Juan Antonio Lossio-Ventura
Hugo Alatrista-Salas (Eds.)

Communications in Computer and Information Science 795

Information Management and Big Data

4th Annual International Symposium, SIMBig 2017
Lima, Peru, September 4–6, 2017
Revised Selected Papers

Springer

Communications in Computer and Information Science 795

More information about this series at http://www.springer.com/series/7899

Juan Antonio Lossio-Ventura
Hugo Alatrista-Salas (Eds.)

Information Management and Big Data

4th Annual International Symposium, SIMBig 2017
Lima, Peru, September 4–6, 2017
Revised Selected Papers

Springer

Editors

Juan Antonio Lossio-Ventura
University of Florida
Gainesville, FL
USA

Hugo Alatrista-Salas
Universidad del Pacífico
Lima
Peru

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-90595-2 ISBN 978-3-319-90596-9 (eBook)
https://doi.org/10.1007/978-3-319-90596-9

Library of Congress Control Number: 2018941551

Printed on acid-free paper

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

Preface

Today, data scientists use the term “big data” to describe the exponential growth and availability of data, which could be structured and unstructured. In this context, techniques used in data science must face a new challenge, which is to extract insights from a large amount of real-time and heterogeneous data (*e.g.*, video, audio, text, image).

Big data has taken place over the past 20 years. For instance, social networks such as Facebook, Twitter, and LinkedIn generate masses of data, which are available to be accessed by other applications. Several domains, including biomedicine, life sciences, and scientific research, have been affected by big data. Therefore there is a need to understand and exploit these data. This process is performed with data science, which is based on methodologies of data mining, natural language processing, Semantic Web, statistics, etc. This allows us to gain new insight through data-driven research. A major problem hampering big data analytics development is the need to process several types of data, such as structured, numeric, and unstructured data (*e.g.*, video, audio, text, image, etc.).

The Annual International Symposium on Information Management and Big Data seeks to present new methods in fields related to the data science for analyzing and managing large volumes of data. SIMBig aims to bring together main — national and international — actors in the field dealing with new technologies dedicated to handling a large amount of information. Moreover, the symposium is a convivial place where these actors present their scientific contributions in the form of full and short papers. This book offers extended versions of the best papers presented at SIMBig 2017[1]. This fourth edition of SIMBig was held in Lima, Peru, during September 4–6. The proceedings are indexed in DBLP[2] [1] and as CEUR Workshop Proceedings[3].

In this special edition, ten long papers were selected from 24 presented in the conference. SIMBig 2017 received 71 submissions.

SIMBig is positioning itself as one of the most important conferences in South America on issues related to information management and big data.

To share the new analysis methods for managing large volumes of data, we encouraged participation from researchers in all fields related to big data, data science, data mining, natural language processing, and the Semantic Web, but also multilingual text processing, and biomedical NLP.

Topics of interest of SIMBig included: data science, big data, data mining, natural language processing, bio-NLP, text mining, information retrieval, machine learning, the Semantic Web, ontologies, Web mining, knowledge representation and linked open data, social networks, social Web, and Web science, information visualization, OLAP,

[1] http://simbig.org/SIMBig2017/.

[2] http://dblp1.uni-trier.de/db/conf/simbig/simbig2017.html.

[3] http://ceur-ws.org/Vol-2029/.

data warehousing, business intelligence, spatiotemporal data, health care, agent-based systems, reasoning and logic, constraints, satisfiability, and search.

SIMBig 2017 was supported mainly by the Universidad del Pacífico, the Pontifical Catholic University of Peru, and the University of Florida.

March 2018

Juan Antonio Lossio-Ventura
Hugo Alatrista-Salas

Reference

1. Juan Antonio Lossio-Ventura and Hugo Alatrista-Salas (eds.), Proceedings of the 4th Annual International Symposium on Information Management and Big Data, SIMBig 2017, Lima, Peru, September 4–6, 2017. CEUR Workshop Proceedings 2029, CEUR-WS.org 2017.

Organization

SIMBig 2017: Organizing Committee

General Organizers

Juan Antonio Lossio-Ventura	University of Florida, USA
Hugo Alatrista-Salas	Universidad del Pacífico, Peru

Local Organizers

Michelle Rodriguez Serra	Universidad del Pacífico, Peru
Cristhian Ganvini Valcarcel	Universidad Andina del Cusco, Peru

SNMAM Track Organizers

Jorge Valverde-Rebaza	University of São Paulo, Brazil
Alneu de Andrade Lopes	University of São Paulo, Brazil

ANLP Track Organizers

Marco Antonio Sobrevilla-Cabezudo	University of São Paulo, Brazil
Félix Arturo Oncevay-Marcos	Pontificia Universidad Católica del Perú, Peru
Félix Armando Fermín-Pérez	UNMSM, Peru

SIMBig 2017: Program Committee

SIMBig Program Committee

Elie Abi-Lahoud	University College Cork, Ireland
César Antonio Aguilar	Pontificia Universidad Católica de Chile, Chile
Sophia Ananiadou	NaCTeM University of Manchester, UK
Jérôme Azé	LIRMM University of Montpellier, France
Riza Batista-Navarro	NaCTeM University of Manchester, UK
Nicolas Béchet	IRISA Université de Bretagne-Sud, France
Jiang Bian	University of Florida, USA
Albert Bifet	MINES ParisTech, France
Sandra Bringay	LIRMM Paul Valéry University, France
Bruno Cremilleux	Université de Caen Normandie, CNRS
Fabio Crestani	University of Lugano, Switzerland
Martín Ariel Domínguez	Universidad Nacional de Córdoba, Argentina
Brett Drury	National University of Ireland Galway, Ireland

Frédéric Flouvat	PPME University of New Caledonia, New Caledonia
Philippe Fournier-Viger	Harbin Institute of Technology, China
Natalia Grabar	University of Lille 3, France
Adrien Guille	Université Lumière Lyon 2, France
Thomas Guyet	IRISA/LACODAM Agrocampus Ouest, France
Phan Nhat Hai	New Jersey Institute of Technology, USA
Jiawei Han	University of Illinois, USA
Sébastien Harispe	Ecole des Mines d'Alès, France
William Hogan	University of Florida, USA
Vijay Ingalalli	Inria Bretagne Atlantique, France
Georgios Kontonatsios	Edge Hill University, UK
Yannis Korkontzelos	Edge Hill University, UK
Ravi Kumar	Google, USA
Christian Libaque-Saenz	Universidad del Pacífico, Peru
Cédric López	VISEO Research and Development Unit, France
André Miralles	SISO Team, France
François Modave	University of Florida, USA
Jordi Nin	BBVA Data & Analytics and Universidad de Barcelona, Spain
Miguel Nuñez-del-Prado-Cortéz	Universidad del Pacífico, Peru
Maciej Ogrodniczuk	Polish Academy of Sciences, Poland
Marco Aurélio Pacheco	Pontifícia Universidade Católica do Rio de Janeiro, Brazil
José Manuel Perea-Ortega	University of Extremadura, Spain
Pascal Poncelet	LIRMM University of Montpellier, France
Julien Rabatel	Catholic University of Leuven, Belgium
José-Luis Redondo-García	Polytechnic University of Madrid, Spain
Mathieu Roche	Cirad - TETIS - LIRMM, France
Nancy Rodriguez	LIRMM University of Montpellier, France
Rafael Rossi	University of São Paulo, Brazil
Fatiha Saïs	Université Paris-Sud 11, France
Arnaud Sallaberry	LIRMM Paul Valéry University, France
Matthew Shardlow	University of Manchester, UK
Gerardo Eugenio Sierra-Martínez	Instituto de Ingeniería, UNAM
Newton Spolaor	Universidade de São Paulo, Brazil
Claude Tadonki	MINES ParisTech, France
Maguelonne Teisseire	Irstea, TETIS, France
Paul Thompson	University of Manchester, UK
Carlos Vàzquez	École de technologie supérieure, Canada
Didier Vega	Universidade de São Paulo, Brazil
Julien Velcin	Université Lumière Lyon 2, France
Maria-Esther Vidal	Universidad Simón Bolívar, Venezuela
Boris Villazon-Terrazas	Fujitsu Laboratories of Europe, Spain

Youyou Wu	Kellogg School of Management, USA
Yang Yang	Kellogg School of Management, Northwestern University, USA
Guo Yi	University of Florida, USA
Amrapali Zaveri	Dumontier Lab, USA
He Zhe	Florida State University, USA

SNMAM Program Committee

Alan Valejo	University of São Paulo, Brazil
Brett Drury	National University of Ireland Galway, Ireland
Celso Kaestner	Federal University of Technology of Paraná, Brazil
Didier Vega-Oliveros	University of São Paulo, Brazil
Hugo Gualdron Colmenares	University of São Paulo, Brazil
José Benito Camiña	Tecnológico de Monterrey, Mexico
Jesús Mena-Chalco	Federal University of ABC, Brazil
Lilian Berton	University of Santa Catarina State, Brazil
Luca Rossi	Aston University, UK
Marcos Domingues	State University of Maringá, Brazil
Marcos G. Quiles	Federal University of São Paulo, Brazil
Mathieu Roche	CIRAD and University of Montpellier, France
Merley Conrado	Intel Corp., USA
Newton Spolaôr	Western Paraná State University, Brazil
Pascal Poncelet	University of Montpellier, France
Rafael Rossi	Federal University of Mato Grosso do Sul, Brazil
Ricardo Campos	Polytechnic Institute of Tomar and LIAAD/INESC TEC, Portugal
Ricardo Marcacini	Federal University of Mato Grosso do Sul, Brazil
Ronaldo C. Prati	Federal University of ABC, Brazil
Sabrine Mallek	Institut Supérieur de Gestion de Tunis, Tunisia
Thiago de Paulo Faleiros	University of Brasilia, Brazil
Vânia Neves	Federal University of Juiz de Fora, Brazil
Victor Stroele	Federal University of Juiz de Fora, Brazil

ANLP Program Committee

Thiago Alexandre Salgueiro Pardo	University of São Paulo, Brazil
Nathan Siegle Hartmman	University of São Paulo, Brazil
Leandro Borges dos Santos	University of São Paulo, Brazil
Fernando Emilio Alva Manchego	University of Sheffield, UK
Paula Christina Figueira Cardoso	Federal University of Lavras, Brazil
Márcio de Souza Dias	Federal University of Goiás, Brazil
Fernando Antônio Asevedo Nóbrega	University of São Paulo, Brazil

Roque Enrique López Condori	Institute for Research in Computer Science and Automation, France
Francis M. Tyers	UiT Norgga árktalaš universitehta, Norway
Shay Cohen	University of Edinburgh, UK
Shashi Narayan	University of Edinburgh, UK

SIMBig 2017: Organizing Institutions and Sponsors

Organizing Institutions

Universidad del Pacífico, Perú[1]
University of Florida, USA[2]
Universidad Andina del Cusco, Perú[3]

Collaborating Institutions

Springer[4]
Banco de Crédito del Perú[5]
Escuela de Post-grado de la Pontificia Universidad Católica del Perú[6]

SNMAM Organizing Institutions

Instituto de Ciências Matemáticas e de Computação, USP, Brazil[7]
Labóratorio de Intêligencia Computacional, ICMC, USP, Brazil[8]
Universidade Federal de São Carlos, Brazil[9]

ANLP Organizing Institutions

Universidad Nacional Mayor de San Marcos, Perú[10]
Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada, PUCP, Perú[11]
Instituto de Ciências Matemáticas e de Computação, USP, Brazil
Universidade Federal de São Carlos, Brazil

[1] http://www.up.edu.pe/.

[2] http://www.ufl.edu/.

[3] http://www.uandina.edu.pe/.

[4] http://www.springer.com/la/.

[5] https://www.viabcp.com/wps/portal/.

[6] http://posgrado.pucp.edu.pe/la-escuela/presentacion/.

[7] http://www.icmc.usp.br/Portal/.

[8] http://labic.icmc.usp.br/.

[9] http://www2.ufscar.br/home/index.php.

[10] http://www.unmsm.edu.pe/.

[11] http://inform.pucp.edu.pe/～grpiaa/.

UNIVERSIDAD DEL PACÍFICO
UNIVERSIDAD DEL PACIFICO LIMA - PERU
UNIVERSIDAD ANDINA DEL CUSCO
UF UNIVERSITY of FLORIDA
UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
ET LUX IN TENEBRIS LUCET MCMXVII
PUCP
ICMC USP SÃO CARLOS
LABIC Laboratório de Inteligência Computacional ICMC. USP
UNIVERSITÉ DE MONTPELLIER
Banco de Crédito BCP
Springer

Contents

Parallelization of Conjunctive Query Answering over Ontologies

E. Patrick Shironoshita[1](✉), Da Zhang[2], Mansur R. Kabuka[1,2], and Jia Xu[2]

[1] INFOTECH Soft, Inc., 1201 Brickell Avenue, Suite 220, Miami, FL 33131, USA
patrick@infotechsoft.com
[2] University of Miami, Coral Gables, FL 33124, USA

Abstract. Efficient query answering over Description Logic (DL) ontologies with very large datasets is becoming increasingly vital. Recent years have seen the development of various approaches to ABox partitioning to enable parallel processing. Instance checking using the enhanced most specific concept (MSC) method is a particularly promising approach. The applicability of these distributed reasoning methods to typical ontologies has been shown mainly through anecdotal observation. In this paper, we present a parallelizable, enhanced MSC method for the answering of ABox conjunctive queries, using a set of syntactic conditions that permit querying of large practical ontologies in reasonable time, and combining it with pattern matching to answer queries over role assertions. We also present execution time and efficiency of an implementation deployed over computing clusters of various sizes, showing the ability of the method to process instance checking for large scale datasets.

1 Introduction

Description Logics (DL) are now widely being used to model and represent structured and semi-structured data in different applications [1], particularly through the use of the Web Ontology Language (OWL). A core task for DL systems is to provide an efficient way to answer queries over the extensional level of the ontology, that is, to compute answers that are not only asserted, but logically implied by the ontology [2]. Considerable efforts have been dedicated to the optimization of algorithms for query answering [2–4]. One of the challenges faced in this era of increasing data wealth is to produce responsive results for queries over very large data sets [5]. However, even as reasoners for very expressive DLs have been created, performing reasoning over very large ABoxes is still prohibitive [2,6,7].

In the last few years, methods for parallel and distributed reasoning over expressive ontologies have been published [5,8–10]; these methods generally seek to make use of fast syntactic checks to generate a set of independent partitions that can be processed in parallel. In [11], we have proposed a method for instance checking, combining the work in [8,9] with the idea of a most specific concept (MSC) [12–14] to move the reasoning task from the very large ABox into a much smaller TBox. Empirical evaluation of the method has shown its ability

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 1–14, 2018.
https://doi.org/10.1007/978-3-319-90596-9_1

to perform sound and complete instance checking over $\mathcal{SHI}$ DLs within reasonable time. Moreover, the method is inherently parallelizable, lending itself to implementation within clusters of commodity hardware.

In this paper, we examine this enhanced MSC method first described in [11], extend it to use in conjunctive queries, and evaluate its parallelization. First, we provide a detailed definition of the MSC method and of the syntactic conditions that enable its use for instance checking. Following this, we describe the use of the MSC method in conjunction with pattern matching to answer ABox conjunctive queries. Subsequently, results of tests performed for accuracy of the enhanced MSC method and of its extension to conjunctive queries are presented, as well as results from experiments over a parallelized implementation. Finally, we discuss our future work and provide our conclusions.

2 The Enhanced Most Specific Concept (MSC) Method

2.1 Basic MSC Method

A Description Logics (DL) knowledge base, also referred to as an ontology, is typically defined a tuple, denoted $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, where the *terminological* component or TBox $\mathcal{T}$ contains definitions of concepts and roles, and the *assertional* component or ABox $\mathcal{A}$ contains assertions about membership of individuals in concepts and about role relations between individuals. The set of roles, concepts, and individual instances in an ontology are denoted respectively by **R**, **C**, and **I**. The discussion in this paper assumes that the reader is familiar with DL concepts and notations; the reader is referred to [15] for details. For the discussion below, we will use the example illustrated in Table 1.

Definition 1 (Most Specific Concept). [12,13] *Let $\mathcal{K} = \{\mathcal{T}, \mathcal{A}\}$ be an ontology, and a be an individual in* **I**. *The most specific concept for a w.r.t. $\mathcal{A}$, written $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a)$, is a concept such that for every concept D where $\mathcal{K} \models D(a)$, $\mathcal{T} \models \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}.a) \sqsubseteq D$.*

If $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a)$ can be derived, then, to test whether $\mathcal{K} \models Q(a)$ holds for an arbitrary concept Q it suffices to test if $(\mathcal{T} \cup \{Q\}) \models \mathrm{MSC}_{\mathcal{T} \cup \{Q\}}(\mathcal{A}, a) \sqsubseteq Q$. We call the concept Q the *query*. Note that Q needs to be inserted into the TBox in simple form, which may require in turn the insertion of additional named concepts as well as general concept inclusion (GCI) axioms to express the necessary equivalences for these inserted concepts. In the remainder of this paper, we will assume that the query Q has been inserted into the TBox so that $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a)$ denotes the MSC calculated including Q.

Computation of the MSC for a given individual a as defined above can be performed using a *rolling up* procedure adapted from the one first introduced in [16].

Definition 2 (Basic MSC Rollup Procedure). *Provided that the ABox does not contain assertion cycles, computation of the MSC can be performed recursively as follows:*

Table 1. Example ABox

TBox
Headmaster ⊑ Professor
Professor ⊑ Person
MagicCourse ⊑ Course
Muggle ⊑ ¬Wizard
takesCourse.MagicCourse ⊑ Wizard
∃isHeadOf.School ⊓ Person ⊑ Headmaster
ABox
School(hogwarts)
Professor(albus)
Professor(severus)
MagicCourse(potions)
MagicCourse(transfiguration)
Course(math)
Student(harry)
Muggle(dudley)
isHeadOf(hogwarts, albus)
takesCourse(harry, transfiguration)
taughtBy(transfiguration, albus)
taughtBy(potions, severus)

1. *for a given individual a, start with an empty* $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a)$*;*
2. *for every concept assertion $C(a)$,*
   $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a) \leftarrow \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a) \sqcap C$*;*
3. *for every role assertion $R(a, b)$,*
   $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a) \leftarrow \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a) \sqcap \exists R.\mathrm{MSC}_{\mathcal{T}}(\mathcal{A} \backslash \{R(a, b)\}, b)$*;*
4. *for every individual equality assertion $a = a'$,*
   $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a) \leftarrow \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a) \sqcap \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a')$*.*

So, in the example ABox above, the MSC for severus is

$$\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, \mathtt{severus}) = \mathtt{Professor} \sqcap \exists \mathtt{taughtBy}^{-}.\mathtt{MagicCourse} \tag{1}$$

It is important to note that while class assertions generate relatively simple concepts, role assertions are capable of generating very complex concepts. Suppose for example that ABox $\mathcal{A}_n$ consists of role assertions $R_0(a_0, a_1)$, $R_1(a_1, a_2)$, $\ldots$, $R_n(a_n, a_{n+1})$; then, the MSC of a_0 is:

$$\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}_n, a_0) = \exists R_1.(\exists R_2.(\ldots(\exists R_n.\mathrm{MSC}_{\mathcal{T}}(a_{n+1}))\ldots)) \tag{2}$$

Thus, in the example ABox of Table 1, the MSC for harry is

$$\begin{aligned}\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, \texttt{harry}) = \texttt{Student} \sqcap \exists \texttt{takesCourse}^{-}.&\\ (\texttt{MagicCourse} \sqcap \exists \texttt{taughtBy}.&\\ (\texttt{Professor} \sqcap \texttt{isHeadOf}^{-}.\texttt{School}))&\end{aligned} \tag{3}$$

Two main issues preclude this basic MSC method from proving fully useful with expressive ontologies. First, if assertion cycles are present in the ABox, the method does not terminate. For example, consider what would happen if the following is inserted in the TBox:

$$\texttt{isProtegeOf}(\texttt{albus}, \texttt{harry}) \tag{4}$$

In this case, a cycle forms among harry, transfiguration, and albus.

Second, unless the ABox is highly disconnected, the method has the potential to generate very large concepts, of size in the same order of the ABox itself. Consider what happens if the following assertion is added to the ABox:

$$\texttt{takesCourse}(\texttt{harry}, \texttt{potions}) \tag{5}$$

In this case, all individuals become connected with each other, and the MSC of every individual ends up being the same size of the ABox.

Xu et al. [11] define a set of improvements that result in the *Enhanced MSC Method*. This enhancement consists of two parts: a mechanism to address assertion cycles, and a set of syntactic conditions to reduce the size of the MSCs in practical ontologies.

2.2 MSC Computation with Assertion Cycles

To address assertion cycles, Xu et al. [11] use nominals to indicate the joint node of a cycle, as previously suggested in [13,17].

Definition 3 (MSC Rollup of Assertion Cycles). *When a cycle is found starting and ending at individual a_c, the individual is represented by its corresponding nominal class $\{a_c\}$, and the conversion of role assertions within the cycle requires modification of the rollup method as follows: If a_c is an individual where a cycle is found, select a direction to go through the cycle and:*

- *for $R(a_c, x)$, $x \neq a_c$*
  $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a_c) \leftarrow \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a_c) \sqcap (\{a_c\} \sqcap \exists R.\mathrm{MSC}_{\mathcal{T}}(\mathcal{A} \backslash \{R(a_c, x)\}, x))$
- *for $R(y, a_c)$, $y \neq a_c$*
  $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, y) \leftarrow \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, y) \sqcap \exists R.\{a_c\}$
- *for $R(a_c, a_c)$,*
  $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a_c) \leftarrow \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a_c) \sqcap \{a_c\}$
  $\sqcap \exists R.\{a_c\}$
- *for any other $R(x, y)$ in the cycle, for $x \neq a_c$ and $y \neq a_c$,*
  $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, x) \leftarrow \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, x) \sqcap \exists R.\mathrm{MSC}_{\mathcal{T}}(\mathcal{A} \backslash \{R(x, y)\}, y)$

Thus, an ABox $\mathcal{A}_c = \{R_0(a_0, a_1), R_1(a_1, a_2), \ldots, R_n(a_n, a_0)\}$ has an assertion cycle, then the MSC is obtained as follows:

$$\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}_c, a_0) = \{a_0\} \sqcap \exists R_1.(\exists R_2.(\ldots(\exists R_n.\{a_0\})\ldots)) \tag{6}$$

So, suppose that the ABox in Table 1 is augmented with the assertion in Eq. (4), then the MSC for `harry` becomes

$$\begin{aligned}\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, \texttt{harry}) =& \{\texttt{harry}\} \sqcap \texttt{Student} \sqcap \exists\texttt{takesCourse}^{-}.\\ &(\texttt{MagicCourse} \sqcap \exists\texttt{taughtBy}.\\ &(\texttt{Professor} \sqcap \texttt{isHeadOf}^{-}.\texttt{School}\\ &\sqcap \texttt{isProtegeOf}^{-}.\{\texttt{harry}\}))\end{aligned} \tag{7}$$

Use of the MSC method to perform instance retrieval is straightforward. Suppose it is desired to query an ABox to retrieve all instances of a concept Q. It suffices to generate $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a)$ for every individual a in the ABox, and then accept every individual where $(\mathcal{T} \cup Q) \models \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a) \sqsubseteq Q$. Note that the query concept Q is added to the TBox being verified.

2.3 Syntactic Conditions

In [11], syntactic conditions verifiable in polynomial time or better were defined, to enable reduction on the size of the MSC and thus permit TBox reasoning in tractable time. These conditions are based on the following:

Lemma 1. *Given two individuals a and b and a role R, a role assertion $R(a, b)$ influences the classification of individual a into concept A if and only if*

$$\mathcal{K} \models \exists R.B \sqcap A_0 \sqsubseteq A \tag{8}$$

where $\mathcal{K} \models B(b)$ and where $A_0 \not\sqsubseteq A$ summarizes information about a not contained in A.

Proposition 1. *Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be a $\mathcal{SHI}$ ontology containing named concept A, concepts A_0 and B, and role R. If Eq. (8) holds, there must exist some GCIs in $\mathcal{T}$ of the form*

$$\exists R'.C_1 \bowtie C_2 \sqsubseteq C_3 \tag{9}$$

where $R \sqsubseteq R'$, $\bowtie$ is a placeholder for either $\sqcap$ or $\sqcup$, and C_i's are concepts.

Proof of this proposition can be found in [8].

Proposition 1 directly leads to a first syntactic condition denoted SYN_COND.

Definition 4 (SYN_COND). *Role assertions of the form $R(a, b)$ are said to be* ***true*** *for* SYN_COND *if role R participates in at least one axiom that can be logically converted to the form of Eq. 9 for some $R \sqsubseteq R'$,* ***false*** *otherwise.*

Assertions $R(a,b)$ with SYN_COND = **false** do not affect the classification of a unless either R or some R' such that $R \subseteq R'$ exist in the query, and can be safely removed from the calculation of $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A},a)$. By symmetry to the inverse role R^{-}, this condition also applies to b. In practice, the axioms more likely to be found are of the form $C \sqsubseteq \exists R.D$, which is equivalent to $C \sqcap \top \sqsubseteq \exists R.D$. Also note that $(C \sqsubseteq \forall R.D) \equiv (\exists R.\neg C \sqsubseteq \neg D)$. In the example in Table 1, assertions with role `isHeadOf` have SYN_COND = **true**, while assertions with roles `takesCourse` and `taughtBy` have SYN_COND = **false** and can be safely removed from consideration, unless the roles exist in the query itself. Note that in this case, the MSC for `harry` is reduced to

$$\mathrm{MSC}_{\mathcal{T}}(\mathcal{A},\texttt{harry}) = \texttt{Student} \sqcap \exists\texttt{takesCourse}^{-}.\texttt{MagicCourse} \tag{10}$$

A second syntactic condition presented in [11] relies on the identification of explicit concept assertions and disjointness axioms that indicate that an individual cannot be classified to an existential restriction:

Definition 5 (SYN_COND_DJ). *Role assertions $R(a,b)$ with* SYN_COND = ***true*** *are said to be* ***true*** *for* SYN_COND_DJ *if there do not exist any of:*

- *an explicit concept assertion $B_0(b)$ such that $\mathcal{K} \models B_0 \sqsubseteq \neg C_1$;*
- *an explicit concept assertion $A_0(a)$ such that $\mathcal{K} \models A_0 \sqsubseteq \neg C_3$; or*
- *an explicit concept assertion $A_0(a)$ such that $\mathcal{K} \models A_0 \sqsubseteq \neg(C_3 \sqcup \neg C_2)$*

for C_1, C_2, an C_3 as in Eq. (9) and $R \sqsubseteq R'$.

