large (non-annotated) corpora. This real-valued vector representation tries to arrange words with similar meanings close to each other based on the occurrences of these words in large-scale corpora. Then, from these representations, interesting features can be explored, such as semantic and syntactic similarities. In our experiments, we used a pre-trained model provided by the *Polyglot*[1] tool [2], in which a neural network architecture was trained with Portuguese *Wikipedia* articles.

In order to obtain a score indicating the similarity between two text fragments, T_i and T_j, we compute the cosine similarity between the vectors that represent each of the text fragments in the high-dimensional space. Each text fragment is projected into the embedding space as $\vec{T_i} = \sum_{k=1}^{n} \vec{e}(w_k) n^{-1}$, where $\vec{e}(w_k)$ represents the embedding vector of the word w_k and n corresponds to the number of words contained in the text fragment T_i. Then, we compute the final value of the cosine similarity $\delta_{\vec{T_i},\vec{T_j}} = \cos(\vec{T_i}, \vec{T_j})$, $\delta_{\vec{T_i},\vec{T_j}} \in [-1, 1]$ followed by the following rescaling and normalization: $(1.0 - \delta_{\vec{T_i},\vec{T_j}})/2.0$. The entailment versor $(\hat{d})$ corresponds to the normalized direction vector obtained by subtracting the projection of T in the embedding space, $\vec{e}(T)$, by the projection of H, $\vec{e}(H)$.

For each classification task, we have run several experiments exploring some well known state-of-the-art algorithms, namely: *Support Vector Machine* (SVM) using linear and polynomial kernels, *Maximum Entropy model* (MaxEnt), *Adaptive Boosting* algorithm (AdaBoost) using *Decision Trees* as weak classifiers, *Random Forrest Classifier* using *Decision Trees* as weak classifiers, and

[1] http://polyglot.readthedocs.io/en/latest/index.html.

Multilayer Perceptron Classifier (Neural Net) with one hidden layer. All the ML algorithms previously mentioned were employed using the *scikit-learn* library [25] for the *Python* programming language. Since the best overall results reported in all the evaluation scenarios were obtained using a *SVM* with a *linear* kernel, all the results reported in Sect. 5 were obtained using this classifier.

5 Experiments and Results

We investigate four evaluation scenarios. First, we report 10-fold cross validation results over all the training examples of the European-Portuguese partition of the ASSIN corpus, using a simple set of features, namely the lexical and syntactic-based features presented in Sect. 4. We also report on the results obtained by the learned model on a separate test set from the ASSIN corpus containing examples annotated in European-Portuguese. The system obtained in this scenario corresponds to our baseline. The second evaluation scenario follows a similar setting but using a more sophisticated set of features, in which semantic-based features were included (complete set of features described in Sect. 4). In this evaluation scenario we aim to determine the impact semantic-based features have in correctly identifying entailment relations. In the third evaluation scenario, we report 10-fold cross validation results over all the training examples available in the ASSIN corpus, including both the European-Portuguese and the Brazilian-Portuguese partitions, using the complete set of features described in Sect. 4. In this evaluation scenario we aim to validate our intuition that increasing the training set with more training data, regardless of the differences between European-Portuguese and Brazilian-Portuguese, should increase the performance of the system for the task of recognizing textual entailment and paraphrases from text written in Portuguese.

Table 4. Evaluation results for each evaluation scenario of the multi-class setting.

	Train						Test	
	N	E	P	Total			Total	
	F1	*F1*	*F1*	*F1*	*Macro-F1*	*Acc.*	*Macro-F1*	*Acc.*
EP	0.89	0.69	**0.60**	0.82	**0.73**	0.823	0.69	0.817
EP and Semantic	**0.9**	**0.7**	0.59	**0.83**	**0.73**	0.824	**0.71**	0.821
EP+BP and Semantic	**0.9**	0.65	0.52	0.82	0.69	0.819	**0.71**	**0.827**

Table 4 summarizes the results obtained in our experiments regarding the multi-class formulation. Each line corresponds to the results obtained in each of the evaluation scenarios previously described. The first three columns correspond to the averaged F1-score evaluation metric obtained after performing 10-fold cross validation on the training data for each label considered in the classification problem, namely *None* (N), *Entailment* (E) and *Paraphrase* (P). The last three columns, also regarding the results obtained in the training set, correspond to the

overall results obtained for each evaluation metric, namely *micro F1-score* (F1), *macro F1-score* (Macro-F1) and *accuracy* (Acc.). Finally, the last two columns correspond to the overall *macro F1-score* and *accuracy* obtained in the test set.

In general, we obtained better overall results in the recognition of the *None* relation (0.9), followed by *Entailment* relations (0.7) and by *Paraphrase* relations (0.6). We associate these results to the higher number of learning instances available in the corpus for each of the labels *None* and *Entailment*, respectively.