Role assertions with SYN_COND_DJ = **false** can be safely removed from the calculation of the MSC of any individual, since they do not affect classification of either a or b. For example, in Table 1, role assertions with role `takesCourse` have SYN_COND = **true** due to the assertion `takesCourse.MagicCourse` $\sqsubseteq$ `Wizard`. However, suppose that the following assertion were inserted into the ABox:

$$\texttt{takesCourse}(\texttt{dudley},\texttt{math}) \tag{11}$$

This assertion has SYN_COND_DJ = **false**, since `dudley` is an instance of `Muggle`, which in turn is disjoint with `Wizard`, the filler concept in the assertion above.

Definition 6 (SYN_COND_SC). *Role assertions $R(a,b)$ with* SYN_COND = ***true*** *are said to be* ***true*** *for* SYN_COND_SC *if, for every GCI of the form in Eq. (9) there exists an explicit concept assertion $A_0(a)$ such that $\mathcal{K} \models A_0 \sqsubseteq C_3$.*

Role assertions with SYN_COND_SC = **true** are redundant for the classification of a, and can therefore be safely removed from the calculation of the MSC of any individual. For example, the role assertion `Headmaster(albus)` would have SYN_COND_SC = **true**.

Application of these three syntactic conditions to the computation of the MSC of a given individual results in a significant reduction in the size of the MSC for practical ontologies. The reasons for this reduction will be established in more detail in the next section, but first we provide a definition for the parallelization of the algorithm.

2.4 Parallelization of the MSC Method

The MSC method, including the syntactic conditions detailed above, is highly parallelizable, due to the following:

Proposition 2. *For a given knowledge base $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, computation of* $\mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a)$ *for an individual a can be performed independently of the computation of the MSC of any other individual.*

Proof of this proposition follows directly from the definition of the MSC computation in Definitions 2 and 3, as well as in the definition of the syntactic conditions, where it should be clear that MSC computation depends only on the initial state of the knowledge base. This means that the calculation of the MSCs for all individuals can be done in parallel. Instance checking then needs to be performed against $\mathcal{T} \cup \mathrm{MSC}_{\mathcal{T}}(\mathcal{A}, a)$ for every individual a.

As is shown later in Sect. 4, parallelization of the MSC method allows for the processing of extremely large ABoxes within clusters of commodity hardware. The resulting computational complexity is sublinear with respect to the size of the ABox, which indicates linear complexity for single-machine implementation.

3 Conjunctive Query Answering and SPARQL

The enhanced MSC method described in the previous section provides an effective, parallelizable algorithm for instance checking. In this section, we discuss its use for ABox conjunctive query answering.

Definition 7 ABox Conjunctive Query [7]**.** *Let* $\mathbf{C}$*,* $\mathbf{R}$ *and* $\mathbf{I}$ *denote the set of concepts, roles, and individuals in the ABox as before, and let* $\mathbf{V}$ *denote a set of variables disjoint with* $\mathbf{C}$*,* $\mathbf{R}$ *and* $\mathbf{I}$*. An* atom *is an expression $C(t)$, $R(t, t')$, or $t = t'$, where t and t' are members of* $\mathbf{V} \cup \mathbf{I}$*. An* ABox conjunctive query *is a collection of atoms.*

The result set of an ABox conjunctive query is a set of tuples, where each tuple contains a set of individuals that map to each variable in a conjunctive query. In the following discussions, we will use the notation $?x$ where necessary to refer specifically to variables, where x is any lowercase letter, borrowing from SPARQL notation.

ABox conjunctive queries are expressible in the SPARQL query language for RDF. SPARQL is designed as an extensible language where different entailment regimes can be employed, thus enabling the use of logic reasoning to derive results. Several works have been published that employ description logic entailment, particularly when using DL entailment regimes over OWL knowledge bases [18–20]. There are also efforts to enable SPARQL querying of large RDF graphs using distributed computing techniques [21,22]; however, these efforts do not include the use of DL entailment and are thus limited to the lesser expressive RDF semantics. It must be noted that SPARQL can be used for other queries - for example, they can be used to query for all class memberships for a single individual [20]. Arguably, however, ABox conjunctive queries are the most widely used, as they are designed to retrieve instances and their associated data values, in a manner analogous to SQL queries in relational databases.

Resolution of ABox conjunctive query atoms of the form $C(t)$ using the MSC method is straightforward, as this is equivalent to instance retrieval for a given class. On the other hand, query atoms of the form $R(t, t')$ and $t = t'$ require additional processing to ensure that all possible individual equalities are taken into account. For example, suppose that in the example in Table 1, the following assertion were added:

$$\texttt{albus} = \texttt{dumbledore} \tag{12}$$

Then, a query for $\texttt{?a} = \texttt{albus}$ should return both `albus` and `dumbledore`. Moreover, a query for `isHeadOf(?b,?a)` would need to return the tuples ($\texttt{hogwarts}$, $\texttt{albus}$) and ($\texttt{hogwarts}, \texttt{dumbledore}$).

For $R(t, t')$, one possible solution is to simply traverse the ABox and perform *pattern matching*:

Definition 8 Pattern Matching. *Triples $R'(a, b)$ are matched to a query $R(t, t')$ if all the following conditions are fulfilled:*

(a) $R' \sqsubseteq R$;
(b) if t (t') is an individual, then $t = a$ ($t' = b$); and
(c) if t (t') is a variable, then ***true****;*

If all conditions are fulfilled, then the triple $R(a, b)$ produces a solution to the query atom by mapping t to a if t is a variable, and similarly t' to b. For example, the query atom $\texttt{taughtBy}(\texttt{?a}, \texttt{?b})$ results in two solution sets, ($\texttt{transfiguration}, \texttt{albus}$) and ($\texttt{potions}, \texttt{severus}$).

Note however that this means that it is necessary to determine if the knowledge base entails individual equality; if the assertion in Eq. (12) were included in the ABox, then ($\texttt{transfiguration}, \texttt{dumbledore}$) would also be a solution to the query. Of course, resolution of equality is also trivially necessary to resolve atoms of the form $t = t'$.

It is possible to avoid this issue if the *unique name assumption* (UNA) is made, similar to database querying. Alternatively, it has been shown that for ontologies in the *DL-Lite* family, reasoning without the UNA can be reduced to reasoning with the UNA in polynomial time, through the resolution of functional properties and of the symmetric-reflexive-transitive closure of the individual equality assertions [23].

Proposition 3. *In a $\mathcal{SHI}$ DL, reasoning without the UNA can be reduced to reasoning with the UNA in polynomial time.*

Proof. The differences between the various *DL-Lite* families, and particularly the $DL\text{-}Lite^{\mathcal{HF}}_{core}$ version, and a $\mathcal{SHI}$ DL, reside in restrictions regarding class constructs. Given that neither *DL-Lite* nor $\mathcal{SHI}$ allow nominals in the TBox, it is not possible for any class assertion to result in an inference of individual equality. Therefore, the results for *DL-Lite* regarding reduction to UNA in [23] are also applicable to $\mathcal{SHI}$ DLs.

Clearly then, ABox conjunctive queries, including those expressed in the SPARQL query language, can be resolved using the enhanced MSC method for class query atom resolution, and pattern matching for role query atom resolution. Since pattern matching is also inherently parallelizable, as it involves verification of every triple in the ABox independent of the others, the entire query answering algorithm can be implemented as an iterative parallel solution.

Corollary 1. *In a $\mathcal{SHIQ}$ DL, reasoning without the UNA can be reduced to reasoning with the UNA in polynomial time.*

The proof follows from Proposition 3.

The MSC method can be extended to DL with expressivity $\mathcal{SHIQ}$ if UNA is assumed, which can be done using the result above. Details of such extension are outside the scope of this paper, but are straightforward, and generally follow the extensions to equality-free module extraction detailed in [8].

4 Experimental Evaluation

4.1 MSC Method Accuracy

To test both the accuracy and to measure parallelization speedup of the enhanced MSC method, we set up clusters of compute-optimized instances through Amazon Web Services (AWS)[1]. The enhanced MSC method with syntactic condition correction was implemented in Java and Scala to work over Apache Spark[2], installed over Hadoop HDFS and YARN.

For the accuracy tests, we used an in-memory version of our enhanced MSC implementation. We set up clusters of two c4.xlarge machines, each containing 4 virtual CPUs and 7.5 GB of memory, to run the enhanced MSC method, and compared the results against the HermiT reasoner version 3.8.1[3], running on a single c4.xlarge machine; HermiT was chosen as a comparison standard due to

Table 2. Times (in seconds) for execution of class and existential restriction queries

	Classes		Existential restrictions	
	MSC	HermiT	MSC	HermiT
Min.	7.82	0.67	7.30	0.76
Max.	19.66	780.20	26.94	2,848.68
Avg.	8.38	19.42	9.14	489.95
Std. Dev.	1.42	90.50	3.24	698.14
Median	8.16	0.79	8.19	135.94

[1] aws.amazon.com.

[2] https://spark.apache.org/.

[3] http://www.hermit-reasoner.com.

its stability and speed; future tests will be done against other reasoners such as Pellet[4] and Konclude[5]. These tests were performed against a single department dataset for the University Ontology Benchmark (UOBM) [24], containing about 150,000 triples. The UOBM TBox was modified to convert cardinality restrictions to existential restrictions, since our current implementation only handles expressivity up to $\mathcal{SHI}$. This modified version can be provided upon request. The UOBM Tbox contains a total of 113 named classes and 35 object properties. We tested our enhanced MSC implementation against all 35 named class queries, and all 3,955 possible single-depth existential restriction queries, and verified 100% agreement between our enhanced MSC method implementation and HermiT. In addition, we recorded the running time for all query executions - the results can be seen in Table 2. It is interesting to note that the enhanced MSC method performs much better than HermiT for existential restrictions. Also note the large standard deviation found when running a hypertableaux-based reasoner like HermiT, where a few queries take a very long time to finish. This result also suggests the possibility of using enhanced MSC in tandem with a traditional reasoner when dealing with smaller datasets.

4.2 Parallelization

To test the parallelization of the MSC method, we used the c3.8xlarge instances, which provide 32 virtual CPUs and 60 GBs of storage. We used the Lehigh University Benchmark (LUBM) [25] to generate data sets of up to 500 million triples. LUBM was chosen as an initial test ontology due to its ability to generate datasets of varying size, while providing a reasonably expressive TBox. These data sets were stored using the TitanDB[6] graph database interface over an Apache HBase[7] backend. Both TitanDB and HBase were installed over Hadoop HDFS 2.7. Our prototype application over Spark accesses TitanDB through its standard Java interface. Data distribution and replication are performed by the database and are transparent to our application.

A test was performed to evaluate the scalability of the parallel MSC method over the number of triples in the ABox. This test was performed over a cluster of 10 c3.8xlarge machines in AWS. The results are shown in the log-log diagram in Fig. 1. As can be observed, the method shows clear sub-linear performance with respect to the size of the data set, as expected from an algorithm with linear performance in a sequential machine.

The performance with respect to the number of machines was evaluated in two parts. First, to evaluate under small cluster conditions, a 500,000 triple set was assembled and used to test against 1 to 10 machines. The execution time and the efficiency with respect to single-machine execution are shown in Fig. 2(a). To obtain evaluation for large numbers of machines, we used the ABox with 500

[4] https://github.com/stardog-union/pellet.

[5] http://derivo.de/produkte/konclude/.

[6] http://thinkaurelius.github.io/titan/.

[7] http://hbase.apache.org/.

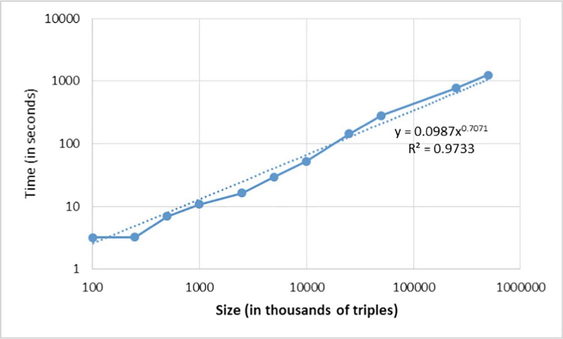
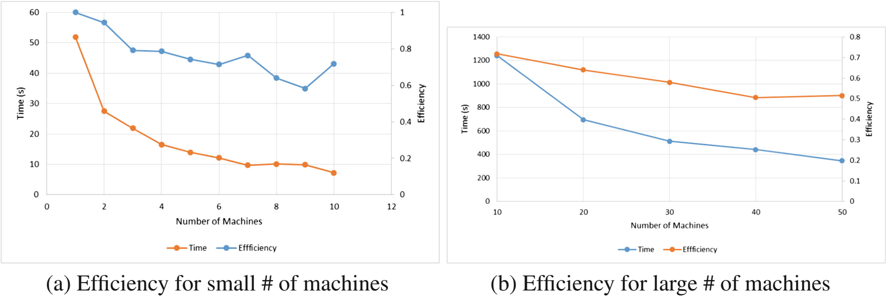


Fig. 1. Execution time vs. data set size with LUBM datasets, in AWS, for 10 c3.8xlarge machines with 32 cores each.

million triples and ran it against 10 to 50 machines; to provide a more realistic estimation, the efficiency value was corrected against the result for 500,000 triples in 10 machines. These higher scalability results are shown in Fig. 2(b).

In terms of raw performance, the parallel enhanced MSC algorithm was capable of performing instance checking for a dataset with 500 million triples and over 110 million individual instances in about 1,240 s, or around 20 min, using a cluster of 10 machines and a total of 320 execution cores. Using 50 machines, the execution time was 346 s, or somewhat less than 6 min. As a comparison, although the difference in algorithms means that execution times are not directly comparable, Oracle reports full ABox inference over 869 million triples in 62 min, and query performance over this pre-reasoned ABox in about 4.3 min, using specialized hardware [26]. It is also important to note that, since tests were performed over an uncontrolled environment, external perturbations could have affected some measurements. Nevertheless, it is clear that the MSC method provides performance comparable with top-of-the-line database technologies and very broad scalability.



(a) Efficiency for small # of machines

(b) Efficiency for large # of machines

Fig. 2. Execution time and efficiency of parallelization.

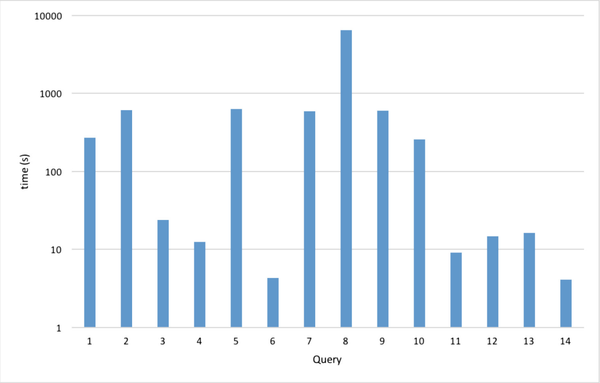


Fig. 3. Execution time for LUBM SPARQL queries, single node.

These results demonstrate that the enhanced MSC method is inherently parallelizable, enabling it to work with large scale ABoxes, and shows that it is useful with practical ontologies.

4.3 SPARQL Query Accuracy

An initial, non-parallelized prototype of a SPARQL query engine using the enhanced MSC method combined with pattern matching has been created. This initial prototype has been tested for accuracy against a LUBM(1, 0) dataset using a set of 14 test queries provided by the LUBM creators [25]. Testing shows 100% accuracy for all queries.

Speed of execution was measured by running the prototype on a VMWare virtual machine configured over a Dell PowerEdge R710 machine using an Intel Xeon E5620 CPU @ 2.40 GHz, running VSphere 4 Hypervisor. The virtual machine was set up for 4 cores and 64 GB of memory. The results are shown in Fig. 3. We expect to achieve significantly faster times using the enhanced MSC method, and are additionally working on the use of optimizations as described in [18].

5 Discussion and Future Work

The enhanced MSC method is inherently parallelizable, since instance checking for every individual in the ABox can be performed independently. Coupled with recent advances in cluster computing such as Apache Spark, large triple stores can be queried efficiently using commodity hardware clusters or cloud platforms. In combination with pattern matching, the method can also be used for answering conjunctive queries over the ABox, including those that can be formulated using the SPARQL query language.

We are currently working on the implementation of the ABox conjunctive query answering prototype to work over Spark, in order to test speedup and efficiency in its parallelization. We are also exploring the use of some optimization mechanisms that can be used to reduce the execution time. A parallelized implementation will also enable us to perform tests with other existing benchmarks such as the DBPedia SPARQL Benchmark (DBPSB). In addition, we are working on improvements in the efficiency of the parallelization of the MSC method. In particular, we are looking into combining multiple individuals that form part of the same connected component in the ABox in the same parallel task, since it can be seen in Definitions 2 and 3 that portions of the MSC computation can be shared among individuals provided that they are connected to each other.

6 Conclusion

In this paper, we have presented a parallel implementation of the enhanced MSC method and an extension to conjunctive queries using pattern matching, and we have evaluated execution time and efficiency as we varied the size of the data and the number of processors used. Since the method performs independent checking for every individual in the ABox, it is inherently parallelizable. The results show sub-linear performance with respect to the size of the ABox, which stems from its performance in linear time in the computation of the MSCs, and almost constant time for reasoning due to the small size of the resulting MSC.

Acknowledgements. This work is supported by grant # R44GM097851 from the National Institute of General Medical Sciences (NIGMS), part of the U.S. National Institutes of Health (NIH).

References

1. Horrocks, I.: Ontologies and the semantic web. Commun. ACM **51**(12), 58–67 (2008)
2. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Data complexity of query answering in description logics. Artif. Intell. **195**, 335–360 (2013)
3. Möller, R., Haarslev, V., Wessel, M.: On the scalability of description logic instance retrieval. In: Freksa, C., Kohlhase, M., Schill, K. (eds.) KI 2006. LNCS (LNAI), vol. 4314, pp. 188–201. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-69912-5_15
4. Motik, B., Shearer, R., Horrocks, I.: Optimized reasoning in description logics using hypertableaux. In: Pfenning, F. (ed.) CADE 2007. LNCS (LNAI), vol. 4603, pp. 67–83. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73595-3_6
5. Priya, S., Guo, Y., Spear, M., Heflin, J.: Partitioning OWL knowledge bases for parallel reasoning, pp. 108–115. IEEE, June 2014
6. Donini, F.M.: Complexity of reasoning. In: Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.) The Description Logic Handbook: Theory, Implementation, and Applications, pp. 96–136. Cambridge University Press, New York (2003)
7. Glimm, B., Horrocks, I., Lutz, C., Sattler, U.: Conjunctive query answering for the description logic SHIQ. arXiv:1111.0049 [cs], October 2011
8. Xu, J., Shironoshita, P., Visser, U., John, N., Kabuka, M.: Extract ABox modules for efficient ontology querying. arXiv:1305.4859 [cs], May 2013

9. Xu, J., Shironoshita, P., Visser, U., John, N., Kabuka, M.: Module extraction for efficient object queries over ontologies with large ABoxes. AIA **2**(1), 8–31 (2015)
10. Wandelt, S., Möller, R.: Towards ABox modularization of semi-expressive description logics. Appl. Ontol. **7**(2), 133–167 (2012)
11. Xu, J., Shironoshita, P., Visser, U., John, N., Kabuka, M.: Converting instance checking to subsumption: a rethink for object queries over practical ontologies. Int. J. Intell. Sci. **05**(01), 44–62 (2015). arXiv:1412.7585
12. Nebel, B.: Reasoning and Revision in Hybrid Representation Systems. LNCS (LNAI), vol. 422. Springer, Heidelberg (1990). https://doi.org/10.1007/BFb0016445
13. Donini, F., Era, A.: Most specific concepts for knowledge bases with incomplete information. In: Proceedings of CIKM, Baltimore, MD, USA, pp. 545–551, November 1992
14. Donini, F.M., Lenzerini, M., Nardi, D., Schaerf, A.: Deduction in concept languages: from subsumption to instance checking. J. Logic Comput. **4**(4), 423–452 (1994)
15. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, New York (2003)
16. Horrocks, I., Tessaris, S.: A conjunctive query language for description logic ABoxes. In: AAAI/IAAI, pp. 399–404 (2000)
17. Schaerf, A.: Reasoning with individuals in concept languages. Data Knowl. Eng. **13**(2), 141–176 (1994)
18. Kollia, I., Glimm, B., Horrocks, I.: SPARQL query answering over OWL ontologies. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011. LNCS, vol. 6643, pp. 382–396. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21034-1_26
19. Jing, Y., Jeong, D., Baik, D.K.: SPARQL graph pattern rewriting for OWL-DL inference queries. Knowl. Inf. Syst. **20**(2), 243–262 (2009)
20. Sirin, E., Parsia, B.: SPARQL-DL: SPARQL query for OWL-DL. In: In 3rd OWL Experiences and Directions Workshop (OWLED-2007) (2007)
21. Myung, J., Yeon, J., Lee, S.: SPARQL basic graph pattern processing with iterative MapReduce. In: Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud, MDAC 2010, pp. 6:1–6:6. ACM, New York (2010)
22. Schätzle, A., Przyjaciel-Zablocki, M., Hornung, T., Lausen, G.: PigSPARQL: a SPARQL query processing baseline for big data. In: Proceedings of the 12th International Semantic Web Conference (Posters & Demonstrations Track), ISWC-PD 2013, Aachen, Germany, vol. 1035, pp. 241–244. CEUR-WS.org (2013)
23. Artale, A., Calvanese, D., Kontchakov, R., Zakharyaschev, M.: The DL-lite family and relations. J. Artif. Int. Res. **36**(1), 1–69 (2009)
24. Ma, L., Yang, Y., Qiu, Z., Xie, G., Pan, Y., Liu, S.: Towards a complete OWL ontology benchmark. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 125–139. Springer, Heidelberg (2006). https://doi.org/10.1007/11762256_12
25. Guo, Y., Pan, Z., Heflin, J.: LUBM: a benchmark for OWL knowledge base systems. Web Semant. **3**(2–3), 158–182 (2005)
26. W3C: Large Triple Stores - W3c Wiki (2015)

Could Machine Learning Improve the Prediction of Child Labor in Peru?

Christian Fernando Libaque-Saenz[1](✉), Juan Lazo[1], Karla Gabriela Lopez-Yucra[2], and Edgardo R. Bravo[1]

[1] Universidad del Pacífico, Avenida Salaverry 2020, Jesús María Lima 11, Peru
{cf.libaques,jg.lazol,er.bravoo}@up.edu.pe
[2] Pontificia Universidad Católica del Perú, Av. Universitaria 1801,
San Miguel Lima 32, Peru
karla.lopez@pucp.pe

Abstract. Child labor is a relevant problem in developing countries because it may have a negative impact on economic growth. Policy makers and government agencies need information to correctly allocate their scarce resources to deal with this problem. Although there is research attempting to predict the causes of child labor, previous studies have used only linear statistical models. Non-linear models may improve predictive capacity and thus optimize resource allocation. However, the use of these techniques in this field remains unexplored. Using data from Peru, our study compares the predictive capability of the traditional logit model with artificial neural networks. Our results show that neural networks could provide better predictions than the logit model. Findings suggest that geographical indicators, income levels, gender, family composition and educational levels significantly predict child labor. Moreover, the neural network suggests the relevance of each factor which could be useful to prioritize strategies. As a whole, the neural network could help government agencies to tailor their strategies and allocate resources more efficiently.

1 Introduction

Child labor is a critical problem in developing countries because it could negatively affect economic growth [1]. Child labor has a negative effect on human capital, which is defined as “the stock of skills that the labor force possesses” [2]. Children who work have a high probability of becoming individuals with a low stock of skills in both quantity and quality [3]. In fact, these children (who work) usually do not dedicate their efforts to study and sometimes they do not even attend school at all. In turn, this low level of human capital and the associated lack of skills have a negative impact on individuals’ earnings and income [1]. Therefore, as a country’s human capital decreases, its economy decreases as well.

The term child labor refers to children working outside the home, paid or unpaid. Usually, unpaid child labor occurs inside the home – for example: working in family-owned enterprises or taking care of family members [4]. In developing countries, child labor predominantly arises because of (a) the need for

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 15–30, 2018.
https://doi.org/10.1007/978-3-319-90596-9_2

additional income to meet families' very basic subsistence requirements, (b) a response to the absence of credit markets, and (c) a perception that education yields low returns [5].

According to the International Labor Organization (ILO), in Latin America this phenomenon reached 12.5 million children and teenagers between 5 and 17 years old in 2014 [6]. Although this number has decreased from 20 million in 2010, an important fact is that the number of children working in dangerous activities has increased from 9 million in 2010 to 9.6 million in 2014 [6]. As for the case of Peru, the National Housing Survey (ENAHO in Spanish) shows that 21% of teenagers between 12 and 17 years old had been working in 2014 [6]. In other words, 1 out of 5 teenagers works in Peru.

Child labor can not only lead to gaps among countries but also within a country. In Peru, for example, the child labor rate in rural areas is twice as high as in urban areas [7]. By assessing child labor by region, Huancavelica presents the highest rate of child labor (58%), which is more than 10 times that for Tumbes (5%) – the latter is the region with the lowest rate of child labor [7]. Therefore, this phenomenon could negatively impact social and economic inclusion by increasing socioeconomic differences.

Programs and policies formulated by governments are the key to addressing these issues. It is also important to determine the probability of a child becoming or continuing as a worker because a correct classification could allow a better allocation of resources. There are various techniques to achieve this goal. We have traditional techniques such as logit models, and machine learning techniques such as neural networks. The principal difference is that the former capture linear effects, while the latter can capture non-linear relationships. It is important to have a model with high predictive capability, and therefore it is necessary to compare the predictive power of the different models.

Table 1 shows a summary of the issues covered by previous research in this field. All these studies used traditional techniques. For example, among theoretical contributions, Lima et al. [8] established that child labor would decrease as household income rises. Meanwhile, Emerson et al. [9] state that prior literature has assumed a trade-off between child labor and human capital accumulation to justify policy interventions. In the case of empirical findings, Susanli et al. [4] found that in Turkey the probability of child labor increases if the oldest child in the household is female, while living in rural areas makes this situation worse. In addition, a study based in Sao Paulo (Brazil) found that working while attending school has a detrimental effect on children's test scores [9]. Moreover, De Paoli and Mendola [10] found that low-educated parents are more likely to send their children to work. Finally, considering that poverty is not the sole antecedent to child labor, D'Alessandro and Fioroni [11] add that inequality also has an important effect. To the best of our knowledge, in this field there are few studies using machine learning techniques such as neural networks. For example, Rodrigues et al. [12] used data from Brazil and decision trees to search for patterns in the variables explaining child labor.

The objective of the present study is to compare the predictive power of traditional and modern models in regard to child labor (i.e., correctly identify

Table 1. Literature review

Author	Topic	Technique
Emerson and Souza (2002) [13]	Impact of gender on child labor	Bivariate probit model
Sapelli and Torche (2004) [14]	Determinants of school desertion (which usually leads to child labor)	Bivariate probit model
Lavado and Gallegos (2005) [15]	Characteristics of children with high probability of leaving the school	Parametric and non-parametric estimation through duration models
García (2006) [16]	Relationship between home responsibilities and work	Simultaneous equations through general Heckman model and bivariate probit model
Gunnarsson et al. (2006) [17]	Link between child labor and academic achievement	Ordered probit model, least square and instrumental equations
Alcázar (2008) [18]	Determinants of school desertion in rural areas	Instrumental variables and bivariate probit model
Rodriguez and Vargas (2009) [19]	Characteristics and nature of economic activity in child labor	Bivariate probit model
Lima et al. (2015) [8]	Effect of family wealth on child labor	Censored quantile instrumental variable
Le and Homel (2015) [20]	Impact of child labor on education performance	System of simultaneous tobit and ordered probit models
He (2016) [21]	Relationship between child labor and academic achievement	Quasi-maximum likelihood estimation
D'Alessandro and Fioroni (2016) [11]	Effect of inequality in political support for child labor regulation	Short-run general equilibrium model and long-run dynamics
De Paoli and Mendola (2016) [10]	Relationship between international labor outflows and child time allocation in developing countries	Panel data with country fixed effects
Susanli et al. (2016) [4]	Determinants of child labor and its relationship with mothers' employment	Bivariate probit models
Emerson and Souza (2017) [9]	Parents' decision to send their sons and daughters to work	Bivariate probit models

those children who work). It is expected that our results will shed light on the difference between models in terms of predictive power. By identifying the antecedents to child labor and the technique with the best predictive power, we will be able to provide recommendations to the Peruvian government.

2 Theoretical Background

Classification problems such as the child labor issue can be addressed by several techniques, both parametric and non-parametric. Parametric techniques (e.g., discriminant analysis, the logit model) require the prior specification of a function (or model) that relates the independent variables (X_i) with the dependent variable (Y). In practical terms, this function may be known – grounded in theory – or assumed. These techniques use observations of Y and X_i to estimate the parameters of the function. Once the parameters have been estimated, they can be used for prediction with new participants. One disadvantage of the parametric techniques is that they have a rigid structure (the mathematical function does not change and it only allows for estimating the parameters). Thus, these techniques may not be appropriate to represent phenomena that do not follow well-known mathematical functions.

In contrast, non-parametric techniques (e.g., artificial neural networks) do not assume a function a priori but instead approximate the function based on observation. Once the function has been approximated, it can be used to predict new cases. One relative advantage of these techniques is that they can represent complex non-linear mathematical functions. In other research arenas, this flexibility of non-parametric techniques has, under certain conditions, demonstrated the superiority of its predictive power over that of parametric techniques (e.g., Abdou et al. [22]; Altman et al. [23]).

Our research compares the logit model (parametric technique) with artificial neural networks (non-parametric technique) in the field of child labor. The application of these models for predictive purposes involves the following steps:

- The sample is randomly divided into two subsamples.
- The parameters of the model are estimated with one of the subsamples.
- The predictive capacity of the model (number of hits over total observations) is assessed.
- With these estimated parameters, prediction of the dependent variable for the other subsample is conducted.
- The predictive capacity of the model (with the test data) is assessed.

2.1 Logit Model

The logit model is a method that uses independent variables to estimate the probability of occurrence of a discrete outcome in the dependent variable [24]. According to the number of discrete outcomes, this technique can be divided into binary logit or multinomial logit models [24,25]. The former defines a dependent variable with two discrete outcomes whereas the latter represents a logit model with more than two discrete outcomes for the dependent variable [24,25]. In both cases, the discrete outcomes for the dependent variable should be mutually exclusive [24].

The logit model "has a straightforward and closed functional form that is easily estimated using maximum likelihood methods" [24, p. 475]. The logit technique does not assume restrictions on the normality of the distribution of variables [26]. Also, independent variables can be both continuous and categorical variables [24]. This technique is a special case of regression, which uses a transformation of the discrete dependent variable. This model assumes: (1) a categorical dependent variable with mutually exclusive outcomes, (2) independent variables can be continuous or categorical, (3) independence of observations, (4) absence of multicollinearity between independent variables, (5) a linear relationship between the continuous independent variables and the logit transformation of the dependent variable, and (6) absence of outliers.

The logit model is defined by the following function:

$$Logit(p_i) = Ln\left(\frac{p_i}{1-p_i}\right) = \alpha + X_i^T\beta + \varepsilon_i \tag{1}$$

where p_i is the probability that an observation takes a specific outcome of the dependent variable, α is the constant term; β is the corresponding vector of the coefficients; and ε_i is the error term.

2.2 Artificial Neural Networks

A neural network is, in a general sense, a machine designed to model the way in which the brain performs a particular task or function of interest [27]. The functioning of the brain is applied in this design because of its "(...) capability to organize its structural constituents, known as neurons, so as to perform certain computations (e.g., pattern recognition, perception, and motor control) many times faster than the fastest digital computer in existence today" [27, p. 23]. Therefore, a neural network resembles the brain mainly in two aspects: (1) the way knowledge is acquired by the network from its environment (i.e., learning process); and (2) the strength of interneuron connections (i.e., synaptic weights), which are used to store the acquired knowledge [27]. Accordingly, an artificial neural network is a physical cellular network that is able to acquire, store, and utilize experiential knowledge [28].

A fundamental unit in the operation of a neural network is the neuron. It is an information-processing unit which has three basic elements: a set of synapses or connecting links, each one with a weight or strength of its own; an adder for summing the input signals; and an activation function for limiting the amplitude of the output of a neuron [27, p. 32]. The neurons perform simple operations, transmitting their results to neighboring processors. Hence, the ability of a neural network to perform non-linear relationships between its inputs and outputs makes it a useful technique for pattern recognition and modeling of complex systems [29].

According to their topology, neural networks can be feedforward or feedback networks. In the former, the mapping goes from an input to an output layer instantaneously since there is no delay between them. This type of network is characterized by its lack of feedback which implies that the neural network has no explicit connection between layers [28]. In contrast, the latter has a connection between the output and input layers [28].

Another typology of neural networks is related to the learning paradigm which distinguishes between supervised learning and non-supervised learning. The first implies that the knowledge of the environment available to ‘the teacher’ is transferred to the neural network through training as fully as possible [27]. Also, it implies an error-correction learning in which the network parameters are adjusted under the combined influence of the training vector (i.e., example) and the error signal (i.e., difference between the desired response and the actual response of the network). This adjustment is carried out step by step in order to make the neural network emulate the teacher [27]. On the other hand, the second does not consider a teacher to oversee the learning process. In this case, there are no labeled examples of the function to be learned by the network. The learning of an input-output mapping is performed through continued interaction with the environment or based on the optimization of its parameters in order to develop the ability to form internal representations [27].

This research uses a Multilayer Perceptron neural network with a back-propagation algorithm which consists of applying a family of gradient-based optimization methods to find the optimal value of the weights based on minimizing the error norm between the desired output and the output calculated by the neural network [30]. In this type of network, the processing is performed by the inputs. The output obtained is compared to the expected output. From the obtained error, a process of adjustment of weights is applied, attempting to minimize the error.

2.3 Child Labor in Peru

The concept of child labor varies from country to country depending on the cultural context. According to the ILO, child labor refers to a work that is dangerous and harmful to the physical, mental, or moral wellness of the child, interfering with his/her education.

In the case of Peru, the minimum age for a child to be allowed to legally work is 14 years old, as long as these activities do not harm their integrity nor negatively impact their studies [6]. Also, they must have the permission of their parents or legal guardians to engage in these activities. In exceptional cases, children between 12 and 14 years old could also work as long as the work meets the same requirements [6]. In the present research, a child was considered to be a worker if he/she helps in the family business, in domestic tasks in a house that is not his or her own, in producing products to be sold, in agriculture activities, in selling products or providing services.

According to the National Housing Survey, child labor between 6 and 13 years old in rural areas (67.5%) is twice as prevalent as child labor in urban areas (32.5%). However, in the range from 14 to 17 years old, the values are similar (49.7% and 50.3% for rural and urban areas, respectively). Another important issue is that child labor rates significantly differ between cities. For example, Huancavelica is the city with the highest rate of child labor with 79.0%, followed by Puno, Huanuco, and Amazonas with 69.0%, 65.0%, and 64.0% respectively. Trujillo has the lowest child labor rate, at about 5.0%, which is significantly lower than the others. Not surprisingly, the cities with the highest rates of child labor are also those with the lowest incomes per capita. Furthermore, according to the National Institute of Statistics and Informatics (INEI in Spanish), economic activity for females (63.3%) is considerable lower than for males (81.4%).

Based on the above paragraph, we included variables capturing: (1) age and gender; (2) type of residence area such as urban/rural, region, stratum, and schooling available; and (3) socioeconomic variables such as expenses, education of the family head, type of housing, housing ownership, and housing status (adequacy, coverage of basic needs, sanitation). In addition, following [6], we included family characteristics as potential antecedents to child labor. Indeed, families where both parents work are less likely to have their children working, while the number of children could increase the probability that one or more children work. In these cases, the oldest child is the one with the highest probability of engaging in economic activities. Finally, current schooling status could also be a potential factor for child labor because those children who are behind in their studies are potentially engaged in other activities.

3 Research Method

3.1 Measurement Model

Table 2 defines our variables and shows the measurement items used in each one.

3.2 Data Collection and Analysis

Data were collected from the Peruvian National Housing Survey (ENAHO) for the year 2014. We eliminated the data for the months of January, February and March to eliminate seasonality. The rationale is that those months are holidays in Peruvian schools and thus the probability of child labor is high but does not imply that children stop studying to carry it out. Data include children between 12 and 17 years old at the national level who meet the following criteria: (1) is the son/daughter of the head of the family, and (2) he/she has not yet finished school.

For analysis, we used logit and neural networks techniques to find the antecedents to child labor and to classify children according to the probability of becoming a worker. We used these two techniques to compare predictive power because a correct prediction may allow governments to correctly allocate resources to deal with this problem. The first technique is based on linear relationships, while the latter can manage non-linear effects. Thus, differences in their results are expected. In the case of the logit model, we randomly divided the full sample into 2 subsamples: (1) a training subsample consisting of 85% of the full sample, and (2) a test subsample made up of the remaining 15%. We used the training subsample to calibrate the model (i.e., estimate the parameters of the function), and the test subsample to assess the predictive power of these results. In the case of neural networks, we randomly divided the sample into 3 subsamples: (1) a training subsample (70% of the total data), (2) a validation

Table 2. Measurement items

Variable	Description
Dependent variable	
Worker (WORK)	1 = If the child works 0 = If the child exclusively studies
Continuous independent variables	
Age (AGE)	Age of the child (in years)
Education of the family head (EDU_HEAD)	Level of schooling of the head of the family (in years)
Younger siblings (SIBLINGS)	Number of children under 5 years' old in the family
Family composition (COMPO)	Ratio of the number of adults (18 years old or older) to the number of children (younger than 18 years old) in the family
Education centers (CENTER)	Ratio of the number of education centers to the number of school-age children in the province of residence of the family
Monthly expense (EXPENSE)	Natural logarithm of the total monthly expense per family member
Categorical independent variables	
Maleness (MALE)	1 = If the child is male 0 = If the child is female
Urban (URBAN)	1 = If the residence of the family is located in the urban area 0 = If the residence of the family is located in a non-urban area
Oldest child (OLD_CHI)	1 = If the child is the oldest in the family 0 = If the child is not the oldest in the family
School backwardness (DELAY)	1 = If the child presents school backwardness 0 = If the child does not present school backwardness
Geographic area (AREA)	1 = North Coast 2 = Center Coast 3 = South Coast 4 = North Highlands 5 = Center Highlands 6 = South Highlands 7 = Jungle 8 = Lima Metropolitan Area

(*continued*)

Table 2. (*continued*)

Variable	Description
Geographic stratum (STRATUM)	1 = More than 100,000 dwellings 2 = From 20,001 to 100,000 dwellings 3 = From 10,001 to 20,000 dwellings 4 = From 4,001 to 10,000 dwellings 5 = From 401 to 4,000 dwellings 6 = 400 dwellings or fewer 7 = Composite rural area 8 = Simple rural area
Type of housing (TYPE)	1 = Independent house 2 = Apartment in building 3 = Chalet 4 = Neighborhood house 5 = Shack or cottage 6 = Improvised housing 7 = Non-housing premises 8 = Other
Housing ownership (OWN)	1 = Rented 2 = Owned by the family, totally paid 3 = Owned by the family, as result of squatting 4 = Owned by the family, paying off a loan 5 = Given by the workplace of one of the members 6 = Given by other family or institution 7 = Other
Housing inadequacy (ADEQ)	1 = If the housing is inadequate 0 = If the housing is adequate
Uncovered basic needs (UNMET)	1 = If the housing has unmet basic needs 0 = If the housing has not unmet basic needs
Absence of sanitation (HYGIENIC)	1 = If the house does not have sanitation 0 = If the housing has sanitation

subsample (15% of the total data), and (3) a test subsample (15% of the total data). We used the training and validation subsamples together to estimate the parameters of the model. To avoid overfitting and guarantee that the results of this stage could be generalized, we validated the predictive quality of the model with only the validation subsample every 1000 interactions. This process allows a better estimation of the weights of the network. Finally, we assessed the predictive power of the model with the test subsample.

4 Results

4.1 Logit Results

We conducted a preliminary analysis including all 17 independent variables. Results show that only 9 variables were statistically significant (variables with coefficients with p-value less than 0.05) in explaining the variance of our dependent variable (WORK). The other 8 variables (p-values higher than 0.05) were not considered in the subsequent analysis given that they do not have any impact on the dependent variable. Retained variables are divided into 6 categorical variables: URBAN, AREA, STRATUM, OWN, ADEQ, and UNMET; and 3 continuous variables: EXPENSE, EDU_HEAD, and SIBLINGS. We calculated the coefficients of the model using Eq. (1), where p_i is the probability that child i becomes a worker.

We assessed whether assumptions of logistic regression were met. Assumptions 1, 2, and 3 were determined by the model and data collection. For assumption 4, we conducted a linear regression to obtain VIF values. All VIF values were lower than 5 (the independent variable URBAN has the highest VIF value at 2.274). Therefore, there is no evidence of multicollinearity problems in our model [31]. For the fifth assumption, we used the Box and Tidwell [32] procedure. This procedure establishes that if the interaction between an independent continuous variable and its natural logarithm transformation is found to be significant, this variable is not linearly related to the logit of the dependent variable. In addition, following Tabachnick and Fidell's [33] recommendation, we used a Bonferroni correction for the statistical significance level by dividing it by the number of independent variables running this test including the constant term. This correction provided a significance level of 0.0038 (i.e., 0.05/13, where 0.05 is the original significance level and 13 is the sum of variables including the constant term: 1 constant term, 6 categorical independent variables, 3 continuous independent variables, and 3 interaction terms). P-values for the interaction terms were 0.688 for EDU_HEAD, 0.999 for SIBLINGS, and 0.0041 for EXPENSE. Based on this assessment, all p-values were over the value of 0.0038 and thus our model satisfied the linearity assumption. For the sixth assumption, we found 4 outliers of concern which were not considered in subsequent analysis.

Results of the logistic model are presented in Table 3. Our model is statistically significant ($\chi^2 = 2300.885, df = 25, p = 0.000$), and explains between 33.2% and 47.8% of the variance in child labor. In terms of predictive value, our model correctly predicted 82.61% of cases, with 55.80% of correct positive classifications (sensitivity) and 93.02% of correct negative classifications (specificity). Accordingly, our model has an efficiency (average of sensitivity and specificity) of 74.41% and a mean absolute percentage error (MAPE) of 17.39%. Although our model has an adequate overall predictive power, the Hosmer and Lemeshow goodness of fit test was significant ($\chi^2 = 39.889, df = 8, p = 0.000$) showing that it is poor at predicting the categorical outcomes. The reason for this finding may be the difference between sensitivity and specificity. Finally, coefficients (B) were found to be significant based on the Wald test. Table 3 also shows the standard error (SE) of the coefficients and their odd ratio (OR).

Table 3. Logistic regression with training sample

Variables	Model 1 (N = 700)			
	B	SE	Wald	OR
URBAN	3.649	0.441	68.563***	38.455
EDU_HEAD	−0.056	0.01	28.969***	0.946
SIBLINGS	0.181	0.081	4.978*	1.198
AREA	SS	SS	360.946***	SS
STRATUM	SS	SS	71.179***	SS
OWN	SS	SS	13.571*	SS
ADEQ	0.644	0.146	19.408***	1.905
UNMET	−0.366	0.117	9.721**	0.693
EXPENSE	−0.262	0.084	9.707**	0.769
Constant	−1.423	0.663	4.604	
−2log likelihood	4459.518			
Chi-square (Model)	2300.885*** (df = 25, p-value = 0.000)			
Hosmer & Lemeshow	39.889*** (df = 8, p-value = 0.000)			
Cox & Snell R2	33.20%			
Nagelkerke R2	47.80%			
Overall predicted %	82.61%			
Sensitivity	55.80%			
Specificity	93.02%			

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$
B = Coefficients; SE = Standard error; OR = Odds ratio
SS = Skipped for simplicity. (For categorical variables with more than 2 categories, there is a coefficient for each category.
We are choosing not to report them all because our focus is the predictive power of the model.)

We then assessed the model with our test subsample. Our model correctly predicted 80.64% of all the cases in this sample, with a sensitivity of 52.78%, a specificity of 91.88%, an efficiency of 72.33%, and a MAPE of 19.36%.

4.2 Neural Network Results

For purposes of comparison, we chose a simple neural network. Accordingly, we used a hidden layer with activation functions *Hyperbolic tangentsigmoidy*, and an output layer with activation functions Log-sigmoid. The value of weights and bias are updated according to gradient descent momentum and an adaptive learning rate. The training parameters of the neural network were: Maximum number of epochs to train: 40000, learning rate: 0.01, momentum constant: 0.7, performance goal: 10-5. These values were set following current literature [27,28]. They were also adjusted during the training process using an adaptive algorithm to find better parameters.

The first neural network used the 17 proposed independent variables (inputs). With the training and validation subsamples we obtained the best neural network made up of 38 neurons in the hidden layer and 1 neuron in the output layer. This model predicted 88.26% of all the cases, with a sensitivity of 90.97% and specificity of 87.21%. The efficiency of the model was thus 89.09%, and the MAPE was 11.74%. When applying this model to the test subsample, it predicted 85.11% of all the cases, with a sensitivity of 90.02%, a specificity of 79.42%, an efficiency of 84.72%, and a MAPE of 14.89%.

In addition, by analyzing the weight of the inputs of the neural network, we ranked the independent variables from the highest to the lowest effect: AREA (7.7), EXPENSE (7.4), HYGIENIC (7.4), STRATUM (7.0), MALE (6.2), OWN (6.0), SIBLINGS (5.9), TYPE (5.9), OLD_CHI (5.9), AGE (5.7), EDU_HEAD (5.5), COMPO (5.5), ADEQ (5.1), DELAY (5.0), CENTER (4.9), URBAN (4.6), and UNMET (4.4).

The second neural network used only the 9 variables that were statistically significant in the logit model for a straight comparison. In this model, with the training and validation subsamples we obtained the best neural network made up of 30 neurons in the hidden layer and 1 neuron in the output layer. Our model achieved 84.45% of correct total predictions, with a sensitivity of 79.61%, a specificity of 86.34%, an efficiency of 82.97%, and a MAPE of 15.55%. When using our model's parameters on the test subsample, it predicted 81.69% of all the cases, with a sensitivity of 78.86%, specificity of 84.23%, efficiency of 81.55%, and a MAPE of 18.31%.

For this model, the ranking of the inputs according to their weights is: AREA (12.9), STRATUM (12.6), EDU_HEAD (12.4), SIBLINGS (12.3), URBAN (10.9), ADEQ (10.6), UNMET (9.7), OWN (9.5), and EXPENSE (9.1).

4.3 Technique Comparison

The results of the previous section are summarized in Table 4. Considering that the logit model used 9 variables (8 were not considered because they have no significant impact on the dependent variable), the neural network used these same 9 variables and the same instances to ensure a fair comparison. In addition, Table 4 shows the results of the neural network technique with the complete 17 variables to assess if this non-linear model could extract important information from those 8 variables without a linear impact on the dependent variable. This table shows that overall neural network technique performed better than the logit model. In fact, the neural network obtained the highest values of accuracy (correct total - positive and negative - predictions). Also, considering that it is more important to predict when a child has high probabilities of becoming a worker than to predict that a child will be non-worker, sensitivity stands as our most important metric when comparing models. By an inspection of Table 4, sensitivity of the neural network technique was superior to the values obtained from the logit model. In spite of these results, the logit model was superior in terms of specificity. However, specificity is a metric for correct predictions of

Table 4. Comparison of results of predictive capacity in the test subsample

Predictive measures	Logit	Neural network	
	9 variables	9 variables	17 variables
Accuracy	80.64%	81.69%	85.11%
Sensitivity	52.78%	78.86%	90.02%
Specificity	91.88%	84.23%	79.42%
Efficiency	72.33%	81.55%	84.72%
MAPE	19.36%	18.31%	14.89%

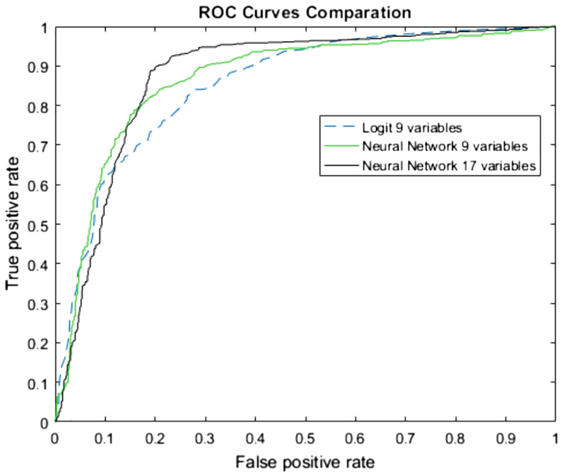


Fig. 1. ROC curves comparation

non-workers, which is not relevant in our case. In addition, Fig. 1 shows the ROC curve of prediction for these techniques.

Considering that the neural network model with 17 variables presents the higher sensitivity (i.e., the model that better predicts whether a child will work or not), we will focus on this model to analyze the relevance of the antecedents to child labor. Table 5 shows these 17 variables ordered by their relevance level. Table 5 also identifies with an "X" those variables with a linear effect captured by the logit model.

5 Discussion

Previous studies in this field have used linear statistical models to predict child labor. Our study shows that the use of machine learning techniques, such as neural networks, could provide better predictions which may lead to better decision making. Indeed, the results show that the neural network technique surpasses the logit technique in predictive capacity of child labor (sensitivity = 90.02% vs.

Table 5. Relevance of antecedents to child labor

Variable	Neural network (17 variables)	Logit (9 variables)
AREA	7.7	X
EXPENSE	7.4	X
HYGIENIC	7.4	
STRATUM	7	X
MALE	6.2	
OWN	6	X
SIBLINGS	5.9	X
TYPE	5.9	
OLD_CHI	5.9	
AGE	5.7	
EDU_HEAD	5.5	X
COMPO	5.5	
ADEQ	5.1	X
DELAY	5	
CENTER	4.9	
URBAN	4.6	X
UNMET	4.4	X

52.78%). This difference in predictive capacity (measured by sensitivity) is not marginal. Therefore, decision makers would do well to use neural networks over logit for the prediction of child labor.

Also, the results suggest that child labor has a more complex structure than that assumed by the logit technique. This complexity could be addressed by the neural network because it can capture non-linear relationships between variables. The logit technique thus discarded variables that the neural network considered relevant for prediction. In fact, logit models reject variables not fitting their base linear function. Neural networks, on the other hand, build relationships between variables in a flexible way, reducing the possibility of discarding relevant antecedents. Specifically, our findings show that the neural network model with 17 variables performed better than the 9-variable models (logit and neural network). This result suggests that this additional set of variables capture an important variability in explaining child labor. In other words, the neural network model with 17 variables does not ignore information that is relevant to the prediction. For example, the logit model dismissed variables such as age, gender and whether the child is the oldest sibling.

The neural network model shows that different types of variables predict child labor. These factors include geographical, economic and educational indicators, family composition and child characteristics. Specifically, the five most relevant variables for the prediction are the area where the child lives (AREA), the level of expenses per family member (EXPENSES), whether the house has hygienic services or not (HYGIENIC), the population density of the area where the child lives (STRATUM), and the gender of the child (MALE). Accordingly, our study highlights the factors that should be considered in formulating public policies and strategies that prevent child labor. For example, policies should consider that resources should be allocated to children living in specific regions and social levels, and that strategies should be different for girls and boys. In addition, various government sectors should participate in the formulation and implementation of strategies to deal with this problem because of its multifactorial nature (e.g., economic, geographic, and educational factors).

In short, considering that our study yields a more accurate prediction of this phenomenon, this could be useful for governmental agencies to develop more effective strategies and to be more efficient in investing scarce resources to deal with this problem. Moreover, our study could be used to determine the relevance of each factor. In turn, this hierarchy could further help government agencies to focus their strategies and resources. In addition, our research warns to decision makers to avoid discarding relevant factors when confronting this phenomenon.