From the analysis of the results we conclude that enhancing the feature set with semantic-based features improved the overall results, but such improvements are not statistically significant. We expected these improvements to be more significant, since it seems intuitive that semantic-based features are relevant for the task of recognizing textual entailment and paraphrases. After performing feature and error analysis, we associate these results with the following: (a) the system gave too much importance to the "percentage of overlapping tokens" feature (*i.e.* when the value of the feature "Overlap_T" is very high the system tends to predict *Paraphrase*, when the feature "Overlap_H" is very high the system tends to predict *Entailment*, and when these values are both very low the system tends to predict *None*); (b) the coverage of semantic-based features is relatively low, causing this feature to have null values in some situations.

Comparing the results obtained by the system using the European-Portuguese and the Brazilian-Portuguese training set of the ASSIN corpus, we observed that increasing the training set with the Brazilian-Portuguese partition reduced the overall performance of the system. These results suggest that some characteristics of entailment and paraphrase relations between two text fragments of the Brazilian-Portuguese partition are different from the European-Portuguese partition. Furthermore, syntactic and semantic differences between the two variants are responsible for the majority of the errors made by the system. The best overall results in the test data were obtained in the last evaluation scenario, which we associate to the highest number of training examples that were provided to the system during the training phase. These resulted in a system that is able to generalize better for unseen data, explaining the results shown in Table 4. Comparing the results reported in this paper with the systems participating in the *ASSIN Challenge*, our approach would be ranked in a second place, obtaining an overall score that is very close to the results presented by the best system: 0.8385 of accuracy and 0.7 of macro F1-score ("L2F/INESC-ID" team).

Finally, in a fourth evaluation scenario, we address the problem in a different perspective, motivated by the characteristics of the ASSIN corpus. As shown in Table 1, the distribution of classes in the ASSIN corpus is very unbalanced, with a much lower number of examples for the *Paraphrase* class. As introduced in Sect. 1, a *Paraphrase* can be formulated as a bidirectional entailment. In this experimental setup we formulate the problem of recognizing textual entailment as a binary classification problem between the classes *Entailment/Paraphrase* and *None*. The training set was built as follows: (a) each *Paraphrase* example from the ASSIN corpus was transformed into two new *Entailment* examples (*i.e.* T entails H and H entails T); (b) the remaining *None* and *Entailment* examples from the ASSIN corpus were added. The test set comprises the same

examples of the ASSIN corpus, where the *Entailment* and *Paraphrase* classes were aggregated in the same class (E+P). We aim to demonstrate the ability of the approach proposed in this paper to distinguish situations where the text sentence (T) entails the hypothesis sentence (H) from when this is not the case. The results obtained in this experimental setup are shown in Table 5. The first two lines correspond to the results obtained for each of the target classes: *None* (N) and *Entailment/Paraphrase* (E+P). For each of the partitions (training and test set) of the ASSIN corpus containing annotations for European-Portuguese, the first column presents the total number of samples used in the experiments and the last two columns correspond to the accuracy and averaged micro F1-score evaluation metrics obtained after performing 10-fold cross validation. The results obtained in the binary formulation show that this binary classification task makes the decision boundaries easier to distinguish.

Table 5. Evaluation results for the binary classification setting

	Train			Test		
	# samples	*Acc.*	*F1*	*# samples*	*Acc.*	*F1*
N	2046	0.87	0.88	1386	0.86	0.88
E + P	1179	0.81	0.79	614	0.78	0.74
Total/avg	3225	0.85	0.85	2000	0.83	0.84

6 Conclusions

In this paper, we presented a preliminary approach to address the NLP task of recognizing textual entailment and paraphrases from text written in the Portuguese language. Firstly, we formulated this task as a multi-class classification problem. The overall results reported in this paper are promising (accuracy of 0.827 in the test set). A close assessment of obtained results shown that the number of annotated sentence pairs may not be sufficient to build a system that generalizes well for unseen data since the implemented classifiers tend to prefer labels that contain more training instances simply because they are more representative of the training data in statistical terms. Looking at the obtained results, we conclude that the overall system performance improved with semantic-based features, but not significantly. Notwithstanding, a detailed analysis points that this is one of the most promising directions for future work. Increasing the training set with the Brazilian-Portuguese partition of the ASSIN corpus had an unexpected impact in the overall performance of the system. We associate this result to syntactic and semantic differences between European and Brazilian Portuguese and because some of the external resources that were employed (*i.e.* fuzzy wordnet, part-of-speech tagger, word embeddings model) are based on the European-Portuguese language. Consequently, some lexical, syntactic and semantic Brazilian-Portuguese linguistic phenomena may be missing or misleading in this approach. Then, we formulate the problem as a binary classification task and demonstrate the ability of the system to recognize textual entailment.

In future work, we would like to enhance the semantic-based features employed in our system, including: metrics to evaluate semantic similarity between fragments of text using the fuzzy wordnet described in this paper, sentence-level representations (e.g. using a dependency parser) and, more sophisticated computations using distributed representation models. These are promising directions for future work that we intend to pursue.

Acknowledgments. The first author is partially supported by a doctoral grant from Doctoral Program in Informatics Engineering (ProDEI) from the Faculty of Engineering of the University of Porto (FEUP).