References

1. Hanushek, E.A.: Economic growth in developing countries: the role of human capital. Econ. Educ. Rev. **37**, 204–212 (2013)
2. Goldin, C.: Human capital. In: Diebolt, C., Haupert, M. (eds.) Handbook of Cliometrics, pp. 55–86. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-642-40406-1_23
3. Becker, G.: Investment in human capital: a theoretical analysis. J. Polit. Econ. **70**, 9–49 (1962)
4. Susanli, Z.B., Inanc-Tuncer, O., Kologlugil, S.: Child domestic labour and mothers' employment in Turkey. Econ. Res. Ekonomska Istraživanja **29**(1), 967–979 (2016)
5. Dimova, R., Epstein, G.S., Gang, I.N.: Migration, transfers and child labor. Rev. Dev. Econ. **19**(3), 735–747 (2015)
6. López Yucra, K.G.: Determinantes del trabajo infantil y la deserción escolar en menores de 12 a 17 años en el perú para los años 2006 y 2014 (2016)
7. Sausa, M.: El trabajo infantil es más alto y más penoso en las zonas rurales. Perú 21 (2016)
8. Lima, L.R., Mesquita, S., Wanamaker, M.: Child labor and the wealth paradox: the role of altruistic parents. Econ. Lett. **130**, 80–82 (2015)
9. Emerson, P.M., Ponczek, V., Souza, A.P.: Child labor and learning. Econ. Dev. Cult. Change **65**(2), 265–296 (2017)
10. De Paoli, A., Mendola, M.: International migration and child labour in developing countries. World Econ. **40**(4), 678–702 (2017)
11. D'alessandro, S., Fioroni, T.: Child labour and inequality. J. Econ. Inequality **14**(1), 63 (2016)
12. Rodrigues, D.C., Prata, D.N., Silva, M.A.: Exploring social data to understand child labor. Int. J. Soc. Sci. Humanity **5**(1), 29 (2015)
13. Emerson, P.M., Souza, A.P., et al.: Bargaining over sons and daughters: child labor, school attendance and intra-household gender bias in Brazil. Technical report, Working Paper (2002)

14. Sapelli, C., Torche, A.: Deserción Escolar y Trabajo Juvenil: ¿Dos Caras de Una Misma Decisión? Cuadernos de economía **41**, 173–198 (2004)
15. Lavado, P., Gallegos, J.: La dinámica de la deserción escolar en el Perú: un enfoque usando modelos de duración. (05–08), September 2005
16. García, L.: The supply of child labor and household work, (31402) (2006)
17. Gunnarsson, V., Orazem, P.F., Sanchez, M.A.: Child labor and school achievement in Latin America. World Bank Econ. Rev. **20**, 31–54 (2006)
18. Alcázar, L.: Asistencia y deserción en escuelas secundarias rurales del Perú, pp. 41–82 (2008)
19. Rodríguez, J., Vargas, S.: Trabajo infantil en el perú. magnitud y perfiles vulnerables. informe nacional 2007–2008 (2009)
20. Le, H.T., Homel, R.: The impact of child labor on children's educational performance: evidence from rural Vietnam. J. Asian Econ. **36**, 1–13 (2015)
21. He, H.: Child labour and academic achievement: evidence from Gansu province in china. China Econ. Rev. **38**(C), 130–150 (2016)
22. Abdou, H., Pointon, J., El-Masry, A.: Neural nets versus conventional techniques in credit scoring in Egyptian banking. Expert Syst. Appl. **35**(3), 1275–1292 (2008)
23. Altman, E.I., Marco, G., Varetto, F.: Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience). J. Banking Finan. **18**(3), 505–529 (1994)
24. Lattin, J.M., Carroll, J.D., Green, P.E.: Analyzing Multivariate Data, vol. 1. Thomson Brooks/Cole, Florence (2003)
25. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. Wiley Series in Probability and Statistics. Wiley, Hoboken (2013)
26. Press, S.J., Wilson, S.: Choosing between logistic regression and discriminant analysis. J. Am. Stat. Assoc. **73**(364), 699–705 (1978)
27. Haykin, S.: A comprehensive foundation. Neural Netw. **2**, 41 (2004)
28. Zurada, J.M.: Introduction to Artificial Neural Systems. West, St. Paul (1992)
29. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press Inc., New York (1995)
30. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, pp. 318–362. MIT Press (1986)
31. Hair, J.F., Ringle, C.M., Sarstedt, M.: PLS-SEM: indeed a silver bullet. J. Mark. Theory Pract. **19**(2), 139–152 (2011)
32. Box, G.E., Tidwell, P.W.: Transformation of the independent variables. Technometrics **4**(4), 531–550 (1962)
33. Tabachnick, B., Fidell, L.: Multivariate analysis of variance and covariance. In: Using Multivariate Statistics. Allyn and Bacon, Boston (2007)

Impact of Entity Graphs on Extracting Semantic Relations

Rashedur Rahman[1](✉), Brigitte Grau[2], and Sophie Rosset[3]

[1] IRT SystemX, LIMSI, CNRS, Université Paris-Saclay, Orsay, France
rashedur.rahman@irt-systemx.fr
[2] LIMSI, CNRS, ENSIIE, Université Paris-Saclay, Orsay, France
brigitte.grau@limsi.fr
[3] LIMSI, CNRS, Université Paris-Saclay, Orsay, France
sophie.rosset@limsi.fr

Abstract. Relation extraction (RE) between a pair of entity mentions from text is an important and challenging task specially for open domain relations. Generally, relations are extracted based on the lexical and syntactical information at the sentence level. However, global information about known entities has not been explored yet for RE task. In this paper, we propose to extract a graph of entities from the overall corpus and to compute features on this graph that are able to capture some evidences of holding relationships between a pair of entities. The proposed features boost the RE performance significantly when these are combined with some linguistic features.

1 Introduction

Relation extraction (RE) from text is a useful task for populating Knowledge Base about entities. Many relations exist between pairs of entities and RE systems learn how the relations between entity-mentions are expressed in texts. RE systems make use of linguistic features based on semantic [1] and syntactic analysis [2,3]. Recently neural networks have been applied for RE task that use word embeddings for semantics without requiring complex feature engineering [4]. These methods use local information at the sentence level but do not account for global information on the entities at collection level. Recent work on Web RE [5] studied global information about the object entity and words around the entity-mentions. Such information facilitates introducing some sort of world knowledge for making choices in addition to modeling the linguistic expression of relation in sentences.

We hypothesize that a pair of entities (subject and object) having a true relationship should share more common neighbors than a false relationship between that particular subject and a different object. For example, the spouse of a person should share more places and relationships with his/her spouse than with a person who has no true relation with him/her. Therefore, we construct a graph of entities based on a corpus that allows us to propose new characterizations of the relations by community graph-based features [6,7], in addition to newly defined

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 31–47, 2018.
https://doi.org/10.1007/978-3-319-90596-9_3

linguistic features. In [8], we have shown the effectiveness of graph based features for validating claimed relations on a relatively small dataset and a large number of relations. Thus we go further in our study in order to enlarge the training and test data and closely observe the impact of graph features on extracting some semantic relations.

For evaluating the relevance of the proposed features, we tested them on a task of relation validation (RV). RV examines the correctness of relations that are extracted by different RE systems. It facilitates to evaluate the new features without developing a complete RE system. In this paper, we add a new kind of evaluation and we evaluate the proposed features on knowledge base population (KBP) where the validation results are used for choosing entities that fill relations linked to a query entity. Experimental results show that the newly proposed features lead to outperform the RV baseline by around 10 points. The KBP evaluation also shows improvement over state-of-the-art system results by 1.5 point.

2 Related Works

Several features have been explored for relation extraction (RE) from texts. Existing RE methods basically extract linguistic evidences of holding relationship between two objects at the text level based on syntactic and semantic analysis.

Syntactic analysis captures the grammatical structures of expressing relations among different words in a sentence. Therefore, syntactic dependency has been widely explored for RE task [9,10]. In [2], dependency tree has been used for defining kernel functions based on the shortest dependency path between two entities. Sometimes, shortest path fails to capture enough information for RE therefore, a context-sensitive convolution tree kernel [9] was proposed to include necessary information outside the shortest path. In open information extraction [11], dependency parsing was employed to define some patterns of relations and to discover verb-clauses at sentence level.

However, syntactic information cannot characterize the semantic type of a relation. Therefore, lexical features i.e. words between and around the mentions are effective [1,12–14] for RE task. Dependency trees and trigger words were combined to take advantage of both syntactic and semantic features for biomedical RE in [15].

Dependency and lexical features have been used in the existing rule based [3,11] and supervised [1,9] methods of RE. Some feature based RE methods [14,16,17] used POS-tags in addition to the dependency and word features. Rule based methods are restricted to extract a small number of relations while supervised methods are very effective but require a large amount of labeled data. Distant supervision [12,16,17] does not require manually labeled data for learning relations. These methods inherit state-of-the-art linguistic features and apply some probabilistic models for extracting relations. Nowadays, patterns and semantic types of relations are learned automatically by employing word embeddings and neural networks [18–20].

However, no existing method used entity level global information for RE task. Some kind of global information about the object entities has been studied in Web RE task [5]. Global information about entities gives some clues how the entities are associated among them. Such information can be explored by representing entities as nodes in a graph.

A graph structure facilitates analyzing paths between nodes and relationships among them. Several graph based methods have been proposed for different tasks i.e. automatically completing existing knowledge base [21,22], automatic trigger identification for slot filling [23], entity linking [6,24] etc.

Several features have been computed on graph i.e. entropy for discovering knowledge in publication networks [25], centrality measurements for finding important and influential nodes in social networks [7] etc.

We construct a graph of entities after extracting named entities from a collection of texts and we propose some new features for RE which are computed on the graph of entities by analyzing the communities of pair of entities.

3 Community Graph of Entities

3.1 Definition of the Graph

Let a graph $G = (E, R)$, a query relation (slot) r_q, a query entity $e_q \epsilon E$, candidate responses $E_c = \{e_{c1}, e_{c2}, \ldots, e_{cn}\} \epsilon E$ where $r_q = r(e_q, e_c) \epsilon R$. The list of candidates is generated by different relation extraction systems. Suppose other relations $r_o \epsilon R$ where $r_o \neq r_q$. We characterize whether a candidate-entity e_{ci} of E_c is correct or not for a query relation (r_q) by analyzing the communities X_q and X_c formed by the query entity and each candidate response. A community X_i contains the neighbors of e_i, and this up to several possible steps.

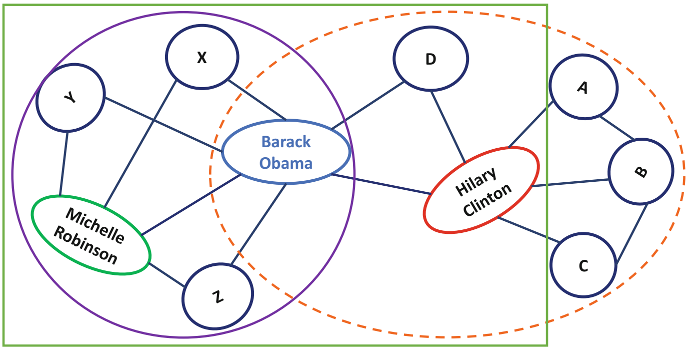


Fig. 1. Community graph (Color figure online)

Figure 1 shows an example of such type of graph where the entity of a query, its type and relationship name are *Barack Obama*, *person* and *spouse* accordingly. The candidate responses are *Michelle Robinson* and *Hilary Clinton*. The objective is to classify *Michelle Robinson* as the correct response based on their community analysis. The communities of *Barack Obama* (green rectangle), *Michelle Robinson* (purple circle) and *Hilary Clinton* (orange ellipse) are defined by *IN_SAME_SENTENCE* relation which means the pair of entities are mentioned in the same sentence in texts. The graph is thus constructed from untyped semantic relationships based on co-occurrences. It would also be possible to use typed semantic relationships provided by a relation extraction system or a knowledge base.

3.2 Construction of the Community Graph

The graph of entities as illustrated in Fig. 1 is created from a graph representing the knowledge extracted from the texts (lower part of Fig. 2) called knowledge graph. The knowledge graph represents documents, sentences, mentions and entities as nodes and the edges between these nodes represent relationships between these elements. This knowledge graph is generated after applying systems of named entity recognition (NER) and sentence splitting.

Recognition of named entities is done by using Stanford system [26] and *Luxid* of ExpertSystem[1]. Luxid is a rule-based NER system that uses some external information sources such as Freebase, geo-names etc. and performs with high precision. It decomposes the entity mentions into components, such as *first name*, *last name* and *title* for a *person* named entity and classifies *location* named entities into *country*, *state/province* and *city*. When the two NER systems disagree, as in (Stanford: location, Luxid: person), we choose the annotation produced by Luxid because it provides more precise information about the detected entity than Stanford does.

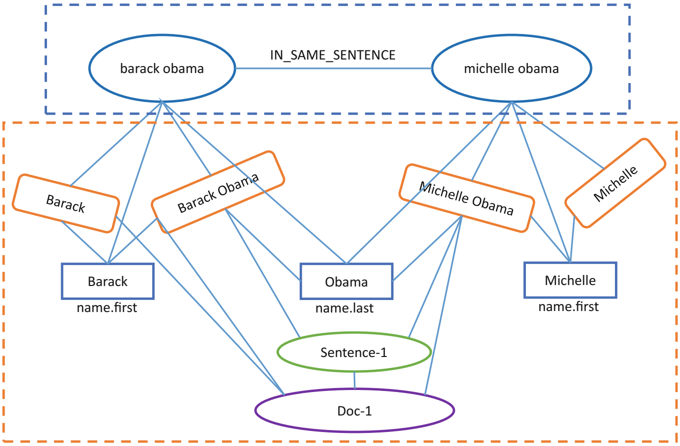


Fig. 2. Knowledge graph

[1] http://www.expertsystem.com/.

Multiple mentions of the same entity found in the same document are connected to the same entity node in the knowledge graph, based on the textual similarity of the references and their possible components, which corresponds to a first step of entity linking on local criteria. This operation is performed by Luxid. However, an entity can be mentioned in different documents also with different forms (e.g. *Barack Obama*, *President Barack Obama*, *President Obama* etc.) which creates redundant nodes in the knowledge graph. Entities are then grouped according to the similarity of their names and the similarity of their neighboring entities calculated by Eq. 4. This step groups the similar entities into a single node in the community graph (upper part of Fig. 2). This latter graph is constructed from the information on the entities and relations present in the knowledge graph and the link with the documents is always maintained. It is thus possible to know the number of occurrences of each entity and each relation. The graph is stored in a Neo4j database, a graph-oriented database, which makes it possible to extract the subgraphs linked to an entity by queries. We only consider as members of the communities the entities of type person, location and organization.

4 Relation Validation

In order to predict whether a relationship is correct or not, we consider this problem as a binary classification task based on three categories of information. We calculate a set of features using the graphs (see Sect. 4.1), to which we add features based on a linguistic analysis of the text that justifies the candidate and describes the relationship (see Sect. 4.2) and an estimation of trust on the candidates (voting) according to the frequencies of them in the responses of each query.

4.1 Graph-Based Features

We explore information at entity level based on community graph analysis. We assume that a true object is an important member in the community of the subject entity. A community X_e of a subject is defined by the sub-graph formed by its neighbors up to several levels. A merging of the communities of two particular entities includes all the neighbors of that pair of entities. We, therefore, define different features related to this hypothesis. We compute 4 features on the community graph (a) network density (b) eigenvector centrality (c) mutual information and (d) network similarity.

Network density (Eq. 1) computes the degree of connectivity among the nodes of the network. A network gets high density score if there are many connections among the neighbor nodes.

$$\rho_{X_e} = \frac{number\, of\, existing\, edges\, with\, e}{number\, of\, possible\, edges} \tag{1}$$

We hypothesize that the density of the community of a true object merged with the community of the subject entity must be higher than the density of a false object community merged with that of the same subject because the subject and true object shares more neighbors that makes more edges among the network nodes. According to the Fig. 1 the merged community of *Michelle Robinson* and *Barack Obama* is more dense than the merged community of *Hilary Clinton* and *Barack Obama.*

Eigenvector centrality [27] measures the influence of a node in a graph. A node will be even more influential if it is connected to other influential nodes. We hypothesize that the subject would be more influenced by a true object than by a false object since the true object shares more community members with the subject and becomes highly influential. We measure the influence of an object in the community of the subject by calculating the absolute difference between the eigenvector centrality scores of the subject and object. We assume that this difference should be smaller for a true object than for a false object because the subject and the true object should get similar score according to their influence to each other. Suppose $A = (a_{i,j})$ is the adjacency matrix of a graph G. The eigenvector centrality x_i of node i is calculated recursively by Eq. 2.

$$x_i = \frac{1}{\lambda} \sum_k a_{k,i} x_k \tag{2}$$

where, $\lambda \neq 0$ *is a constant and the equation can be expressed in matrix form:* $\lambda x = xA$.

Mutual information quantifies the amount of information gained by a random variable compare to another random variable. We compute mutual information gained by the community of an object through the community of the subject to capture the evidence of having relationship between them by using the Eq. 3.

$$MI(X_q, X_c) = H(X_q) + H(X_c) - H(X_q, X_c) \tag{3}$$

$$where,\ H(X) = -\sum_{i=1}^{n} p(e_i)\ log_2(p(e_i))$$

$$and\ p(e) = \frac{number\,of\,edges\,of\,e}{number\,of\,edges\,of\,X}$$

and X_q and X_c *are the communities of the subject and object entity respectively, and p(e) is the probability of degree of centrality of a community-member.*

We hypothesize that the mutual information of a true object should be higher than a false object because its community shares more edges to the community of the subject.

Network similarity computes the similarity between two communities in term of common neighbors by using Eq. 4.

$$similarity = \frac{|X_q \cap X_c|}{\sqrt{|X_q||X_c|}} \tag{4}$$

where, X_q and X_c *are the community members of the subject and object entity accordingly.*

The similarity gets higher score if both communities share large number of neighbors. Thus we hypothesize that the similarity of the communities of a subject and a true object should be higher than with a false object.

However, sometimes the value of the network density or similarity between the communities of a subject and a false object may be higher than that between the subject and a true object. For example, the network density or similarity score between *Barack Obama* and *Hillary Clinton* can be higher than the score between *Barack Obama* and *Michelle Obama* based on the existence of other members of their networks in the corpus. Moreover, the same pair of entities may have multiple relations that cannot be distinguished by graph analysis. Basically, graph based features are useful to compute the degree of association between the pairs of entities but these features do not hold the semantics of different relation types. Therefore, we need to define some linguistic features for characterizing the true meaning of relations.

4.2 Linguistic Features

We analyze two kinds of linguistic features for RE: syntactic and semantic. Syntactic features are able to give some clues for assessing if a relation exists between a pair of mentions. The semantics of a relation is captured by trigger words and thus we will analyze the sentence at lexical level.

Syntactic dependency analysis facilitates to compute the syntactic features, i.e., the parser [26] provides a tree in which nodes are the words of the sentence and the edges between them are labeled by their syntactic role. The consecutive dependency labels between a pair of mentions in a sentence form a pattern of the relation that is expressed by the sentence. Such pattern can be repeated for expressing the same relationship between a different pair of mentions.

We extract a list of dependency patterns for each relation and simplify them as we did in [8]. However, it is hard to capture all the dependency patterns since the relations are expressed in many different ways. Therefore, we compute minimal edit distance of an unknown pattern to the known patterns so that the unknown pattern can be considered as a member of one of the known pattern groups by an approximation. We propose the minimum edit distance as a feature and call it dependency pattern edit distance (DPED). Since relations are often expressed in short dependency paths, the length of the simplified dependency path is considered as a feature.

The semantic analysis is performed based on positive triggers associated to the relation types. Positive triggers refer to the keywords that strongly characterize a particular relation. For example, *wife, husband, married* are positive triggers for a *spouse* relation.

Since the relations are expressed by a variety of words, it is hard to collect all the positive triggers for a relation. Therefore, we associate a word embedding to each trigger by using the *GloVe*[2] model. Thus, deciding if a word is a trigger or not relies on the similarity of their embeddings. Suppose, *[a, b]* are two words between the subject and object mentions in a sentence and *[x, y, z]* the pre-collected positive triggers for the claimed relation. We compute the cosine similarity between the vectors of each pair of *[(a,x), (a,y), (a,z), (b,x), (b,y), (b,z)]* and take the best similarity score as a feature. If a word between the pair of mentions completely matches to one of the pre-collected positive triggers the similarity score is 1.0. We inspect the existence of any positive trigger in three cases: (1) between the mentions at surface level (2) in the dependency path and (3) in the minimum subtree as in [15]. We define the baseline feature set by combining the path length of simplified dependency with these three features.

5 Data

This section describes two kinds of data. Firstly, data that have been used for computing the linguistic and graph features and secondly, the datasets used for training and testing several models.

5.1 Data for Computing Features

For linguistic features, we use the 2014 assessed corpus of TAC-KBP English cold start slot filling (CSSF) which contains examples of correct relations for around 38 relationships between 1,020 entity pairs. Trigger words and dependency patterns (as discussed in Sect. 4.2) of different relations were collected from this dataset. We collected the words between the subject and object pairs of positive responses of each kind of relation and ranked them by counting their frequencies. We observed that for some relations (i.e. *spouse*) the top 5 words are discriminating (i.e. *married*) for characterizing the semantics of the relation. We call these relations trigger-dependent relations. In contrast, we notice that for other relations (i.e. *city_of_residence*) top 5 words are either prepositions or other words that are not able to distinguish any semantic relation. Such relations are called trigger-independent relations. Thus we selected 12 trigger-dependent relations (first column of Table 1) for our study. We obtain in total 76 trigger words and 286 patterns for these relations.

For computing the entity graph as discussed in Sect. 3.2, we used the reference corpus of TAC-KBP CSSF evaluation. We parsed around 50,000 and 30,000 documents provided for the CSSF-2015 and CSSF-2016 evaluation tasks accordingly. Both corpus included texts from newswire and discussion forums. Two knowledge graphs have been built from these datasets and there are 152,583 and 65,389 entities (*person, organization* and *location*) in the 2015 and 2016 graphs accordingly. Moreover, the knowledge graphs consist of 805,216 and 488,198 edges of *IN_SAME_SENTENCE* relations among different entity mentions accordingly.

[2] https://nlp.stanford.edu/projects/glove/.

5.2 Data for Training and Testing the Models

In our experiments, we use subsets of TAC-KBP CSSF datasets of 2015 and 2016 for training and testing accordingly. The CSSF task requires a participant system to respond to a set of queries. Each query is about an entity (subject), associated with the slot (relationship) to fill. A system responses to a query by providing an object value, an object type, an object offset, the relation provenance offset and a confidence score. The relation provenance is an excerpt of a document that justifies the claimed relation. The relation provenance offsets of a response is not guaranteed to delimit a complete sentence. Thus we extract the complete sentence corresponding to the relation provenance offset snippet from the source document as several features have to be computed on the complete sentences.

A lot of queries have been answered with only wrong responses by different systems. Therefore, we keep queries that have been answered with at least one correct response. We also filter out queries that are not relevant to the relations we study in our experiments. In order to build a set of positive and negative examples for training and testing, we extracted answers corresponding to those queries from the systems assessment files. An assessment file contains the indication whether an object and a relation provenance text of a query relation are correct or not. There are many wrong responses to the queries regarding the amount of correct ones. Therefore, we reduce the number negative examples to construct a balanced training dataset. We randomly select a subset of wrong responses from each query of the training dataset. After removing the duplicate responses, the ratio of positive and negative responses is around 2/3. Similar process of extracting positive and negative examples has been applied for building the test dataset but we do not filter any negative example.

Since graph based features depends on the performance of NER systems that could be not good enough to detect all the named entities mentioned in the queries and responses, our system could compute graph features for a small number of responses. In order to compute features on the graphs, we defined two strategies i.e. *hard constraint* and *relaxed constraint* in terms of connectivity in a graph between the subject and object entities of a relation under inspection.

Hard Constraint: Usually, a relation between two entities is expressed when both of the entities are mentioned in the same sentence. Therefore, in our preliminary study, we constrained the system to find an IN_SAME_SENTENCE link (in the knowledge graph) between the subject and object entities of a relationship under observation for computing graph features. Thus we could compute graph features for around 14% of the responses. We obtained 2,274 (827 positive and 1,447 negative) instances for training from 130 queries of the 12 trigger-dependent relations. In this setting, the test dataset counted 3,429 (262 positive and 1,167 negative) instances from 63 queries for the same number of relations. The number of training instances (positive and negative) for different relations are very different. Moreover, some relations (i.e. *per:spouse, org:member_of*) count very small number of training instances and some (i.e. *per:children, country_of_death*) have no training example at all. Therefore, in the setting of hard constraint we include training examples of some other relations to the instances of the 12 selected relations. We obtained in total 3,481 (1,268 positive and 2,213 negative) instances from 260 queries of 19 relations. Our experiment on this dataset obtained poor result (it will be discussed in Sect. 6.1) because of small number of training examples. Therefore, we defined a strategy for increasing the number of examples.

Relaxed Constraint: We relax the constraint of IN_SAME_SENTENCE between the subject and object entities of a relation. If the entity pairs are not connected by an IN_SAME_SENTENCE link in the graph we forcefully connect them by creating the link before computing graph features and delete the link after completing the feature computation of the entity pair. Thus our system could compute graph features for around 50% of the responses. Relaxed constraint significantly increases both training and test instances for all the relations as shown in Table 1. We obtain a training dataset that counts in total 14,804 (5,933 positive and 8,871 negative) instances from 411 queries of the 12 trigger-dependent relations. In similar way, our test dataset counts 1,109 and 4,827 positive and negative instances accordingly from 223 queries.

6 Results and Discussion

We performed experiments on the 12 trigger-dependent relations as discussed in Sect. 5.1. We select these relations to measure the effect of the proposed graph-based features when they are combined with linguistic features.

We evaluate our features for characterizing a relation in the setting of a relation validation (RV) task. The RV method takes as input a text snippet, subject, object and the claimed relation name and outputs true if the snippet holds the claimed relation otherwise false. We include a voting feature (as done in [8]) to the linguistic and graph features to observe the trustworthy influence of multiple systems on validating a claimed relation.

Furthermore, we evaluate the contribution of our RV model for a knowledge base population (KBP) task that will be discussed in Sect. 6.2. Both tasks RV and KBP are evaluated by standard precision, recall and F-score.

6.1 Results on the Relation Validation Task

In this section, we want to observe the effectiveness of adding more training data on relation validation task, performance of different classifiers for binary classification and impact of the proposed features on relation extraction which is evaluated as a relation validation task.

We want to inspect the efficiency of relaxed constraint of IN_SAME_SENTENCE link between subject and object entities over hard constraint to improve the classification performance. We expect that the relation validation system would learn and perform better by training with and testing on more data accordingly.

Table 1. Comparison of relation validation performance between hard and relaxed constraints

Relation name	Hard constraint						Relaxed constraint					
	# Train data		# Test data		F	Acc.	# Train data		# Test data		F	Acc.
	Pos.	Neg.	Pos.	Neg.			Pos.	Neg.	Pos.	Neg.		
per:parents	35	12	17	15	66.7	65.6	148	229	94	386	50.0	77.1
per:children	0	0	0	221	0.0	99.1	67	93	37	630	65.9	95.7
per:spouse	2	1	0	0	-	-	155	298	25	106	49.1	79.4
per:country_of_death	0	0	6	114	53.3	94.2	77	148	72	189	88.3	93.1
per:country_of_birth	14	28	1	65	8.0	65.2	108	140	5	260	80.0	99.3
per:city_of_death	56	100	0	0	-	-	243	398	30	227	51.4	86.0
per:city_of_birth	141	281	70	22	86.8	77.2	485	814	139	90	91.8	90.4
per:employee_or_member_of	211	355	16	232	15.4	73.4	2,517	3,267	287	1,538	50.4	80.8
org:top_members_employees	68	78	29	30	96.7	96.6	461	743	61	277	63.4	79.9
org:member_of	8	14	0	0	-	-	571	917	27	389	58.1	91.4
org:country_of_headquarters	158	310	82	334	23.0	74.3	471	822	140	362	54.3	78.9
org:city_of_headquarters	134	268	41	134	44.6	58.9	630	1,002	192	373	74.7	83.0
All Together	827	1,447	262	1,167	51.5	78.2	**5,933**	**8,871**	**1,109**	**4,827**	**63.3**	**84.8**

Table 1 represents the statistics of training and test dataset, F-score (F) and accuracy (Acc.) regarding both hard and relaxed constraints. This table shows the scores obtained by Random Forest classifier which is trained by the best feature combination, i.e. $Voting + Linguistic + Graph$. Relaxed constraint significantly increases both training and test instances for all the relations as discussed in Sect. 5.2. The F-score and accuracy are gained over almost all the relations by relaxing the IN_SAME_SENTENCE constraint. In the results on hard constrained dataset, we notice that several relations do not have any training examples (as *per:children* or *per:country_of_death*), or test data (as *per:spouse, per:city_of_death, org:member_of*). We see that relaxed constraint results better F-score for all the relations except *per:parents* and *org:top_members_employees*. We obtain overall F-score and accuracy of 51.5 and 78.2 accordingly by hard constraint. In contrast, the relaxed constraint improves these scores by around 12 and 6 points accordingly. We achieve overall F-score and accuracy of 63.3 and 84.8 accordingly by relaxing the constraint. Thus relaxation of IN_SAME_SENTENCE constraint between subject and object entities facilitates to train a model with more data that significantly improves the performance to classify the correct and wrong relations.

We also compare the classification performances of different classifiers e.g. LibLinear, SVM, Naive Bayes, MaxEnt and Random Forest based on the best feature combination (Voting+Linguistic+Graph) on the dataset of relaxed constraint as shown in Table 2. We achieve the best precision (57.8), F-score (63.3) and accuracy (84.8) by Random Forest classifier although it gets a lower recall (70.1) compared to other classifiers. The best recall of 76.5 is resulted by Naive Bayes which obtains the third highest F-score (59.0) and accuracy (80.6).

However, it may not be system-independent to use the voting feature to realize a task of relation extraction. Usually, a relation extractor does not employ

Table 2. Relation validation performances by different classifiers

Classifier	Precision	Recall	F-measure	Accuracy
LibLinear	45.6	73.9	56.1	79.1
SVM	49.8	73.5	59.4	81.6
Naive Bayes	48.0	76.5	59.0	80.6
MaxEnt	48.3	69.5	57.0	80.6
Random forest	**57.8**	**70.1**	**63.3**	**84.8**

multiple systems for generating relation hypothesis. Therefore, we discard the voting feature in this RV evaluation to realize the contribution of proposed features for relation extraction. We define a baseline (BL) by four linguistic (semantic and syntactic) features as discussed in Sect. 4.2 and observe relation validation performances through the linguistic baseline, the proposed linguistic and graph features and their combinations. Since Random Forest results the best score over several classifiers, we observe the performances of different feature sets by this classifier.

Table 3 represents the classification scores where we observe that the combination of BL and proposed graph features outperforms the BL almost for all the relations except *per:country_of_death*. We obtain overall F-score of 58.60 by BL+Graph that is around 9 points higher than the BL. The experimental results also show that the combination of BL and dependency pattern edit distance (DPED) improves the overall F-score by 1.79 point over the BL. This combination achieves higher F-score for 7 relations (among 12) which indicates the effectiveness of DPED for RV task. Basically, we gain higher precision by allowing a slight drop of recall that results better F-score over the BL. The best F-score is achieved by the combination of BL, DPED and graph (BL+DPED+Graph). This combination results overall F-score of 59.90 which is around 10 points higher than the BL. We observe that BL+DPED+Graph obtains higher F-score for 11 relations compare to the BL. For only one relation (*per:country_of_death*) the classification performance remains same as the BL.

We notice in Table 3 that BL+Graph and BL+DPED+Graph obtain a surprising performance for *per:country_of_birth* over the BL. Both BL+Graph and BL+DPED+Graph achieve an F-score of 88.89 which is around 69 points higher than the BL. The reason behind this result is that we have a very small number of true instances for this relation compare to the number of false instances (as shown in Table 4) and a high precision is resulted by discarding large number of false relations.

We achieve the highest precision almost for all the relations by BL+DPED+Graph. BL+DPED+Graph achieves overall precision of 66.02 that is around 21 points higher than the BL that indicates the proposed features discard large number of false relation instances correctly. A little drop of recall is caused by BL+DPED+Graph which is around 1 point less than the BL. The recall of 55.73 and 54.82 are resulted by the BL and BL+DPED+Graph

Table 3. Classification performances by different feature sets

Relation name	BL			BL + DPED			BL + Graph			BL + DPED + Graph		
	P	R	F	P	R	F	P	R	F	P	R	F
per:parents	37.30	73.40	49.46	40.12	69.15	50.78	51.75	78.72	**62.45**	45.65	67.02	54.31
per:children	62.22	75.68	68.29	60.87	75.68	67.47	70.00	75.68	72.73	93.33	75.68	**83.58**
per:spouse	36.23	100	53.19	73.53	100	84.75	65.00	52.00	57.78	68.42	52.00	**59.09**
per:country_of_death	98.55	94.44	**96.45**	98.55	94.44	**96.45**	97.50	54.17	69.64	98.55	94.44	**96.45**
per:country_of_birth	11.43	80.00	20.00	12.90	80.00	22.22	100	80.00	**88.89**	100	80.00	**88.89**
per:city_of_death	58.00	96.67	72.50	71.05	90.00	79.41	75.00	90.00	**81.82**	75.00	90.00	**81.82**
per:city_of_birth	97.20	100	98.58	97.18	99.28	98.22	100	99.28	**99.64**	100	99.28	**99.64**
per:employee_or_member_of	20.74	29.27	24.28	20.63	27.53	23.58	34.85	24.04	28.45	32.88	25.09	**28.46**
org:top_members_employees	39.39	63.93	48.75	35.78	63.93	45.88	52.38	90.16	66.27	59.14	90.16	**71.43**
org:member_of	28.57	44.44	34.78	38.71	44.44	41.38	48.00	44.44	46.15	50.00	44.44	**47.06**
org:country_of_headquarters	52.17	25.71	34.45	59.02	25.71	35.82	75.93	29.29	**42.27**	75.93	29.29	**42.27**
org:city_of_headquarters	50.00	44.27	46.96	60.14	43.23	50.30	86.67	47.4	**61.28**	89.69	45.31	60.21
All Together	44.75	**55.73**	49.64	48.55	54.46	51.34	65.09	53.29	58.60	**66.02**	54.82	**59.90**

accordingly. The drop of recall indicates the limitations of graph features to hold the semantic evidences of some relations.

Table 4 illustrates the confusion matrix resulted by BL and BL+DPED+Graph where we compare the number of true positive (TP), false negative (FN), false positive (FP), true negative (TN) and accuracy (Acc.). We see that the baseline and BL+DPED+Graph methods correctly classify overall 618 and 608 true relation instances accordingly among 1,109. That means BL+DPED+Graph discards 501 true relation instances which is around 1% more than the BL. However, the BL and BL+DPED+Graph correctly discard overall 4,064 and 4,514 false relation instances respectively among 4,827. The rate of discarding false relation instances by BL+DPED+Graph is around 9% higher than the BL which contributes to increase the overall precision and finally achieves a high accuracy. While observing the accuracy relation-by-relation we see a significant improvement achieved by BL+DPED+Graph over the BL for all the relations.

Table 5 presents classification results on some claimed relations from the test data that helps to realize the performance of our RV model. The first and second row show two correctly classified true claims of *spouse* and *children* relation accordingly. Furthermore, a false claim of *spouse* relation has been detected as wrong as shown in the third row. In contrast, our system fails to correctly classify a true *children* relation as shown in the fourth row. However, our system achieves overall descent scores compared to the baseline. All the experimental results on RV task show that global information about the entities captured by the community-graph based features are significantly effective for RE task.

Table 4. Comparison of the confusion matrices resulted by BL and BL+DPED+Graph

Relation name	BL					BL + DPED + Graph				
	TP	FN	FP	TN	Acc.	TP	FN	FP	TN	Acc.
per:spouse	25	0	44	62	66.41	13	12	6	100	**86.28**
per:parents	69	25	116	270	70.62	63	31	75	311	**77.92**
per:children	28	9	17	613	96.10	28	9	2	628	**98.35**
per:country_of_death	68	4	1	188	98.08	68	4	1	188	**98.08**
per:country_of_birth	4	1	31	229	87.92	4	1	0	260	**99.62**
per:city_of_death	29	1	21	206	91.44	27	3	9	218	**95.33**
per:city_of_birth	139	0	4	86	98.25	138	1	0	90	**99.56**
org:top_members_employees	39	22	60	217	75.74	55	6	38	239	**86.98**
org:member_of	12	15	30	359	89.18	12	15	12	377	**93.51**
org:country_of_headquarters	36	104	33	329	72.71	41	99	13	349	**77.69**
org:city_of_headquarters	85	107	85	288	66.02	87	105	10	363	**79.65**
per:employee_or_member_of	84	203	321	1217	71.29	72	215	147	1391	**80.10**
All Together	618	491	763	4064	78.87	608	501	313	4514	**86.29**

Table 5. True positive (TP), true negative (TN) and false negative (FN) examples after validating relations

Claimed relation	Justification sentence	RV
spouse(Willem-Alexander, Maxima Zorreguieta Cerruti)	**Willem-Alexander** married **Maxima Zorreguieta Cerruti** from Argentina and they have three daughters: Princess Catharina-Amalia, Princess Alexia and Princess Ariane	TP
children(Margaret Thatcher, Mark)	In a statement to the public, **Thatcher**'s son **Mark** Thatcher said his twin sister Carol and the rest of their family had been overwhelmed by messages of support they had received from around the globe	TP
spouse(Willem-Alexander, Alexia)	**Willem-Alexander** married Maxima Zorreguieta Cerruti from Argentina and they have three daughters: Princess Catharina-Amalia, Princess **Alexia** and Princess Ariane	TN
children(Margaret Thatcher, Carol)	In a statement to the public, **Thatcher**'s son Mark Thatcher said his twin sister **Carol** and the rest of their family had been overwhelmed by messages of support they had received from around the globe	FN

6.2 Results of Knowledge Base Population Task

One objective of RE is the population of knowledge base. Since existing RE systems generate a large number of wrong relationships, it is interesting to know whether the validation step allows for building a better KB.

For evaluating KBP task, we define a *ground truth* (GT) for all the queries that contains different correct objects for each of the queries. An object is considered as correct if the excerpt containing the subject and object justifies their relation, otherwise wrong. A system should not repeat an answer (object) for the same query. If a system repeats an object for the same query only one instance

Table 6. KBP performances by some top ranked systems (upper part) and our RV models (lower part)

	Precision	Recall	F-score
System-1	36.73	22.78	28.12
System-2	32.07	24.89	28.03
System-3	37.50	21.52	27.35
Voting+Linguistic+Graph	**38.51**	**24.05**	**29.61**
Linguistic+Graph	29.53	18.57	22.80
Voting	24.88	21.10	22.83

of that object would be considered as correct and others would be wrong. Moreover, there are some queries such as *city_of_birth* whose object should be a single value. Therefore, a system has to response with a single object for such query. In our KBP system, we select an object randomly if several candidates are validated as correct for such relation. We compute the KBP performances of the single systems that participated to the TAC KBP evaluation on our test dataset for comparison. The test dataset given by the TAC KBP organizers provides the assessments of the slot filling responses of all the participating systems. Therefore it allows us to compute their results on the subset of queries of our test set. The top 3 TAC KBP systems on our test set individually obtained F-score of 28.12, 28.03 and 27.35 accordingly (see upper part of Table 6).

Since different relation extraction systems can be employed for the KBP task, we can use the *voting* feature to take advantage of the agreements on the outcomes by several relation extraction systems. Therefore, we built a RV model by using a single voting feature. Since the best performance of RV is achieved by Voting+Linguistic+Graph features, we use the RV model trained by this feature combination for the KBP task.

In the lower part of Table 6, we see that the voting based KBP system obtains an F-score of 22.83 which indicates the importance of this feature. Interestingly, our Voting+Linguistic+Graph based KBP system achieves a F-score of 29.61 which is higher than each individual KBP system. We also observe that Voting+Linguistic+Graph based KBP system achieves the highest precision of 38.51 that is almost 2 points higher than the best KBP system. The precision improvement indicates that our model discards many wrong relations which are resulted by different RE systems. Moreover, Voting+Linguistic+Graph based KBP system obtains the recall of 24.05 that is around 1.27 point higher than the best RE based KBP system and around 3 points higher than the voting based KBP system. These results justify that our system enables to fill more relations in knowledge base than the existing ones specially for the trigger dependent relations.

7 Conclusion

In this paper, we have presented community-based features for RE task that are able to capture some global information about the entities in a relationship. The proposed features are computed on a community graph extracted from a corpus and they measure how two entities are associated globally when they are in a relationship. Since such kind of measurements cannot characterize the semantics of relations, we combine these with some linguistic features that are able to characterize the type of a relation. We have shown that the proposed graph based features significantly improve the performance of relation extraction over the baseline. The proposed features also enable to globally select a large number of true values for populating a knowledge base and helps to obtain better scores over the sate-of-the-art system for the relations we studied.

One of our objectives is now to explore graph algorithm to exploit our graph representation for relation validation and KBP tasks in an unsupervised fashion.

References

1. GuoDong, Z., Jian, S., Jie, Z., Min, Z.: Exploring various knowledge in relation extraction. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 427–434. Association for Computational Linguistics (2005)
2. Bunescu, R.C., Mooney, R.J.: A shortest path dependency kernel for relation extraction. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 724–731. Association for Computational Linguistics (2005)
3. Fundel, K., Küffner, R., Zimmer, R.: Relex–relation extraction using dependency parse trees. Bioinformatics **23**(3), 365–371 (2007)
4. Nguyen, T.H., Grishman, R.: Relation extraction: Perspective from convolutional neural networks. In: VS@ HLT-NAACL, pp. 39–48 (2015)
5. Augenstein, I.: Web Relation Extraction with Distant Supervision. Ph.D. thesis, University of Sheffield (2016)
6. Han, X., Sun, L., Zhao, J.: Collective entity linking in web text: a graph-based method. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 765–774. ACM (2011)
7. Friedl, D.M.B., Heidemann, J., et al.: A critical review of centrality measures in social networks. Bus. Inf. Syst. Eng. **2**(6), 371–385 (2010)
8. Rahman, R., Grau, B., Rosset, S.: Community graph and linguistic analysis to validate relationships for knowledge base population. In: 4th Annual International Symposium on Information Management and Big Data (SIMBig) (2017)
9. Zhou, G., Zhang, M., Ji, D., Zhu, Q.: Tree kernel-based relation extraction with context-sensitive structured parse tree information. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (2007)
10. Jiang, J., Zhai, C.: A systematic exploration of the feature space for relation extraction. In: HLT-NAACL, pp. 113–120 (2007)
11. 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. Association for Computational Linguistics (2012)

12. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 541–550. Association for Computational Linguistics (2011)
13. Mooney, R.J., Bunescu, R.C.: Subsequence kernels for relation extraction. In: Advances in Neural Information Processing Systems, pp. 171–178 (2006)
14. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. 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, vol. 2, pp. 1003–1011. Association for Computational Linguistics (2009)
15. Chowdhury, M.F.M., Lavelli, A.: Combining tree structures, flat features and patterns for biomedical relation extraction. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 420–429. Association for Computational Linguistics (2012)
16. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 148–163. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15939-8_10
17. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 455–465. Association for Computational Linguistics (2012)
18. Vu, N.T., Adel, H., Gupta, P., Schütze, H.: Combining recurrent and convolutional neural networks for relation classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, pp. 534–539. Association for Computational Linguistics, June 2016
19. Zheng, S., Xu, J., Zhou, P., Bao, H., Qi, Z., Xu, B.: A neural network framework for relation extraction: learning entity semantic and relation pattern. Knowl. Based Syst. **114**, 12–23 (2016)
20. Dligach, D., Miller, T., Lin, C., Bethard, S., Savova, G.: Neural temporal relation extraction. In: EACL 2017, p. 746 (2017)
21. Gardner, M., Mitchell, T.M.: Efficient and expressive knowledge base completion using subgraph feature extraction. In: EMNLP, pp. 1488–1498 (2015)
22. Wang, Q., Liu, J., Luo, Y., Wang, B., Lin, C.: Knowledge base completion via coupled path ranking. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1308–1318 (2016)
23. Yu, D., Ji, H.: Unsupervised person slot filling based on graph mining. In: ACL (2016)
24. Guo, Y., Che, W., Liu, T., Li, S.: A graph-based method for entity linking. In: IJCNLP, Citeseer, pp. 1010–1018 (2011)
25. Holzinger, A., Ofner, B., Stocker, C., Calero Valdez, A., Schaar, A.K., Ziefle, M., Dehmer, M.: On graph entropy measures for knowledge discovery from publication network data. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 354–362. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40511-2_25
26. 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)
27. Bonacich, P., Lloyd, P.: Eigenvector-like measures of centrality for asymmetric relations. Soc. Netw. **23**(3), 191–201 (2001)

Predicting Invariant Nodes in Large Scale Semantic Knowledge Graphs

Damian Barsotti, Martin Ariel Dominguez, and Pablo Ariel Duboue(✉)

FaMAF-UNC, Cordoba, Argentina
{damian,mdoming,pablod}@famaf.unc.edu.ar

Abstract. Understanding and predicting how large scale knowledge graphs change over time has direct implications in software and hardware associated with their maintenance and storage. An important subproblem is predicting invariant nodes, that is, nodes within the graph will not have any edges deleted or changed (add-only nodes) or will not have any edges added or changed (del-only nodes). Predicting add-only nodes correctly has practical importance, as such nodes can then be cached or represented using a more efficient data structure. This paper presents a logistic regression approach using attribute-values as features that achieves 90%+ precision on DBpedia yearly changes trained using Apache Spark. The paper concludes by outlining how we plan to use these models for evaluating Natural Language Generation algorithms.

1 Introduction

We are interested in understanding and predicting how large scale knowledge graphs change over time. An important subproblem is predicting which nodes within the graph will not have any edges deleted or changed (what we call add-only nodes) or undergo any changes at all (what we call constant nodes) and which ones will not have any edges added or changed (del-only). Predicting add-only nodes correctly has practical importance, as such nodes can then be cached or represented using a more efficient data structure. In this paper we show a logistic regression approach using attribute-values as features that achieves 90%+ precision on DBpedia[1] yearly changes, as trained using Apache Spark. We conclude by outlining how we plan to use these models for Natural Language Generation. Add-only nodes can be efficiently represented as static information, for example by leveraging large scale perfect hashes [3].

Our intuition is that in large scale semantic graphs holding an imperfect representation of the real world, there are two types of changes, (1) **model enhancements,** where the truth about the world is better captured by the model (for example, the model missed any data about the children of an actor and that has been subsequently included) and (2) **model corrections,** where the world has changed and the model is updated (for example, an actress gave

[1] http://dbpedia.org.

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 48–60, 2018.
https://doi.org/10.1007/978-3-319-90596-9_4

birth to a newborn and that gets included in her page). Updates of the first type result in new information added to the graph, without modifying existing data. Finding such nodes is the objective of our work.

Our experiments employed eight different versions of DBpedia, a large scale knowledge graph with several million nodes. Our results show that for updates similar in size, models training in past data can help identify add-only nodes in future data. From our experiments, we believe better data in terms of stable extraction scripts and equally spaced regular intervals is needed rather than relying in the data dumps graciously provided by the DBpedia project.

This paper is structured as follows: in the next section we summarize related work. In Sect. 3 we discuss DBpedia, the semantic graph we used for our experiments, and how we build the datasets for our experiments. In Sect. 4 we describe the machine learning and other methods used for this work. The result of experiments are described in Sect. 5. We close with a discussion of our intended application in Natural Language Generation.

2 Related Work

Mining graphs for nodes with special properties is not new to Big Data mining [5]. With the development of DBpedia, much research has been devoted to exploiting this resource in AI tasks as well as to model its changes. For example, Lehman and others' work was on modeling DBpedia's currency [18], that is, the age of the data it contains and the speed at which those changes can be captured by any system. Although currency could be computed based on the modification/creation dates of the resources, this information is not always present in Wikipedia pages. To overcome this problem, the authors propose a model to estimate currency combining information from the original related pages and currency metrics measuring the speed of retrieval by a system and basic currency or timestamp. Their experiments suggest that entities with high system currency are associated with more complete DBpedia resources and entities with low system currency are associated with Wikipedia pages that are not easily tractable (or that "could not provide real world information" according with the authors). While we also look into changes in DBpedia, we are interested in changes that for the most part do not result from changes in the real world, as Lehman and others are interested. Using the nomenclature from the introduction, they focus on model corrections while we are also interested in model enhancements.

The need to account for changes in ontologies has long been acknowledged, given that they may not be useful in real world applications if the representation of the knowledge they contain is outdated. Eder and Koncilia [9] present a formalism to represent ontologies as graphs that contain a time model including time intervals and valid times for concepts. They base their formalism on techniques developed for temporal databases, namely the versioning of databases instead of their evolution and they provide some guidelines about its possible implementation. Our work can be used to improve the internal representation of such temporal databases [4].

Another source of ontology transformation is spatio-temporal changes. Dealing with spatial changes in historical data (or over time series) is crucial for some NLP tasks, such as information retrieval [10]. The authors deal with the evolution of the ontology's underlying domain instead of its versioning or evolution due to developments or refinements. Their main result is the definition of partial overlaps between concepts in a given time series, which was applied to build a Finnish Temporal Region Ontology, showing promising results.

Finally, we see parallelisms between tracking change in DBpedia and in other large graphs, namely, object graphs in garbage collection systems. State of the art garbage collection will single out objects that survive multiple garbage collections [19] and stop considering them for collection. We expect that it is this type of optimizations where the detection of invariable nodes will help semantic graphs updates.

3 Data

We investigate the task of prediction invariant nodes by using DBpedia, a knowledge graph derived from the Wikipedia collaborative encyclopedia started in January 2001 at present containing over 37 million articles in 284 languages. We selected DBpedia because it is a large scale naturally occurring knowledge graph with a rich update history.

Given that the content in Wikipedia pages contains structured information in their side boxes ("infoboxes"), it is possible to extract and organize it in an ontology-like manner as implemented in the DBpedia community project. This is accomplished by mapping Wikipedia infoboxes from each page to a curated shared ontology that contains 529 classes and around 2,300 different properties. These mappings are themselves organized as a wiki and updated over time. DBpedia contains the knowledge from 111 different language editions of Wikipedia and, for English the knowledge base consists of more than 400 million facts describing 3.7 million things [13]. A noble feature of this resource is that it is freely available to download in the form of *dumps* or it can be consulted using specific tools developed to query it.

These dumps contain the information in a language called Resource Description Framework (RDF) [12]. The WWW Consortium (W3C) has developed RDF to encode the knowledge present in web pages, so that it is comprehensible and exploitable by agents during any information search. RDF is based on the concept of making statements about (web) resources using expressions in the subject-predicate-object form. These expressions are known as triples, where the subject denotes the resource being described, the predicate denotes a characteristic of the subject and describes the relation between the subject and the object. A collection of such RDF declarations can be formally represented as a labeled directed multi-graph, naturally appropriate to represent ontologies.

Next, we formally define this concept.

Definition. Given a multigraph G_0 with named edges such that each source node S is linked through an edge labeled V to a target node O, which we will call a *triple* $\langle S, V, O \rangle$, we will say a given node S is an *add-only node* if in a next version (G_1) of the multigraph, all triples starting on S in G_0 are also in G_1. That is, S is $add - only$ iff :

$$\forall v, o / \langle S, v, o \rangle \in G_0 \Rightarrow \langle S, v, o \rangle \in G_1$$

Similarly, S is $del - only$ iff:

$$\forall v, o / \langle S, v, o \rangle \in G_1 \Rightarrow \langle S, v, o \rangle \in G_0$$

Table 1 shows the different years employed in this work. The DBpedia project obtains its data through a series of scripts run over Wikipedia, which on itself is a user-generated resource. Changes to the DBpedia scripts or to Wikipedia itself sometimes result in dramatic differences from one year to the next resulting in many nodes to disappear and new nodes to be created (Table 2), in addition to natural changes in Wikipedia. Besides the overall sizes, what is relevant to this work is the total number of additions and deletions, shown in Table 3.

Table 1. Data sizes with percentage increase from previous version.

Version	# Unique subjects	# Links
2010-3.6	1,638,799	13,608,535
2011-3.7	1,827,598 (111.52%)	17,228,764 (126.60%)
2012-3.8	2,343,007 (128.20%)	20,174,709 (117.09%)
2013-3.9	3,241,132 (138.33%)	25,346,359 (125.63%)
2014	4,172,476 (128.73%)	33,296,974 (131.36%)
2015-04	4,030,472 (96.59%)	32,018,293 (96.15%)
2015-10	4,903,501 (121.66%)	35,436,595 (110.67%)
2016-04	4,954,767 (101.04%)	35,085,378 (99.00%)

Table 2. Percentage of entities that disappear or appear anew in the next version.

Consecutive		Old subjects	New subjects
2010-3.6	2011-3.7	13.53%	22.47%
2011-3.7	2012-3.8	19.38%	37.12%
2012-3.8	2013-3.9	8.98%	34.20%
2013-3.9	2014	5.25%	26.39%
2014	2015-04	27.29%	24.72%
2015-04	2015-10	13.77%	29.12%
2015-10	2016-04	4.03%	5.02%

4 Methods

Our prediction system is implemented using Apache Spark[2] using the distributed Logistic Regression package in MLlib [15]. In our algorithm the feature vector itself is comprised of binary features indicating whether or not a given relation object holds for the subject in OLD; that is, we do not look at whether the $\langle V_i, O_i \rangle$ have changed, just their existence in OLD. The class is, given a node in subject position, S:

add-only: $\{(V_i, O_i)\}_{OLD} \subseteq \{(V_i, O_i)\}_{NEW}$
constant: $\{(V_i, O_i)\}_{OLD} = \{(V_i, O_i)\}_{NEW}$
del-only: $\{(V_i, O_i)\}_{OLD} \supseteq \{(V_i, O_i)\}_{NEW}$

The full feature vector has a dimension of $\|V\| \times \|O\|$, a potentially very large number given the millions of values in O (for DBpedia $\|V\|$ is in the order of a thousand and $\|O\|$ is larger than the number of unique subjects in Table 1 but of comparable size). We leverage Apache Spark Mlib pipelines to filter out this extremely large feature vector to the top 200,000 entries.

We encode the DBpedia mapping files into Spark native Parquet file format [14], totaling 8.6 Gb on disk, for all DBpedia versions used. To compute the class labels efficiently, we profit from Spark SQL join operations [2]: we join the two consecutive versions and keep all unique subjects present on the old version. We then find whether there are any triples in the new version that has been added or deleted. Depending on the existence of such triples, we obtain three binary labels and use them to train three different classifiers (constant, add-only and del-only). The "other" class in Table 3 are the nodes that are negative for all three categories. After training the logistic regression a threshold on its value that maximizes F1 is determined by cross validation on the train data.