References

1. Agirre, E., Banea, C., Cardie, C., Cer, D.M., Diab, M.T., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., Wiebe, J.: Semeval-2015 task 2: semantic textual similarity, english, spanish and pilot on interpretability. In: Cer, D.M., Jurgens, D., Nakov, P., Zesch, T. (eds.) Proceedings of the 9th International Workshop on Semantic Evaluation, Denver, USA, pp. 252–263. ACL (2015)
2. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: distributed word representations for multilingual NLP. In: Proceedings of Seventeenth Conference on Computational Natural Language Learning, pp. 183–192. ACL, Sofia, Bulgaria, August 2013
3. Alves, A.O., Oliveira, H., Rodrigues, R.: ASAPP: Alinhamento Semântico Automático de Palavras aplicado ao Português. Linguamática **8**(2), 43–58 (2016)
4. Androutsopoulos, I., Malakasiotis, P.: A survey of paraphrasing and textual entailment methods. J. Artif. Int. Res. **38**(1), 135–187 (2010)
5. Beltagy, I., Roller, S., Cheng, P., Erk, K., Mooney, R.J.: Representing meaning with a combination of logical and distributional models. Comput. Linguist. **42**(4), 763–808 (2016)
6. Bentivogli, L., Dagan, I., Dang, H.T., Giampiccolo, D., Magnini, B.: Fifth PASCAL recognizing textual entailment challenge. In: Proceedings of Text Analysis Conference (2009)
7. 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). doi:10.1007/11736790_9
8. Dagan, I., Roth, D., Sammons, M., Zanzotto, F.M.: Recognizing Textual Entailment: Models and Applications. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael (2013)
9. De Marneffe, M., Rafferty, A.N., Manning, C.D.: Finding contradictions in text. In: Association for Computational Linguistics (2008)
10. Fellbaum, C. (ed.): WordNet: an electronic lexical database Language, speech, and communication. MIT Press, Cambridge (1998)
11. Fialho, P., Marques, R., Martins, B., Coheur, L., Quaresma, P.: INESC-ID@ASSIN: Medição de Similaridade Semântica e Reconhecimento de Inferência Textual. Linguamática **8**(2), 33–42 (2016)
12. Fonseca, E., Santos, L., Criscuolo, M., Aluisio, S.: ASSIN: avaliacao de similaridade semantica e inferencia textual. In: Computational Processing of the Portuguese Language - 12th International Conference, Tomar, Portugal, 13–15 July (2016)

13. Garcia, M., Gamallo, P.: Yet another suite of multilingual NLP tools. In: Sierra-Rodríguez, J.-L., Leal, J.P., Simões, A. (eds.) SLATE 2015. CCIS, vol. 563, pp. 65–75. Springer, Cham (2015). doi:10.1007/978-3-319-27653-3_7
14. 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). doi:10.1007/978-3-319-41552-9_29
15. Hartmann, N.S.: Solo Queue at ASSIN: Combinando Abordagens Tradicionais e Emergentes. Linguamática **8**(2), 59–64 (2016)
16. Lai, A., Hockenmaier, J.: Illinois-LH: a denotational and distributional approach to semantics. In: Proceedings of 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 329–334. ACL, Dublin, Ireland, August 2014
17. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of 42nd Annual Meeting Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
18. Lippi, M., Torroni, P.: Argumentation mining: state of the art and emerging trends. ACM Trans. Internet Technol. **16**(2), 10:1–10:25 (2016)
19. Madnani, N., Dorr, B.J.: Generating phrasal and sentential paraphrases: a survey of data-driven methods. Comput. Linguist. **36**(3), 341–387 (2010)
20. Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., Zamparelli, R.: Semeval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In: Nakov, P., Zesch, T. (eds.) Proceedings of 8th International Workshop on Semantic Evaluation, COLING, Dublin, Ireland, pp. 1–8. ACL (2014)
21. Moens, M.F.: Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer, Heidelberg (2009)
22. Mollá, D., Vicedo, J.L.: Question answering in restricted domains: an overview. Comput. Linguist. **33**(1), 41–61 (2007)
23. Padó, S., Galley, M., Jurafsky, D., Manning, C.: Robust machine translation evaluation with entailment features. In: Proceedings of Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 1, pp. 297–305. ACL, Stroudsburg, PA, USA (2009)
24. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of 40th Annual Meeting Association Computational Linguistics, pp. 311–318. ACL, Stroudsburg, PA, USA (2002)
25. 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)
26. Rocha, G., Lopes Cardoso, H., Teixeira, J.: ArgMine: a framework for argumentation mining. In: 12th International Conference on Computational Processing of the Portuguese Language - PROPOR 2016, Student Research Workshop, Tomar, Portugal, 13–15 July (2016)
27. Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kociský, T., Blunsom, P.: Reasoning about entailment with neural attention. CoRR abs/1509.06664 (2015)
28. Sammons, M., Vydiswaran, V., Roth, D.: Recognizing textual entailment. In: Bikel, D.M., Zitouni, I. (eds.) Multilingual Natural Language Applications: From Theory to Practice, pp. 209–258. Prentice Hall, Upper Saddle River (2012)