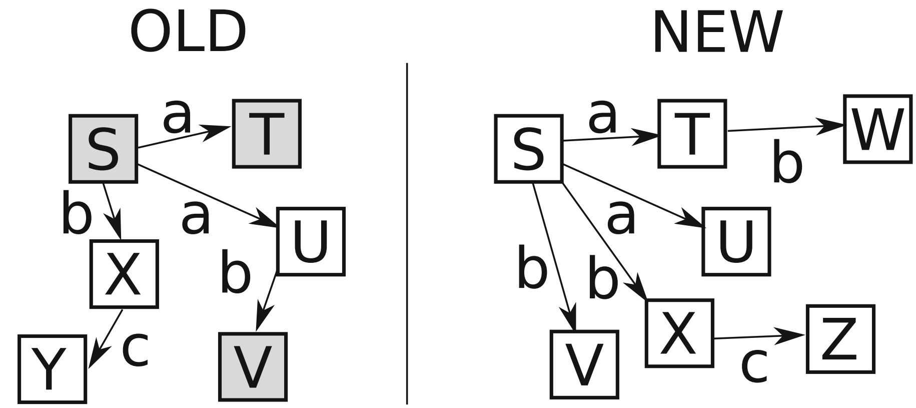
We also performed a drill-down for the errors for entities of different types. We used the types for evaluation but we did not use them for training as we did not want to bias the system unnecessarily but we revisit this decision in the conclusions.

Figure 1 shows a small example of feature and class extraction. A four node graph OLD evolves into a five node graph NEW. The classes for each node are computed over OLD, using three binary features.

5 Results

Using the machine learning method described in the previous section we took three consecutive year, $G_{y_1}, G_{y_2}, G_{y_3}$, built a model M on $G_{y_1} \rightarrow G_{y_2}$, apply M on G_{y_2}, obtaining G'_{y_3} and evaluate it by comparing it to G_{y_3}. Table 4 shows our results which are better understood by comparing them to Table 3 on the same page. Besides G_{y_3}, we also apply M to $G_{y_k} \rightarrow G_{y_{k+1}}$ for all $y_{k+1} > y_3$. All the experiments described here took a week to run using 20 Gb of RAM on a

[2] https://spark.apache.org/.



Node	Features	Target		
S	a=T, a=U	add-only	¬ constant	¬ del-only
T	∅	add-only	¬ constant	¬ del-only
U	b=V	¬ add-only	¬ constant	del-only
V	∅	add-only	constant	del-only
X	c=Y	¬ add-only	¬ constant	¬ del-only
Y	∅	¬ add-only	¬ constant	del-only

Fig. 1. Feature generation from an ontology *OLD* and *NEW*. In this example most nodes are add-only (shaded in *OLD*), only U loses a relation and it is thus not add-only. Note that W and Z are not considered as they do not exist in *OLD*.

8-core server. We can see that for pairs close in size (from Table 1) we obtain a precision close to 90% with recall ranging from 42% to 90% (and F-measure as high as 70%). *A priori,* we could expect that a model trained on data closer in time should predict better that one trained farther in time. The table, however, shows that models trained in periods with fewer changes perform better when applied to times with fewer changes and similarly for large changes. That is, in these experiments priors trump recency. Now, the uneven nature of the priors is an artifact of using DBpedia dumps as released by the project, which happen at irregular intervals and contain changes in Wikipedia concurrent with changes in the code used to obtain the ontology. We believe further experimentation on this problem should be done running a fixed version of the DBpedia scripts over Wikipedia historical data directly, a topic we discuss in the conclusions.

The low numbers in the bottom of the table can be attributed to ontological re-structuring on Wikipedia/DBpedia on 2015-04/10 period (second to last row on Table 3) were few entities remained constant through 2015.

From the table we can also see that detecting constant nodes, on the other hand, is a much more difficult task given the fact that Wikipedia is always growing and no model is ever 100% accurate. As such, model enhancements are always to be expected. Del-only results are much difficult to interpret because they contain an extra class of entities, which dominate the numbers (the “old subjects” column in Table 2, or the difference between del-only and proper del-only plus constant in Table 3): the entities that got removed from one version to another. While the numbers themselves are higher than add-only, we do not believe them to be comparable. The large number of deletions can be attributed to changes in the DBpedia scripts (from example, renaming “Chicago (City)” to “Chicago, Illinois”) that result in large scale entity deletion and creation.

Table 8 shows some examples of correctly and incorrectly predicted nodes.

5.1 Type Analysis

We also took all types present in at least six versions and combined all the predictions across all triples of consecutive years. We then analyze the behavior of our approach on the top most frequent 50 types, focusing on the ones above and below average (Table 5). From the 22 types below average (Table 6), 17 are related to people or locations. The table also highlights that the numbers vary substantially from year to year although certain years are clearly worse across the board: 2012–2013 look reasonably good while subsequent years experimented plenty of ontological changes due to the introduction of the OWL ontology [1] and the launch of the Wikidata project [21]. The analysis for relatively good

Table 3. Percentage of entities which remains constant or add-only calculated between two consecutive versions of DBpedia. Here, “proper” means add-only that are not constant and del-only that are neither constant nor disappeared in the next version.

Consecutive		Constant	Add-only		Del-only		Other
Versions				Proper		Proper	
2010-3.6	2011-3.7	12.14%	32.13%	19.98%	30.11%	4.43%	49.89%
		199,047	526,550	327,503	493,556	72,646	817,740
2011-3.7	2012-3.8	44.69%	49.10%	4.40%	65.91%	1.83%	29.67%
		816,830	897,386	80,556	1,204,727	33,558	542,315
2012-3.8	2013-3.9	57.04%	65.80%	8.75%	72.61%	6.59%	18.62%
		1,336,488	1,541,727	205,239	1,701,442	154,523	436,326
2013-3.9	2014	22.24%	32.09%	9.84%	33.11%	5.61%	57.04%
		721,117	1,040,162	319,045	1,073,183	181,899	1,848,904
2014	2015-04	13.77%	16.08%	2.31%	42.50%	1.43%	55.18%
		574,767	671,268	96,501	1,773,392	59,924	2,302,583
2015-04	2015-10	20.60%	22.73%	2.12%	36.65%	2.28%	61.21%
		830,463	916,305	85,842	1,477,551	92,034	2,467,079
2015-10	2016-04	81.84%	84.16%	2.31%	91.03%	5.16%	6.64%
		4,013,131	4,126,835	113,704	4,464,119	253,261	325,678

Table 4. Training in two consecutive years and evaluating on a third. Training maximizing F1.

Train		Eval	System								
Source	Target	Target	Constant			Add-only			Del-only		
			Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
2010-3.6	2011-3.7	2012-3.8	47.03	19.02	**27.09**	56.82	52.34	**54.49**	78.04	41.17	**53.91**
		2013-3.9	70.10	18.96	**29.85**	73.98	44.43	**55.52**	83.99	51.38	**63.76**
		2014	43.05	22.94	**29.93**	31.80	38.00	**34.63**	49.38	75.49	**59.71**
		2015-04	29.11	23.38	**25.93**	13.79	33.55	**19.55**	55.93	67.60	**61.22**
		2015-10	35.06	19.05	**24.69**	18.60	30.36	**23.07**	48.66	70.30	**57.51**
		2016-04	84.16	11.48	**20.20**	84.53	35.48	**49.98**	93.47	59.10	**72.41**
2011-3.7	2012-3.8	2013-3.9	68.30	66.69	**67.49**	73.06	66.48	**69.61**	81.16	83.30	**82.22**
		2014	25.13	73.11	**37.41**	34.78	74.64	**47.46**	37.88	89.05	**53.15**
		2015-04	13.67	72.51	**23.00**	15.60	74.65	**25.81**	44.98	87.58	**59.43**
		2015-10	21.84	79.63	**34.29**	22.66	78.69	**35.19**	38.92	89.40	**54.23**
		2016-04	84.44	78.84	**81.54**	85.61	81.26	**83.38**	92.49	87.36	**89.85**
2012-3.8	2013-3.9	2014	28.60	91.60	**43.59**	37.29	94.06	**53.41**	38.57	94.45	**54.77**
		2015-04	15.60	90.91	**26.64**	16.92	94.89	**28.72**	45.50	90.62	**60.59**
		2015-10	23.49	92.60	**37.47**	24.06	95.53	**38.44**	39.49	91.95	**55.25**
		2016-04	84.53	85.54	**85.03**	84.93	92.32	**88.47**	92.75	88.58	**90.62**
2013-3.9	2014	2015-04	39.34	82.54	**53.28**	29.66	90.93	**44.73**	75.81	60.85	**67.51**
		2015-10	51.20	85.04	**63.92**	36.41	90.59	**51.95**	59.05	75.68	**66.33**
		2016-04	90.26	40.54	**55.95**	85.63	57.02	**68.45**	94.52	53.62	**68.42**
2014	2015-04	2015-10	63.18	57.75	**60.34**	63.20	78.61	**70.07**	71.66	63.78	**67.49**
		2016-04	87.67	20.31	**32.98**	91.76	33.37	**48.94**	92.73	70.61	**80.17**
2015-04	2015-10	2016-04	90.22	39.27	**54.72**	91.86	42.06	**57.70**	94.12	41.50	**57.60**

types (Table 7) shows more stability per year. The types in both lists make conceptual sense: insects and bodies of water will usually have missing information added over time but data is seldom modified. On other hand, soccer players have statistics continually updated and they change teams fairly often.

6 Application to Natural Language Generation

In Natural Language Generation [17], Referring Expressions Generation (REG), is the task of, given an entity (the **referent**) and a set of competing entities (the **set of distractors**), creation of a mention to the referent so that, in the eyes of the reader, it is clearly distinguishable from any other entity in the set of distractors. Therefore REG algorithms are expected to select attributes that unambiguously identify an entity with respect to a set of distractors.

Table 5. First type analysis: for all types appearing in all years, pick the top most frequent 50 and see their overall precision/recall/F1 over all three consecutive years. Rows in italic are 20% above average in F1, rows in bold are 20% below average.

Type	Count	Constant			Add-only			Del-only		
		Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
AVERAGE	482,492	60.91	65.37	63.06	57.24	53.54	55.33	64.01	64.72	64.36
owl Thing	5,019,008	65.02	66.48	65.74	56.83	51.09	53.81	66.51	68.78	67.63
Person	3,412,161	**37.82**	**47.54**	**42.13**	**31.30**	**28.01**	**29.56**	**51.96**	**48.15**	**49.98**
Settlement	1,686,605	49.07	55.54	52.11	**40.99**	**31.44**	**35.59**	**46.83**	**35.48**	**40.37**
Place	1,682,897	53.76	80.45	64.45	46.04	51.83	48.77	49.18	57.33	52.95
Village	727,057	57.55	48.11	52.41	**43.80**	**25.18**	**31.98**	**52.13**	**30.08**	**38.15**
PersonFunction	648,343	*88.44*	*88.56*	*88.50*	*87.90*	*79.30*	*83.38*	*88.76*	*98.70*	*93.47*
SportsPerson	639,620	74.73	74.36	74.55	*72.58*	*72.25*	*72.41*	*91.85*	*96.26*	*94.00*
Athlete	608,903	50.60	57.10	53.65	**29.11**	**35.54**	**32.00**	**38.97**	**57.46**	**46.45**
Species	599,425	*75.80*	*75.56*	*75.68*	*73.25*	*74.93*	*74.08*	73.34	75.46	74.38
Eukaryote	579,287	75.34	75.17	75.25	*72.73*	*74.51*	*73.61*	72.76	75.02	73.87
Album	578,843	*79.85*	*75.83*	*77.79*	*75.15*	*72.84*	*73.98*	*77.59*	*88.28*	*82.59*
Organisation	547,605	51.14	60.83	55.56	46.78	52.02	49.26	52.50	65.41	58.25
Insect	514,096	*78.97*	*84.93*	*81.84*	*75.92*	*85.19*	*80.29*	*79.04*	*82.70*	*80.83*
SoccerPlayer	501,995	**31.63**	**19.51**	**24.13**	**10.22**	**2.37**	**3.85**	**25.63**	**11.56**	**15.94**
Film	439,197	*74.07*	*80.18*	*77.01*	61.83	65.88	63.79	68.29	77.36	72.54
Animal	435,377	*75.47*	*78.81*	*77.10*	*72.81*	*78.14*	*75.38*	72.69	79.47	75.93
MusicalWork	389,151	69.17	53.34	60.23	62.77	49.98	55.65	64.77	80.89	71.93
ArchitectStruc	318,087	54.22	79.84	64.58	45.08	55.16	49.61	49.18	65.08	56.03
Plant	272,468	76.63	72.96	74.75	*73.69*	*71.10*	*72.37*	75.35	68.92	71.99
Building	269,209	63.36	68.37	65.77	52.00	44.95	48.22	55.12	52.11	53.57
Company	252,438	73.66	77.13	75.36	*69.18*	*72.39*	*70.75*	73.88	79.03	76.37
MusicalArtist	234,646	**27.91**	**28.66**	**28.28**	**28.20**	**24.08**	**25.98**	**37.01**	**23.01**	**28.38**
Town	232,155	**44.94**	**32.61**	**37.80**	**34.67**	**17.86**	**23.58**	**43.37**	**23.75**	**30.69**
Artist	228,849	44.47	62.81	52.07	**33.96**	**46.72**	**39.33**	**42.72**	**55.96**	**48.45**
OfficeHolder	224,708	**31.13**	**25.94**	**28.29**	**16.50**	**12.18**	**14.01**	**22.51**	**13.22**	**16.65**
Single	216,311	61.07	53.82	57.22	54.06	42.39	47.52	53.20	57.94	55.47
Band	163,539	**30.43**	**25.33**	**27.65**	**24.89**	**16.62**	**19.93**	**36.54**	**29.68**	**32.76**
Book	160,711	**40.86**	**32.28**	**36.07**	**38.77**	**31.03**	**34.47**	**50.24**	**50.36**	**50.30**
Mollusca	155,990	75.24	71.29	73.21	*73.73*	*73.05*	*73.38*	*75.60*	*79.30*	*77.40*
School	143,682	**39.85**	**33.81**	**36.58**	**35.18**	**26.34**	**30.13**	**41.13**	**38.76**	**39.91**
TelevisionShow	142,129	63.72	64.11	63.92	55.28	50.22	52.63	64.36	69.32	66.75
MilitaryPerson	140,326	**22.63**	**14.99**	**18.04**	**15.11**	**7.31**	**9.85**	**30.50**	**20.88**	**24.78**
Ship	123,098	**46.53**	**30.97**	**37.19**	**30.95**	**9.40**	**14.42**	**27.37**	**28.81**	**28.07**
Station	121,190	**49.14**	**49.32**	**49.23**	**33.93**	**22.69**	**27.19**	**38.78**	**35.19**	**36.90**
River	120,146	74.50	70.20	72.29	*71.90*	*62.40*	*66.81*	*83.51*	*81.45*	*82.47*
Politician	119,184	**40.72**	**45.77**	**43.10**	**33.65**	**33.89**	**33.77**	**40.36**	**35.40**	**37.71**
AdminRegion	118,995	*74.76*	*86.82*	*80.34*	66.75	58.77	62.51	73.99	60.68	66.68
Writer	118,306	**45.34**	**53.65**	**49.15**	**39.74**	**33.31**	**36.24**	**47.71**	**33.35**	**39.26**
WrittenWork	116,619	**46.79**	**54.21**	**50.23**	43.00	48.75	45.69	53.27	71.48	61.04
EducationalInst	116,006	46.49	61.64	53.00	42.19	51.94	46.57	46.93	64.43	54.30
City	111,997	**32.94**	**21.21**	**25.80**	**23.74**	**10.80**	**14.84**	**36.70**	**16.97**	**23.21**

Table 6. Year drill-down for “bad” types (types 20% below average in F1) on the add-only category. The majority are related to either people (P) or location (L).

Type	2012-3.8	2013-3.9	2014	2015-04	2015-10	2016-04
(P) Artist	0.53	0.72	0.52	0.07	0.38	0.07
(P) Athlete	0.39	0.64	0.61	0.01	0.16	0.14
Band	0.66	0.49	0.00	0.36	0.27	0.01
(P) BaseballPlayer	0.00	0.63	0.00	0.17	0.17	0.01
Book	0.14	0.47	0.22	0.43	0.51	0.34
(L) City	0.46	0.45	0.16	0.55	0.16	0.10
(P) MilitaryPerson	0.13	0.31	0.40	0.07	0.22	0.04
(P) MusicalArtist	0.55	0.55	0.01	0.06	0.48	0.02
(P) OfficeHolder	0.36	0.36	0.47	0.07	0.05	0.05
(P) Person	0.46	0.66	0.62	0.13	0.04	0.10
(P) Politician	0.50	0.59	0.44	0.12	0.55	0.17
(L) Road	0.08	0.79	0.03	0.47	0.55	0.19
School	0.39	0.63	0.16	0.41	0.34	0.25
(P) Scientist	0.71	0.84	0.87	0.09	0.19	0.05
(L) Settlement	0.80	0.68	0.16	0.71	0.63	0.14
Ship	0.26	0.32	0.17	0.20	0.13	0.62
(P) SoccerPlayer	0.33	0.44	0.07	0.02	0.07	0.06
(L) Station	0.74	0.72	0.23	0.57	0.37	0.27
(L) Town	0.73	0.56	0.19	0.59	0.63	0.06
(L) Village	0.95	0.83	0.23	0.27	0.81	0.02
(P) Writer	0.62	0.83	0.78	0.12	0.30	0.08
WrittenWork	0.29	0.55	0.47	0.87	0.93	0.84

Our current work is part of a plan to simulate natural perturbations on the data in order to find the conditions on which REG algorithms start to fail (for example, a simulated DBpedia 25 years in the future).

In previous work we explored the robustness for the particular instances of REG algorithms by means of different versions of an ontology [8]. In [7] we presented experiments on two types of entities (people and organizations) and using different versions of DBpedia we found that robustness of the tuned algorithm and its parameters do coincide but more work is needed to learn these parameters from data in a generalizable fashion.

For that task, successfully predicting add-only nodes help us immediately with the performance of the prediction system. Our high precision results will then carry over to direct improvements in our full system: if our system has an error rate of 30% and there are 25% of add-only nodes, our current system will reduce error by up to 12% (in the case of 50% recall).

Table 7. Year drill-down for “good” types (types 20% above average in F1) on the add-only category.

Type	2012-3.8	2013-3.9	2014	2015-04	2015-10	2016-04
AdministrativeRegion	0.88	0.92	0.65	0.79	0.51	0.81
Album	0.40	0.28	0.86	0.79	0.86	0.91
Animal	0.59	0.92	0.69	0.89	0.90	0.93
AutomobileEngine	0.00	0.53	0.81	0.77	0.85	0.86
BodyOfWater	**0.75**	**0.82**	**0.72**	**0.75**	**0.75**	**0.84**
Company	0.32	0.38	0.81	0.79	0.81	0.87
Eukaryote	0.49	0.93	0.70	0.87	0.72	0.71
Film	0.65	0.59	0.81	0.72	0.80	0.84
Fish	0.59	0.93	0.68	0.80	0.87	0.72
Insect	**0.77**	**0.95**	**0.76**	**0.74**	**0.79**	**0.84**
Mollusca	0.04	0.86	0.71	0.76	0.77	0.81
PersonFunction	0.00	0.93	0.93	0.82	0.94	0.97
Plant	0.15	0.96	0.69	0.82	0.79	0.77
River	0.71	0.80	0.86	0.54	0.64	0.65
Species	0.49	0.93	0.70	0.88	0.88	0.91
SportsTeamMember		0.52	0.17	0.46	0.69	0.89

Table 8. Example predictions and mispredictions, using 2015-04 → 2015-10 as training and tested on 2016-04.

Correctly predicted add-only	
Iasi_Botanical_Garden	Constant
USS_Breakwater_(SP-681)	Constant
Interborough_Handicap	Constant
Thode_Island	Added archipelago→Marshall_Archipelago
Colonel_Reeves_Stakes	Added location→Perth Added location→Australia
Incorrectly predicted as add-only	
Beverly_Hills_Handicap	Disappears due to name change
First_Ward_Park	Disappears due to name change
2012_Shonan_Bellmare_season	Changes league→2012_J._League_Division_2 to league→2012_J.League_Division_2

We plan to extend the current model with a specific model for additions and deletions using techniques from statistical machine translation [11] and investigate techniques based on knowledge embedding models [22]. In general, mining knowledge from changes in the data has been a successful meta-heuristic in Natural Language Generation [6].

7 Conclusions

Working on the problem of predicting changes in large knowledge graphs, we have found promise on a logistic regression approach to detect invariant nodes, particularly nodes within the graph that will not have any edges deleted or changed (add-only nodes). Our experiments employed eight different versions of DBpedia, a large scale knowledge graph with several million nodes. Our results show that for updates similar in size, models training in past data can help identify add-only nodes in future data.

From our experiments, we believe better data in terms of stable extraction scripts and equally spaced regular intervals is needed rather than relying in the data dumps graciously provided by the DBpedia project. Moreover, our experiments suggest that enriching the training features with type information could be a valuable direction. Alternatively, for future experiments, it might suffice to focus on people and locations as both types are plentiful and capture fine grained issues that should allow differentiating better models of the data.

We believe our techniques can be of interest to anomaly detection for security purposes and spam detection [20] and for maintenance of the data. The add-only nodes and relations can be pre-cached using more efficient data structures such as perfect hashes [3]. For example, dividing storage and algorithms for entities of different freshness is a standard solution the recommendation systems world [16] (Chap. 5: "Taking recommenders to production").

Acknowledgments. The authors would like to thank the Secretaria de Ciencia y Tecnica of Cordoba Province for support and Annie Ying and the anonymous reviewers for helpful comments and suggestions. They would also like to extend their gratitude to the organizers of the SIMBig Symposium.

References

1. Antoniou, G., Van Harmelen, F.: Web ontology language: OWL. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 67–92. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24750-0_4
2. Armbrust, M., Xin, R.S., Lian, C., Huai, Y., Liu, D., Bradley, J.K., Meng, X., Kaftan, T., Franklin, M.J., Ghodsi, A., et al.: Spark SQL: relational data processing in spark. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1383–1394. ACM (2015)
3. Botelho, F.C., Ziviani, N.: External perfect hashing for very large key sets. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM 2007, pp. 653–662. ACM, New York (2007). http://doi.acm.org/10.1145/1321440.1321532
4. Cheng, S., Termehchy, A., Hristidis, V.: Efficient prediction of difficult keyword queries over databases. IEEE Trans. Knowl. Data Eng. **26**(6), 1507–1520 (2014)
5. Drury, B., Valverde-Rebaza, J.C., de Andrade Lopes, A.: Causation generalization through the identification of equivalent nodes in causal sparse graphs constructed from text using node similarity strategies. In: Proceedings of SIMBig, Peru (2015)

6. Duboue, P.A., McKeown, K.R.: Statistical acquisition of content selection rules for natural language generation. In: Proceedings of the 2003 Conference on Empirical Methods for Natural Language Processing, EMNLP 2003, Sapporo, Japan, July 2003
7. Duboue, P.A., Domínguez, M.A.: Using robustness to learn to order semantic properties in referring expression generation. In: Montes-y-Gómez, M., Escalante, H.J., Segura, A., Murillo, J.D. (eds.) IBERAMIA 2016. LNCS (LNAI), vol. 10022, pp. 163–174. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47955-2_14
8. Duboue, P.A., Domínguez, M.A., Estrella, P.: On the robustness of standalone referring expression generation algorithms using RDF data. In: WebNLG 2016, p. 17 (2016)
9. Eder, J., Koncilia, C.: Modelling changes in ontologies. In: Meersman, R., Tari, Z., Corsaro, A. (eds.) OTM 2004. LNCS, vol. 3292, pp. 662–673. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30470-8_77
10. Kauppinen, T., Hyvnen, E.: Modeling and reasoning about changes in ontology time series. In: Sharman, R., Kishore, R., Ramesh, R. (eds.) Ontologies. Integrated Series in Information Systems, pp. 319–338. Springer, Boston (2007). https://doi.org/10.1007/978-0-387-37022-4_11
11. Koehn, P.: Statistical Machine Translation, 1st edn. Cambridge University Press, New York (2010)
12. Lassila, O., Swick, R.R., Wide, W., Consortium, W.: Resource description framework (RDF) model and syntax specification (1998)
13. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web J. **6**(2), 167–195 (2015)
14. Li, X., Zhou, W.: Performance comparison of Hive, Impala and Spark SQL. In: 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 1, pp. 418–423. IEEE (2015)
15. Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., et al.: MLlib: machine learning in Apache Spark. J. Mach. Learn. Res. **17**(1), 1235–1241 (2016)
16. Owen, S.: Mahout in Action. Manning, Shelter Island (2012)
17. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press, Cambridge (2000)
18. Rula, A., Panziera, L., Palmonari, M., Maurino, A.: Capturing the currency of DBpedia descriptions and get insight into their validity. In: Proceedings of the 5th International Workshop on Consuming Linked Data (COLD 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, 20 October 2014 (2014)
19. Stefanović, D., McKinley, K.S., Moss, J.E.B.: Age-based garbage collection. ACM SIGPLAN Not. **34**(10), 370–381 (1999)
20. Tsai, C.F., Hsu, Y.F., Lin, C.Y., Lin, W.Y.: Intrusion detection by machine learning: a review. Expert Syst. Appl. **36**(10), 11994–12000 (2009)
21. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
22. Xie, Q., Ma, X., Dai, Z., Hovy, E.: An interpretable knowledge transfer model for knowledge base completion. In: ACL 2017: Proceedings of the Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Vancouver (2017)

Privacy-Aware Data Gathering for Urban Analytics

Miguel Nunez-del-Prado(✉), Bruno Esposito, Ana Luna,
and Juandiego Morzan

Universidad del Pacífico, Av. Salaverry 2020, Lima, Peru
{m.nunezdelpradoc,bn.espositoa,ae.lunaa,j.morzansamame}@up.edu.pe

Abstract. Nowadays, there are a mature set of tools and techniques for data analytics, which help Data Scientists to extract knowledge from raw heterogeneous data. Nonetheless, there is still a lack of spatiotemporal historical dataset allowing to study everyday life phenomena, such as vehicular congestion, press influence, the effect of politicians comments on stock exchange markets, the relation between food prices evolution and temperatures or rainfall, social structure resilience against extreme climate events, among others. Unfortunately, few datasets are combining from different sources of urban data to carry out studies of phenomena occurring in cities (*i.e.*, Urban Analytics). To solve this problem, we have implemented a Web crawler platform for gathering a different kind of available public datasets.

Keywords: Privacy · Data collection · Urban analytics · Open data

1 Introduction

Providing citizens with free access to raw data is one of the new global trends. These data, generated on a daily basis, could come from different sources such as governmental entities, NGOs, companies or Public Administration Entities, social networks, newspapers, congestion services, *etc.* Therefore, the data format must be a standard to make easier the access, use, generation of information and sharing. Thus, it is crucial that governments and private organizations, which have valuable data in their systems, servers, and databases make available these datasets for common benefits while taking into account citizen's privacy.

Unfortunately, Latin American countries, in particular, Peru, lacks historical data available to citizens. In May 2017, the government published Legislative Decree[1] to create the National Authority for Transparency and Access to Public Information; and, strengthen the Regime of Personal Data Protection. This first step would allow not only greater transparency of the state institutions, but

[1] Legislative Decree 1353: "Decreto legislativo que crea la autoridad nacional de transparencia y acceso a la información pública, fortalece el régimen de protección de datos personales y la regulación de la gestión de intereses".

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 61–75, 2018.
https://doi.org/10.1007/978-3-319-90596-9_5

also the possibility of the citizens of becoming a partner and author of solutions that could improve the life quality of our society. Peru is beginning to generate an open data culture and developing an open data portal at the national level[2]. Nonetheless, some phenomena need fine-grained data. For instance, in a research paper [11], the author highlights the need to collect segregated data of urban poor for inclusive urban planning. The accentuated scarcity of segregated data does not allow to make a comprehensive understanding of their vulnerabilities. Information about the characteristics and disadvantages of the urban poor population is essential for (i) inclusive planning and (ii) building sustainable cities. Undoubtedly, there are many benefits of having segregated data of urban poor in urban planning, not only for inclusive planning but also to understand the vulnerability, to know the contribution of urban poor in urban economy and to prioritize actions.

Our main contribution is presenting an alternative to gathering daily basis generated data (from public sources for storing, organizing and sharing) to perform Urban Big Data Analytics under an aware privacy framework in a developing country such as Peru. The information gathered is public but scattered around multiple websites and institutions. The data collected has followed a sanitization process, which assures the identity safety of citizens and brands located in Peru. This platform aims to provide an urban dataset for finding correlations and studying different phenomena in urban environments such as urban planning, online emergency detection, vulnerability, climate change, resilience and even poverty.

The present paper is organized as follows: Sect. 2 describes the related works, Sects. 3, 4 and 5 detail the framework architecture, the collected data and data statistics of some datasets, respectively. Then, Sect. 6 shows a simple application of Urban Analytics. We find the locations where crime news affect the sentiment of the people the most. Finally, Sect. 7 concludes our work.

2 Related Works

Open Data promotes innovation thoughts societal participation with the use of the data. Such datasets include measurement data from city-wide sensor networks on smart cities as well as from citizen sensors. In the current section, we present some efforts for data collection to tackle urban planning problems, to detect emergencies and to show city insights in real time.

Concerning data collection for urban planning, [8] propose a smart city data collection platform. This platform gathers information about floods, water usage, traffic, vehicular mobility traces, parking lots, pollution, social networks and weather from smart homes, smart parking, vehicular networking, water & weather and environmental pollution monitoring systems. The authors use the collected information for urban planning decision-making. Nevertheless, in [9], authors claim a need to give some context to this kind of measurements. Thus, they proposed the Human-Aware Sensor Network Ontology for Smart Cities

[2] Sistema Nacional de Información Ambiental: sinia.minam.gob.pe/.

(HASNetO-SC) to describe knowledge associated with data collection from city-wide sensor networks with an appropriate level of contextual metadata for data understanding. Therefore, they implemented the architecture for data collection in an urban metropolitan area in Brazil. Consequently, the platform opens the possibility that citizens, who have a little to no knowledge about the collection environment and the collected data, to access and process the information.

About emergencies detection, the work of [12] proposes a mechanism to gather information from *Weibo*[3] about urban emergency events. The platform discovers What, Where, When, Who, and Why of a given emergency event from Weibo comments. Thus, to complete these pieces of information, the platform relies on *Social Sensors* and *Crowdsourcing* layers. The former receives textual data from Weibo users. The latter extracts the basic elements of an emergency event (what, when, where, who, and why) to provide information for rescue services or decision making.

Regarding urban data gathering, which is an essential element of modern cities, a great challenge appears, such as data volume, velocity, data quality, privacy, and security, among others. In the paper [7], authors describe the development of a set of techniques that aim at effective and efficient urban data management in real settings in Dublin city. The solutions were integrated into a system that is currently used by the city. The system can detect multiple types of incidents, each one focusing on a different input source. Hence, the solutions can identify events by analyzing in real-time GPS trajectories, data coming from sensors installed injunctions, or textual information coming from social media. Authors developed analysis modules to forecast load so that they can manage efficiently the volume and velocity. Besides, they combine information to infer events from data anomalies. Noisy data and erroneous measurements are also dealt. Moreover, machine learning algorithms are used to identified relevant tweets avoiding in that way low-quality data. Another real-time data collection application can be found in London where the information can be viewed in a dashboard[4]. More details on this work can be found in the reference [5], where its main idea is to understand the city dynamics in a better way. The system gathers data from different third-party entities and open data platforms given as *CSV* file, *JSON* object or an *HTML*.

At last, there is also a multi-source dataset of urban life in Milano and Trentino [2]. The authors put together different data sets, such as spatial grid, social pulse, telecommunications, precipitations, weather, electricity, news, and social pulse. The scientists locate spatially all the data in a grid, which allows comparing datasets generated in various companies using different standards. The idea behind this dataset is to make a testbed for different solutions to urban problems like energy consumption, mobility planning, tourism and migrant flows, urban structures and interactions, event detection, urban well-being, etc.

In this work, we present, as far as we could investigate, the first Web crawler platform with a wide variety of open data exclusive to Peru. Also, it is incredibly

[3] Weibo website: tw.weibo.com.

[4] London city dashboard: citydashboard.org/london.

versatile because it allows the possibility of adding information when desired. The broad spectrum and diversity of data enable the exploitation of them with multiple approaches, and they are not limited to a single topic, which offers a significant advantage over previously presented works.

In the next section, we detail the framework proposed in the present effort.

3 Data Gathering Framework

In the present section, we describe the different components of the architecture of the Web crawler. Figure 1 depicts the different parts of the architecture, where each part is responsible for a given task as follows:

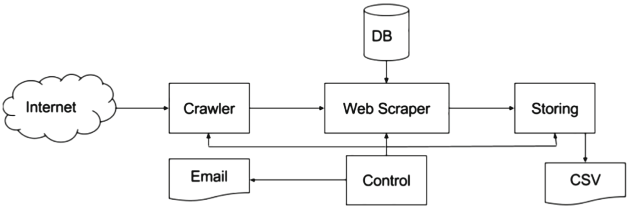


Fig. 1. Data gathering framework

Crawler: this artifact is responsible for reading the Uniform Resource Locators (*URLs*) from the database to download the target web pages from different websites.

Database: this database engine stores a list of *URL* provided by the user and the rules for reading and extracting relevant fields.

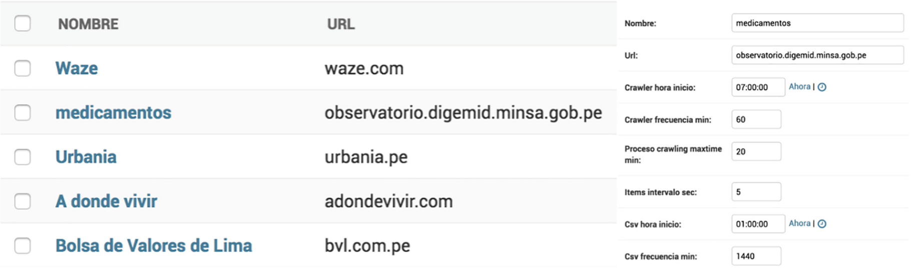
Web Scraper: it receives the downloaded web pages and the extracting rules to parse the web pages for gathering the needed fields of a given web page.

Storing: generate the Comma Separated Vector (*CSV*) files for each treated website.

Control: verifies the parts above (*i.e.*, Web Scraper, Crawler and Storing) are alive and are able to perform their task without problems.

The process begins with the manual creation of a list of target *URLs* and the rules needed to extract relevant fields from these. Then, that information is stored in the *Data Base*. The target *URLs* are chosen by the user. The data extraction starts with the *Crawler* reading the (*URLs*) from the *Data Base* to download target Web pages. Then, the *Web Scrapper* receives a set of Web pages as input. Consequently, it generates an index I of all the Web pages in each registered Website. Next, the *Web Scrapper* extract the data from the different gathered web pages, listed in the index I, using the particular structure settled for a given website. It is worth noting that websites structures for extracting data are stored in the database beforehand. For implementing this part of the crawler, we relied on the *Scrappy* library[5]. The Scrappy framework allows scheduling events for crawling [10]. Therefore, once the scrapy engine receives the downloaded website it sends to the spiders. Then, spiders extract particular information from the downloaded file.

Extracted data is temporally stored in the database to generate a Comma Separated Vector *CSV* file at the end of a day. Therefore, we store the datasets on a daily basis to build a historical repository. Finally, the last part of the framework is the *Control* mechanism that verifies the state of the crawler, scraper and file generation to notify by email if something goes wrong while gathering data from the different Websites. To setup the websites for data collecting we provide a friendly GUI to schedule the scrapping as shown in Fig. 2. Thus, the lefts part shows the list of websites to crawl and the right side illustrates the frequency to crawl, the beginning time, the url, etc.



□	NOMBRE	URL
□	Waze	waze.com
□	medicamentos	observatorio.digemid.minsa.gob.pe
□	Urbania	urbania.pe
□	A donde vivir	adondevivir.com
□	Bolsa de Valores de Lima	bvl.com.pe

Nombre:	medicamentos
Url:	observatorio.digemid.minsa.gob.pe
Crawler hora inicio:	07:00:00 Ahora | 🕒
Crawler frecuencia min:	60
Proceso crawling maxtime min:	20
Items intervalo sec:	5
Csv hora inicio:	01:00:00 Ahora | 🕒
Csv frecuencia min:	1440

Fig. 2. Crawler Graphical User Interface

In the next section, we detail the different datasets collected by the platform.

4 Datasets

In this section, we detail the variables of the collected datasets, which are available[6]. It is important to note that a sanitization process was performed on the datasets. Therefore, we consider as sensible information people's and brands' names, which are pseudonymized and erased, respectively. Concerning the opinion dataset, the comment field is sanitized by erasing stop words and sorting words alphabetically to reduce the impact of a De-anonymization re-link attack [4]. Consequently, these sanitization processes are carried out to prevent a privacy breach. Please note that the sanitization process is performed off-line. Thus, the sanitization does not limit the extraction of useful information.

It is possible to extract unstructured datasets from the websites targeted (i.e., news from newspapers, satellite imagery, etc.). In this work, we present two

[5] Scrapy: scrapy.org.

[6] BITMAP Urban Analytics: bitmap.com.pe/urbands.html.

non-tabular datasets: newspapers and social networks. Nonetheless, a tabular structure has been applied to them for easier readability.

In the following paragraphs, we described each of the nine categories of data sources as well as the datasets in each category.

Beauty consists of a description, date, price and category of cosmetics products (*c.f.* Table 1).

Table 1. Sample of beauty dataset.

category	date	price	title
mujer	2016-08-04	69.0	Noir de Nuit
mujer	2016-08-04	0.0	Orianité

Table 2. Sample of climate dataset.

CO	NO	NO2	NOX	O3
0.6	13.9	19.3	33.2	3.5
PM10	PM2.5	SO2	date	hour
51.6	34.0	2.5	2016-08-04	00:00

Climate category contains data from monitoring stations, atmospheric pollutants and radiation, which were provided by the National Meteorological and Hydro-graphic Service of Peru (SENAMHI).

Table 2 shows atmospheric pollutants (CO, NO, NO_2, NOX, O_3, PM_{10}, $PM_{2.5}$, SO_2) acquired from several monitoring networks distributed in Metropolitan Lima and the Province of Lima. Each measurement includes the time and date when it was executed.

Table 3 describes the set of meteorological data (humidity and temperature) of different monitoring stations. It also reports the station name where it was measured, the date and the UTM location of the record.

Table 3. Sample of station dataset.

altitude	district	date	hum
3928	AYAVIRI	16-08-04	9
lat	lon	temper.	type
14°52′22″	70°35′34″	19	Met

Table 4. Sample of UV radiation dataset.

arequipa	cajamarca	cusco	puno
8.0	7.0	9.0	9.0
date	ica	junin	tacna
2016-08-04	6.0	9.0	4.0
lima	moquegua	piura	
2.0	8.0	8.0	

Table 4 describes the data corresponding to solar radiation levels in different regions in Peru.

Markets category reports maximum and minimum prices and description of different products from three different suppliers in Lima. Table 5 contains an example of the products data. We report the name, category, minimum price, maximum price, average price and date of each good.

Table 5. Sample of a market dataset.

title	type	min_price
acelga	acelga	4.0
max_price	avg_price	date
5.0	4.5	2016-08-04

Table 6. Sample of real estate market dataset.

title	section	description
text1	alquiler	text2
location	area	price
Rep. Panama	320 m	$5000
lon	lat	date
−77.032584	−12.0431	2016-08-06

Medicines category contains prices of medicines in Lima. This information is provided by the Ministry of Health of Peru (*c.f.*, MINSA). Table 7 shows and example of the dataset, which contains the condition of the drug, address where it can be purchased, technical director of the drug, pharmacy name, price, commercial name, country, and date of manufacture.

Table 7. Sample of medicament dataset.

Condition	address	director
Con receta	C. Cordoba 2300	Luis Guia
manufacturer	date	Working hours
Hersil	2016-08-27	L-V 9:00-20:20
name	country	regulation
Bot. Pharmalys	Peru	NG1279
price	register	phone
2.5	NG1279	2660488
holder	location	
Hersil	Lince - Lima	

Newspapers category contains news from online newspapers in Peru. The dataset contains publication date, author, and the category on the newspaper, as shown in Table 8.

Table 8. Sample of written press dataset.

content	date	author
long text	2014-02-14 13:33:47	journalist name
section	location	title
mundo	mundo/eeuu	Edward Snowden

Table 9. Sample of opinion dataset.

user id	timestamp	language
1059254686	1476728629010	es
lon	lat	region
−77.0364	−12.0513	Lima, Pe
comment		
alza bar cafe centro el futuro gifs los puño		

Real Estate Market contains prices for houses, apartments, and offices sales nationwide. Table 6 shows an example of the real estate data. We capture information about the state of the property (sale or rental), description, address and geolocalization.

Social Networks category contains geo-referenced comments from people on different topics and social relations.

Table 9 shows opinions of various users, the language in which the opinion was made as well as the region of origin. It is worth noting that user id was pseudonymized using a hash function. In the same spirit, comments were sanitized to reduce re-identification risk.

Table 10 represents the friendship links between users of the social network.

Table 10. Sample of social links dataset.

user id1	user id2
1059254686	1059254367
1059254686	2259254876

Table 11. Sample of money exchange dataset.

currency	buy	sell	date
Swiss franc	3.346	3.686	2016-08-04
Euro	3.684	3.819	2016-08-04

Stock Market category contains two datasets for money exchange rates and stock exchange markets. The former contains the different historical exchange rates, from Soles to other foreign exchange as shown in Table 11.

The latter dataset also contains the transactions of the Stock Exchange of Lima. This dataset contains the price of the last transaction, opening price, purchase, sale, company, dates, amount, the amount of the stock, operations, and price variation. These variables are detailed in Table 12.

Table 12. Sample of stock exchange dataset.

pre	open	sale	company	date
7.8	7.7	7.56	Alicorp	2016-08-04
currency	amountNeg	mnemonic	noShare	noOper
S/	1 325 647	ALICORC1	173 677	16
sector	segm	last	var	sale
IND	RV1	7.56	−3.08	7.7

Transportation contains information about transport in Lima. There are reports of traffic jams and alerts on the main avenues of the city. Also, there is data about air transport at Jorge Chavez airport in Lima.

Tables 13 and 14 detail datasets of both alerts and congestion, respectively. These datasets contain information of some points in Lima city. It details the street, city, date, latitude and longitude coordinates of the alerts or the level of congestion. Table 13 also indicates the level of traffic, the node where the traffic level is reset as well as the speed of the traffic.

In the case of Table 14, the types and subtypes of alerts are gathered in addition to the above-mentioned data.

Table 15 shows the dataset for both arrivals and departures flights from Jorge Chavez airport in Lima, Peru. The name of the airline, the city of origin or destination, the status of the flight, the belt in which the suitcases are delivered. The estimated and scheduled time, and the door and flight number are also reported.

Table 13. Sample of jams dataset.

street	city	date
Av. Los Frutales	La Molina	2016-08-23
latitude	longitude	traffic level
−12.071628	−76.964632	2.0
node	speed	
Calatrava	4.719	

Table 14. Sample of alerts dataset.

Street	city	date
Av. Aviación	San Borja	2016-08-23
latitude	longitude	type
−12.086116	−77.003996	JAM
subtype		
JAM HEAVY TRAFFIC		

Table 15. Sample of airport traffic dataset.

Airline	city	state
Peruvian	Cusco	Landing
belt	date	estimated time
4	2016-07-22	10:50
scheduled time	door	flight
11:00	3	210

Table 16 shows the size of the data, the number of attributes and records, the characteristics linked to data types and temporal space granularity. The temporal granularity of these datasets ranges between 1.02 and 6.84 days. Additionally, there are seven datasets georeferenced with UTM coordinates (*i.e.*, latitude and longitude).

In the next section, we describe some statistics of our datasets.

5 Dataset Statistics

In this section, we detail the statistics of the different dataset in the described categories in Sect. 3.

Beauty. Table 17 shows the characteristics of the beauty dataset, described in Table 1. Two of the four attributes of the table contain categorical values (discrete values) for which the mode is the most important. There is no attrition in this dataset (*c.f.*, NAs).

Table 16. Summary of different data sets size.

Data set	Data Frame	Data points	Attributes	Size (Mb)	Temporal granularity	Geo referenced	Spatial granularity[a,b]	Types
Beauty	Beauty	22,190	4	1.2	1.02	False	None	float, str
Stock Market	Money ex	997	4	0.0	3.19	False	None	float, str
Stock Market	Stock ex	7,253	16	0.8	1.16	False	None	float, int, str
Weather	Ate	2,952	10	0.2	1.18	False	None	float, TSTP
Weather	Campo de marte	3,291	10	0.2	1.18	False	None	float, TSTP
Weather	Carabayllo	2,796	10	0.1	1.18	False	None	float, TSTP
Weather	Stations	291,839	11	19.3	1.16	True	UTM	float, int, str
Weather	Huachipa	2,322	10	0.1	1.18	False	None	float, TSTP
Weather	Puente piedra	2,968	10	0.2	1.18	False	None	float, TSTP
Weather	Radiacion UV	3,181	11	0.2	1.16	False	Region	int
Weather	San borja	3,180	10	0.2	1.18	False	None	float, TSTP
Weather	S.J. Lurigancho	2,565	10	0.1	1.18	False	None	float, TSTP
Weather	S.M. de Porres	2,890	10	0.2	1.18	False	None	float, TSTP
Weather	Sta. Anita	3,233	10	0.2	1.18	False	None	float, TSTP
Weather	V.M. del Triunfo	3,195	10	0.2	1.18	False	None	float, TSTP
Real estate	Real estate3	415,225	9	172.6	3.13	True	LL	float, str
Real estate	Real estate1	159,665	9	107.3	2.52	True	LL	float, str
Real estate	Real estate2	162,736	9	140.7	2.57	True	LL	float, int, str,
Medicines	Medicines	555,7353	14	1,311.6	1.17	False	District	float, int., str
Markets	Commerce1	136,433	6	12.3	1.2	False	None	float, str
Markets	Commerce2	490,582	5	37.1	1.2	False	None	float, str
Markets	Markets	11,178	6	0.7	3.09	False	None	float, str
Opinion	Opinion	6’979,829	7	239	-	True	LL	float, TSTP
News	News1	1’560,134	6	1,129.6	3.89	False	None	str
News	News2	13,392	6	23.4	1.18	False	None	str, int
News	News3	27,060	6	11.4	1.35	False	None	int, str
News	News4	40,451	6	16.1	6.84	False	None	int, str
Transport	Arrival	726,624	9	39.3	1.21	False	None	TSTP, str, int
Transport	Departure	679,260	9	41.2	1.17	False	None	TSTP, str, int
Transport	Alerts	351,690	7	33.3	1.16	True	LL	float, str
Transport	Jams	1’643,277	8	187.5	1.16	True	LL	float, int, str

[a] UTM is the Universal Transverse Mercator coordinate system.

[b] LL is the Latitude - Longitude coordinate system.

Table 17. Statistics of beauty dataset.

N	variable	type	mean	median	std	mode	min	max	NAs	%NAs
0	category	str	-	-	-	arrugas	-	-	0	0
1	date	date	-	-	-	-	-	-	0	0
2	price	float	935	45	35.1	0	0	1400	0	0
3	article	str	-	-	-	Essential cutis graso	-	-	0	0

Climate. We describe the characteristics associated with climate data. Two datasets will be described: (i) data from meteorological stations and their measurements (Table 18), and (ii) data on pollutants by districts (Table 19).

Table 18. Statistics of stations dataset.

N	variable	type	mean	median	std	mode	min	max	NAs	%NAs
0	altitude	integer	2143.8	2485.0	1560.8	3812.0	0.0	5192.0	0	0
1	department	str	-	-	-	-	-	-	291839	1
2	district	str	-	-	-	-	-	-	291839	1
3	station	str	-	-	-	CABO INGA	-	-	272508	1
4	date	date	-	-	-	-	-	-	0	0
5	humidity	float	62.6	62.0	226.4	75 0′44.52″	-	-	0	
6	latitude	str	-	-	-	12 46′17.86″	-	-	0	0
7	longitude	str	-	-	-	75 0′44.52″	-	-	0	0
8	province	str	-	-	-	-	-	-	291839	1
9	temperature	float	17.0	16.7	41.3	20.8	−30.8	4974.2	37258	0
10	type	str	-	-	-	Meteorologica	-	-	272508	1

Table 19. Statistics of pollutants dataset.

N	Var	type	mean	median	std	mode	min	max	NAs	%NAs
0	CO	float	1.2	1.2	7.6	1.4	0.0	410.9	64	0.02
1	NO	float	46.6	40.1	31.3	27.0	1.2	263.4	708	0.24
2	NO2	float	18.0	16.5	9.7	19.6	0.1	164.3	720	0.24
3	NOX	float	64.6	58.95	35.1	63.9	0.8	328.7	708	0.24
4	O3	float	7.6	5.4	9.3	0.5	0.3	198.3	9	0.0
5	PM10	float	115.7	107.25	55.9	94.5	0.0	948.0	10	0.0
6	PM2.5	float	35.1	28.9	27.4	0.0	0.0	203.0	518	0.18
7	SO2	float	11.2	9.6	10.4	8.7	2.7	353.3	290	0.1
8	date	date	-	-	-	-	-	-	0	0.0
9	horas	TSTP	-	-	-	-	-	-	1	0.0

Markets. This category contains information about four main markets in Peru. Table 20 shows statistics about the products supplied. The most popular product is *potato* and the most mentioned type is *red-headed onion*. The average price of the goods in this dataset is $2.746 PEN$.

Newspapers. Table 21 shows that the most cited author, section, and title are *Carlos Battle*, *executive zone*, *world/Current* and *5 tips for a start-up to survive*, respectively. Unfortunately, there are high levels of attrition reported in this dataset: The percentage of missing values of content, author, and location are 81%, 82% and 95%.

Real Estate. This category contains information about the real state market in Peru. We have collected information from the two biggest housing websites.

Table 20. Statistics of market dataset.

N	Var	type	mean	median	std	mode	min	max	NAs	%NAs
0	title	str	-	-	-	PAPA	-	-	0	0.0
1	type	str	-	-	-	CEBOLLA CABEZA ROJA	-	-	0	0.0
2	min_price	float	2.4	1.5	2.3	2.0	0.57	13.0	0	0.0
3	max_price	float	3.0	2.0	2.8	2.0	0.71	14.0	0	0.0
4	av._price	float	2.7	1.75	2.5	1.25	0.61	13.5	0	0.0
5	fecha	date	-	-	-	-	-	-	0	0.0

Table 21. Statistics of newspapers dataset.

N	variable	type	mean	median	std	mode	min	max	NAs	%NAs
0	content	str	-	-	-	text1	-	-	1264992	0.81
1	date	date	-	-	-	-	-	-	0	0
2	author	str	-	-	-	Carlos Batalla	-	-	1275035	0.82
3	section	str	-	-	-	zona-ejecutiva	-	-	0	0
4	location	str	-	-	-	mun./act	-	-	1480648	0.95
5	title	str	-	-	-	5 consejos para que una startup sobreviva	-	-	0	0

In Table 22 we have only two attributes with numerical values associated with the location of the property. The most offered properties are those with 100 m^2. Regarding the price, the most frequent value is $900.

Transportation. We describe the statistics of the datasets related to transportation in Peru. For car transport, Tables 23 and 24 report the statistics of alerts and congestion, respectively.

Table 23 shows that street with most alerts is *Av. Javier Prado* and *San Isidro* is the most reported district. Table 24 shows that the street most con-

Table 22. Statistics of real estate dataset.

N	variable	type	mean	median	std	mode	min	max	NAs	%NAs
0	title	str	-	-	-	Alquiler de	-	-	0	0.0
1	section	str	-	-	-	alquiler	-	-	0	0.0
2	description	str	-	-	-	Rento	-	-	2524	0.02
3	location	str	-	-	-	Ubicación	-	-	27912	0.17
4	area	str	-	-	-	100 m	-	-	2790	0.02
5	price	str	-	-	-	US$ 900	-	-	0	0.0
6	longitude	float	-77.01	-77.03	0.07	-77.03	-77.76	-76.13	0	0.0
7	latitude	float	-77.01	-77.03	0.07	-77.03	-77.76	-76.13	0	0.0
8	date	date	-	-	-	-	-	-	0	0.0

Table 23. Statistics of alerts dataset.

N	variable	type	mean	median	std	mode	min	max	NAs	%NAs
0	rue	str	-	-	-	Av. Javier Prado Este	-	-	0	0.0
1	city	str	-	-	-	San Isidro	-	-	0	0.0
2	date	date	-	-	-	-	-	-	0	0.0
3	latitude	float	−12.08	−12.09	0.008	−77.00	−77.07	−76.949	0	0.0
4	longitude	float	−77.01	−77.01	0.03	−77.00	−77.07	−76.949	0	0.06
5	subtype	str	-	-	-	JAM HEAVY TRAFFIC	-	-	28659	0.08
6	type	str	-	-	-	JAM	-	-	0	0.0

Table 24. Statistics of jams dataset.

N	variable	type	mean	median	std	mode	min	max	NAs	%NAs
0	calle	str	-	-	-	Street 1	-	-	0	0.0
1	ciudad	str	-	-	-	La Molina	-	-	0	0.0
2	fecha	date	-	-	-	-	-	-	0	0.0
3	latitud	float	−12.08	−12.08	0.00	−12.076	−12.11	−12.06	0	0.0
4	longitud	float	−76.99	−76.99	0.03	−76.96	−77.08	−76.94	0	0.0
5	nivel de trafico	integer	3.286	3.0	0.713	3.0	1.0	5.0	2	0.0
6	nodo	str	-	-	-	Street 1	-	-	0	0.0
7	velocidad	float	2.5	2.3	1.3	2.0	0.1	12.2	14288	0.01

gestioned in Lima is *Av. Circunvalación del Golf Los Incas* (street 1) in the *la Molina* district. The average speed in a jam is 2.59 km/h.

In the next section, we propose an example of Urban Analytics to show the potential of our datasets.

6 Sentiment Analysis of Crimes in Peru

One possible application of the datasets is sentiment analysis of crimes. For this purpose, we took a subset of 12 767 news items from four different newspapers in Peru. The sentiments analysis from the articles are divided into three different classes, namely positive, neutral, and negative. As seen in Fig. 3, most of the news in Peru generate a neutral opinion on the population.

In Fig. 4, we show the result of the sentiment analysis using the Indico API[7] based on multinomial logistic regression on ngrams with tf-idf features[8]. This algorithm classifies the news related to crime in positive, neutral and negative. Therefore, we have detected the spatial entities in texts for geo-referencing the

[7] Indico API: https://indico.io/.

[8] Source: https://indico.io/blog/sentimenthq-new-accuracy-standard/.

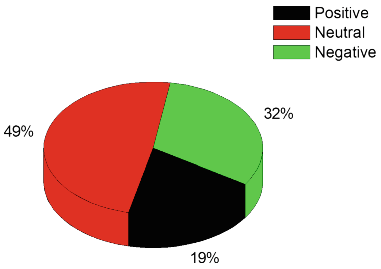
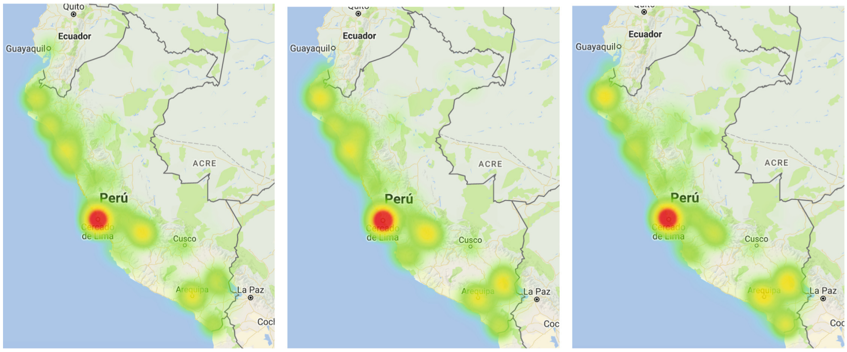


Fig. 3. Percentages of positive, neutral and negative sentiments.



(a) Heat map of positive sentiments.

(b) Heat map of neutral sentiments.

(c) Heat map of negative sentiments.

Fig. 4. Heat map sentiment analysis in Peru.

news' sentiments. We have plotted the heat map of each class over a cartography. The south of Peru often has a negative sentiment about crime news.

A simple analysis like the one presented in this work could reveal useful insights. With these data, it is possible to create heat maps like the ones shown in Fig. 4 and where they can be visualized and identified specific areas that could be of interest to a certain audience. This is just one of many Urban Analytics applications we are able to do using these urban datasets described in Sect. 4.

7 Conclusion

In the present work, we have described the architecture of a Web crawler platform to gather information about nine different categories of datasets to make urban analytics. The main contribution of this work is the provision of information to the scientific community and policymakers for analyzing and studying social behavior and urban phenomena in a developing country such as Peru. We have collected data following a privacy-aware structure. Sensible information about the citizens or brand-names has been sanitized. These datasets have enabled us to implement a Knowledge Tier Platform for Graph Mining [6] and perform urban analytics [3] or study urban resilience [1]. We have shown an example of a simple application. Undeniably, there is still more information to exploit. In the future, we plan to extend the crawler to collect more information from new public available Web sites. We also plan to make a privacy risk analysis of the described datasets.

References

1. Abbar, S., Zanouda, T., Borge-Holthoefer, J.: Robustness and resilience of cities around the world. arXiv preprint arXiv:1608.01709 (2016)
2. Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., Lepri, B.: A multi-source dataset of urban life in the city of milan and the province of trentino. Sci. Data **2**, 150055 (2015)
3. Di Clemente, R., Luengo-Oroz, M., Travizano, M., Vaitla, B., Gonzalez, M.C.: Sequence of purchases in credit card data reveal life styles in urban populations. arXiv preprint arXiv:1703.00409 (2017)
4. Gambs, S., Killijian, M.O., del Prado Cortez, M.N.: De-anonymization attack on geolocated data. J. Comput. Syst. Sci. **80**(8), 1597–1614 (2014)
5. Gray, S., O'Brien, O., Hügel, S.: Collecting and visualizing real-time urban data through city dashboards. Built Environ. **42**(3), 498–509 (2016)
6. Nunez-del Prado, M., Bravo, E., Sierra, M., Canchay, M., Hoyos, I.: Knowledge tier platform for graph mining in (smart) cities. In: Proceedings of Symposium on Information Management and Big Data (2016)
7. Panagiotou, N., et al.: Intelligent urban data monitoring for smart cities. In: Berendt, B., Bringmann, B., Fromont, É., Garriga, G., Miettinen, P., Tatti, N., Tresp, V. (eds.) ECML PKDD 2016. LNCS (LNAI), vol. 9853, pp. 177–192. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46131-1_23
8. Rathore, M.M., Ahmad, A., Paul, A., Rho, S.: Urban planning and building smart cities based on the internet of things using big data analytics. Comput. Netw. **101**, 63–80 (2016)
9. Santos, H., Furtado, V., Pinheiro, P., McGuinness, D.L.: Contextual data collection for smart cities. arXiv preprint arXiv:1704.01802 (2017)
10. Scrapy: Scrapy API. https://doc.scrapy.org/en/latest/topics/architecture.html
11. Srivastava, A.K.: Segregated data of urban poor for inclusive urban planning in India: needs and challenges. SAGE Open **7**(1), 2158244016689377 (2017)
12. Xu, Z., Liu, Y., Yen, N., Mei, L., Luo, X., Wei, X., Hu, C.: Crowdsourcing based description of urban emergency events using social media big data. IEEE Trans. Cloud Comput. **99**(PP), 1–10 (2016)

Purely Synthetic and Domain Independent Consistency-Guaranteed Populations in $\mathcal{SHIQ}^{(\mathcal{D})}$

Jean-Rémi Bourguet[1,2](✉)

[1] Dipartimento di Scienze Politiche ed Ingegneria dell'Informazione, Università degli Studi di Sassari (UNISS), Sassari, Italy
[2] Núcleo de Estudos em Modelagem Conceitual e Ontologias, Federal University of Espírito Santo (UFES), Vitória, Brazil
jean-remi.bourguet@ufes.br

Abstract. The elaborations of artificial knowledge bases can represent a clever solution to test new semantics-based infrastructures before deploying them and a precious support to the design of some prototypes. One major challenge of such synthetic data generations is to guarantee the acquisition of sound knowledge bases able to pass the equivalent of a Turing test. That's why populations have to be restricted to guarantee the consistency until a certain fragment of expressivity. In a past work, we released a first version of a populator guaranteeing the consistency and populating knowledge bases founded on TBoxes expressed in $\mathcal{ALCQ}^{(\mathcal{D})}$. This purely syntactic and domain independent populator is based on a random process of concept, role and limited data instantiations. In this paper, we propose to extend the expressivity covering by the populator until the fragment $\mathcal{SHIQ}^{(\mathcal{D})}$. This extension deals with RBoxes conforming the consistency of the role assertions with respect to the domains/ranges, the universal quantifications and the maximal cardinalities of all the super and inverse roles. Finally, an evaluation of some performances of the populator has been performed.

1 Introduction

Populating artificial knowledge bases is a well-known sideline to test the performances of diverse reasoning tasks [1]. For a long while, the solutions considered by default were semi-automated nay manual populations which had the disadvantages of being time-consuming and prone to inconsistencies. Moreover, quite all the solutions were dependent on the content of a domain conceptualization accessible through an explicit and formal description in terms of a concrete artifact, which is classified as a domain ontology by Guizzardi in [2]. According to the international academic publisher IGI-global[1], the ontology population is defined as the process of creating instances for an ontology *usually* involving the linking of various data sources to the elements of the ontology. However, this domain has often been tarnished like the task of recognizing the new elements that should go

[1] https://www.igi-global.com/dictionary/ontology-population/21134.

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 76–89, 2018.
https://doi.org/10.1007/978-3-319-90596-9_6

into a domain ontology [3]. In fact, an Ontology Populator can operate either as an Extracted Data Generator (EDG), or as a Synthetic Data Generator (SDG). In turn, the SDGs can perform as i-domain dependent SDGs i.e. those performing only populations depending on a given domain ontology or as ii-domain independent SDGs i.e. those performing populations independent of domains of ontologies. According to Guarino and Shneider [4], the complexity also called ontological depth (or precision) can range ontologies categorizing them as catalog, glossary, taxonomy, thesaurus, database scheme or axiomatic theories. As pointed out by Baader et al. in [5] the expressive power of an ontology should be adequate for defining the relevant concepts in enough detail, but not too expressive to make reasoning infeasible. Moreover, they argued that Description Logics (DL), a family of knowledge representation languages, should be ideal candidates for ontology languages because they provide both well-defined semantics and powerful reasoning tools. Early in the mid-1990s, DL were characterized in [6] by four fundamental aspects: the set of constructs used in concept and role expressions, the kind of assertions allowed in the TBox (assertions on concepts) and the Abox (assertions on individuals), and the inference mechanisms for reasoning on both TBox and Abox. In a past work [7], we proposed a tool that is Just another Populator of TBox (`JPoT`) based on a random process of assertions on individuals i.e. concept, role and data instantiations. `JPoT` performed populations like a domain independent SDG and guaranteeing consistency of the knowledge bases (that is in fact defined as the union of the TBox and the Abox) in a DL fragment called $\mathcal{ALCQ}^{(\mathcal{D})}$. This DL fragment is based on an extension of the well known DL $\mathcal{ALC}$ [8] with qualifying number restrictions ($\mathcal{Q}$) and datatypes ($^{(\mathcal{D})}$). In this paper we propose an extension of our populator to guarantee the consistency until the fragment $\mathcal{SHIQ}^{(\mathcal{D})}$. This DL fragment is in fact an extension of the DL fragment $\mathcal{ALCQ}^{(\mathcal{D})}$ including transitively closed primitive roles [9], inverse roles ($\mathcal{I}$) and role hierarchies ($\mathcal{H}$). The subsequent parts are as follows: Sect. 2 presents `JPoT`, Sect. 3 performs an evaluation, Sect. 4 presents the related works and Sect. 5 draws some perspectives.

2 Synthetic Data Generation

Formally, members of DL can be based on some finite sets like for example: a set $\mathbf{C_A}$ of concepts names, a set $\mathbf{D_T}$ of datatype names, a set $\mathbf{R_A}$ of abstract role names and a set $\mathbf{R_T}$ of concrete role names. Baader and Nutt introduced the notion of interpretation in first-order logic [10].

Definition 1. *An interpretation $\mathcal{I}$ is a tuple such that $\mathcal{I} = \langle \Delta^{\mathcal{I}}, \Delta_{\mathcal{D}}, \cdot^{\mathcal{I}} \rangle$ where:*

- *$\Delta^{\mathcal{I}}$ is the domain, i.e. a set of individuals,*
- *$\Delta_{\mathcal{D}}$ is a data-type domain disjoint with $\Delta^{\mathcal{I}}$,*
- *$\cdot^{\mathcal{I}}$ is the interpretation function which maps:*
  - *each $A \in \mathbf{C_A}$ to a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$,*
  - *each $r \in \mathbf{R_A}$ to a relation $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$,*
  - *each $D \in \mathbf{D_T}$ to a value space $D^{\mathcal{I}} \subseteq \Delta_{\mathcal{D}}$,*
  - *each $t \in \mathbf{R_T}$ to a relation $t^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{D}}$.*

The previous definition is based on the interpretation presented in the OWL2 direct semantic[2]. The two following definitions are adapted from [11,12].

Definition 2 (Rbox). *Let $\mathbf{R_A}$ be a set of role names and a set $\mathbf{R_+}$ of transitive roles s.t. $\mathbf{R_+} \subseteq \mathbf{R_A}$. The set of all $\mathcal{SHIQ}^{(\mathcal{D})}$-roles is $\mathbf{R_A} \cup \{r^- | r \in \mathbf{R_A}\} \cup \mathbf{R_T}$, where r^- is called the inverse of the role r. A role inclusion axiom (RIA) is of the form $r \sqsubseteq s$, where r, s are $\mathcal{SHIQ}^{(\mathcal{D})}$-roles. A role hierarchy is a finite set of RIAs. The interpretation $\mathcal{I}$ is extended for an* RBOX *to satisfy a role hierarchy $\mathcal{R}$ iff $r^\mathcal{I} \subseteq s^\mathcal{I}$ for each $r \sqsubseteq s \in \mathcal{R}$ and such that $\forall p \in \mathbf{R_A}$ and $\forall r \in \mathbf{R_+}$:*

$$(x, y) \in p^\mathcal{I} \; iff \, (y, x) \in (p^-)^\mathcal{I},$$
$$if \, (x, y) \in r^\mathcal{I} \; and \, (y, z) \in r^\mathcal{I} \; then \, (x, z) \in r^\mathcal{I}.$$

The $\mathcal{SHIQ}^{(\mathcal{D})}$-roles are provided with the following functions and relations:

- $\texttt{Inv}(r) := \begin{cases} r^- \text{ if } r \text{ is a role name} \\ s \quad \text{ if } r = s^- \text{ for a role name } s \end{cases}$
- $\sqsubseteq^+_\mathcal{R}$ is the transitive closure of $\mathcal{R} \cup \{\texttt{Inv}(r) \sqsubseteq \texttt{Inv}(s) | r \sqsubseteq s \in \mathcal{R} \wedge r, s \in \mathbf{R_A}\}$
- $\texttt{Tra}(s) := \begin{cases} \text{TRUE if } r \in \mathbf{R_+} \text{ or } \texttt{Inv}(r) \in \mathbf{R_+} \text{ for some } r \text{ with } r \equiv^+_\mathcal{R} s \\ \text{FALSE otherwise} \end{cases}$
- A role $r \in \mathbf{R_S}$ ($\mathbf{R_{NS}}$ if no) w.r.t. $\mathcal{R}$ iff $\texttt{Tra}(s) = \text{FALSE}$ for all $s \sqsubseteq^+_\mathcal{R} r$

JPoT can actually populate guaranteeing consistency the fragment $\mathcal{SHIQ}^{(\mathcal{D})}$.

Definition 3 (Syntax and semantics of $\mathcal{SHIQ}^{(\mathcal{D})}$).

$\mathbf{R_A}$ ::= $\mathbf{R_S} \mid \mathbf{R_{NS}}$
$\mathbf{R}$ ::= $\top_\mathbf{R} \mid \mathbf{R_A} \mid \mathbf{R_T}$
$\mathbf{D}$ ::= $\mathbf{C} \mid \mathbf{D_T}$
$\mathbf{C}$ ::= $\mathbf{C_A} \mid \neg\mathbf{C} \mid \top_\mathbf{C} \mid \bot_\mathbf{C} \mid (\mathbf{C} \sqcap \mathbf{C}) \mid (\mathbf{C} \sqcup \mathbf{C}) \mid \exists\mathbf{R}.\mathbf{D} \mid \forall\mathbf{R}.\mathbf{D} \mid \leq n\mathbf{R_S}.\mathbf{D} \mid$
$\geq n\mathbf{R_S}.\mathbf{D} \mid = n\mathbf{R_S}.\mathbf{D} \mid \leq n\mathbf{R_T}.\mathbf{D_T} \mid \geq n\mathbf{R_T}.\mathbf{D_T} \mid = n\mathbf{R_T}.\mathbf{D_T}$

Definition 4 (Unequivocal TBox [10]). *Given $C \in \mathbf{C_A}$, $D \in \mathbf{C}$, an unequivocal* TBOX *$\mathcal{T}$ is a finite set of i-concept equivalences '$C \equiv D$' or ii-concept inclusions '$C \sqsubseteq D$'. $\mathcal{I}$ is a model of $\mathcal{T}$ iff for all the axioms $\varphi \in \mathcal{T}$, $\mathcal{I} \Vdash \varphi$, with:*
- *$\mathcal{I} \Vdash (C \sqsubseteq D)$ iff $C^\mathcal{I} \subseteq D^\mathcal{I}$*
- *$\mathcal{I} \Vdash (C \equiv D)$ iff $C^\mathcal{I} = D^\mathcal{I}$*

Definition 5 (Abox). *Given a, $b \in \Delta^\mathcal{I}$, $d \in \Delta_\mathcal{D}$, $C \in \mathbf{C}$, $r \in \mathbf{R_A}$, $t \in \mathbf{R_T}$, an* ABOX *$\mathcal{A}$ is a finite set of class, role or data assertions. An interpretation $\mathcal{I}$ is a model of an $\mathcal{A}$ iff for all the assertions $\varphi \in \mathcal{A}$, $\mathcal{I} \Vdash \varphi$, with:*
- *$\mathcal{I} \Vdash C(a)$ iff $a \in C^\mathcal{I}$* Concept Assertion
- *$\mathcal{I} \Vdash r(a, b)$ iff $(a, b) \in r^\mathcal{I}$* Role Assertion
- *$\mathcal{I} \Vdash t(a, d)$ iff $(a, d) \in t^\mathcal{I}$* Data Assertion

[2] https://www.w3.org/TR/owl2-direct-semantics/.

JPoT is designed to populate TBox in function of:

Definition 6 (Parameters of population). *Given the set of all the concept, role and data assertions are respectively denoted A_c, A_r and A_d:*

$\boldsymbol{n} = |\Delta^{\mathcal{I}}|$ - *number of potential individuals*
$\boldsymbol{m} = |A_c \cup A_r \cup A_d|$ - *number of assertions*
$\boldsymbol{\tau} = \frac{|A_c|}{|A_c \cup A_r \cup A_d|}$ - *ratio of the number of concept assertions on that of assertions*
$\boldsymbol{\rho} = \frac{|A_r|}{|A_r \cup A_d|}$ - *ratio of the number of role assertions on that of data assertions.*

2.1 Concept Assertions

The Algorithm 1 describes the concept assertions phase of JPoT initiating with the computation of the set of all the disjoint class denoted $\mho_T$ such that $\mho_T = \{\{C, D\} | C \sqsubseteq^+ \neg D\}$ with $\sqsubseteq^+$ is the transitive closure of $\sqsubseteq$. Next, each individual drawn will instantiate the first concept drawn (with the function Draw) that is not disjointed with the concepts the individual already instantiates.

Algorithm 1. Concept Assertions (CAs)

```
Data: TBox, $m$, $\tau$
Result: round$(m \cdot \tau)$ CAs
x= 0;
while $x \leq$ round$(m \cdot \tau)$ do
    $\delta_i$ = Draw$(\Delta^{\mathcal{I}})$;
    $\Theta$=False;
    while $\neg\Theta$ do
        $C_j$ = Draw$(\mathbf{C})$;
        if $\{\{C, C_j\} | \delta_i \in C^{\mathcal{I}}\} \cap \mho_T = \emptyset$ then $\Theta$=True;
    $C_j(\delta_i)$;
    x++;
```

2.2 Role Assertions

JPoT deals only with unequivocal TBox [10] as a finite set of equivalences or inclusions for which the left-hand side of each axiom is an atomic concept and such that for every atomic concept there is at most one axiom where it occurs on the left-hand side. As evoked in [13], An $\mathcal{ALC}$ TBox can be translated into an equi-satisfiable TBox using structural transformations. For this, a TBox can be converted into negation normal form in which every concept C that occurs immediately below a role restriction is replaced by a definer D (adding the axiom $D \sqsubseteq C$ for each such subconcept). The resulting TBox does not contain any nested role restriction and can be brought into $\mathcal{ALC}$-conjunctive normal.

We define a GCI part of an unequivocal TBox as the set of axioms obtained by removing the concept inclusions of split concept equivalences $((C \equiv D) \Leftrightarrow (C \sqsubseteq D) \wedge (D \sqsubseteq C))$ where the single concept names are located on the right-hand side of the axioms. If a GCI part of an unequivocal TBox is in $\mathcal{SHIQ}^{(\mathcal{D})}$-conjunctive form (denoted TBox$'$) then every axiom t_i is of the form $\top \sqsubseteq L \sqcup L_1 \sqcup \cdots \sqcup L_n$ with L_j is a $\mathcal{SHIQ}^{(\mathcal{D})}$-literal and L is of the form $\neg A$ with A a named concept. A $\mathcal{SHIQ}^{(\mathcal{D})}$-literal is a concept description of the form A, $\neg A$, $\forall r.D$, $\exists r.D$, $\leq nr.D$, $\geq nr.D$ or $=nr.D$ where A is a named concept and D is a definer. A derivation of the $\mathcal{SHIQ}^{(\mathcal{D})}$-conjunctive form of a GCI part of an unequivocal TBox is defined like that: TBox$'' = \{\mathcal{M}(ti) | t_i \in$ TBox$'\}$ s.t. $\mathcal{M}(t_i) = \bigcup_j (\neg L \triangleq L_j)$ with L_j a concept description of the form $\forall r.D$, $\exists r.D$, $\leq nr.D$, $\geq nr.D$ or $= nr.D$.

Roughly speaking, `JPoT` draws abstract roles and tries to instantiate them. The Algorithm 2 details the heuristic of the population. Once an abstract role is drawn, an individual subject is drawn and the concepts it already instantiates are confronted to the set composed by the domains of the drawn abstract role and the domains of all its super roles. In case of disjointness another individual is drawn. After this step, the concepts of the chosen individual are confronted with the universal quantification axioms implying the drawn abstract role and all its super roles in order to build a set of mandatory concepts for the individual object that has to be drawn in the next step. In the event the drawn individual i-has a concept in common with each mandatory one, ii-doesn't have concepts disjoints with the range of the drawn abstract roles and all its super roles and iii-doesn't violate possible max cardinality restrictions occurring with the concepts of the individual subject and the drawn abstract roles (if it's a simple role present in $\mathbf{R_S}$) and all its super roles, this individual object will continue as the candidate for the instantiation. In case of a violation of some of the cardinality restrictions, another abstract role will be drawn and the process of selection will restart. Note that in case of an empty set of domains or ranges for the drawn abstract role and all its super roles, the populator uses a set of pairs containing the concepts of the axioms implying an existential quantification with the drawn abstract role and all its super roles. One pair is randomly drawn and will serve to guide the assertions if needed. In $\mathcal{SHIQ}$, a consequence of a role assertion is the possible generation of other role assertions due to the presence of inverse roles of the drawn abstract role that could possibly break the consistency of the knowledge base due to possible role restrictions (e.g. maximal cardinalities). A parallel set Abox' (using the same concept name and role name than the Tbox followed by a') is then maintained along the process in which are stored all the aforementioned generated role assertions. Thus, before proceeding to the role instantiation in case of an inverse role definition, the corresponding inverse role assertion is added to Abox'. Then, the populator checks if the individual object of the generated assertion doesn't violate possible role restrictions. In case of non violation, the drawn individual object will participate to the instantiation with the abstract role and the individual subject.

Algorithm 2. Role Assertions (RAs)

```
Data: TBox, TBox'', m, $\tau$, $\rho$
Result: $\mathsf{round}(m \cdot (1-\tau) \cdot \rho)$ RAs
y= 0;
while $y \leq \mathsf{round}(m \cdot (1-\tau) \cdot \rho)$ do
  $\Xi$ =False;
  while $\neg\Xi$ do
    $\Xi$ =True;
    $r_k = \mathsf{Draw}(\mathbf{R_A})$;
    $P_\exists = \{(C,D)|(C \triangleq \exists s_p.D) \wedge (r_k \sqsubseteq^+_{\mathcal{R}} s_p)\}$;
    $(A,B) = \mathsf{Draw}(P_\exists)$;
    if $(\bigcup_p$ domain$(s_p)$ $s.t.$ $r_k \sqsubseteq^+_{\mathcal{R}} s_p) = \emptyset$ then $I = \{A\}$ else $I = \bigcup_p$
     domain$(s_p)$;
    $\Theta$=False;
    while $\neg\Theta$ do
      $\delta_i = \mathsf{Draw}(\Delta^{\mathcal{I}})$ s.t. $\exists C.\ \delta_i \in C^{\mathcal{I}}$;
      if $\{\{C,I\}|\delta_i \in C^{\mathcal{I}}\} \cap \Omega_T = \emptyset$ then $\Theta$=True;
    $P_\forall = \{(C,D)|(C \triangleq \forall s_p.D) \wedge (r_k \sqsubseteq^+_{\mathcal{R}} s_p)\}$;
    $\forall\ c \in \{C|\delta_i \in C^{\mathcal{I}}\}$
      if $\exists(E,F) \in P_\forall$ $s.t.$ $c \sqsubseteq E$ then $F \in W$;
    if $(\bigcup_p$ range$(s_p)$ $s.t.$ $r_k \sqsubseteq^+_{\mathcal{R}} s_p) = \emptyset$ then $J = \{B\}$ else $J = \bigcup_p$
     range$(s_p)$;
    $\Theta$=False;
    while $\neg\Theta$ do
      $\delta_j = \mathsf{Draw}(\Delta^{\mathcal{I}})$ s.t. $\exists C.\ \delta_j \in C^{\mathcal{I}}$;
      if $\{\{C,J\}|\delta_j \in C^{\mathcal{I}}\} \cap \Omega_T = \emptyset$ AND $W \subseteq \{C|\delta_j \in C^{\mathcal{I}}\}$ then
       $\Theta$=True;
    $T$=$\{\langle C,D,l\rangle|((C \triangleq \leq l s_p.D) \vee (C \triangleq = l s_p.D)) \wedge (r_k \sqsubseteq^+_{\mathcal{R}} s_p)\}$;
    $l = 0$;
    if $r_k \in \mathbf{R_S}$ then
      $\forall\ c \in \{C|(\delta_i \in C^{\mathcal{I}}) \wedge (\nexists(B \sqsubseteq C).\delta_i \in B^{\mathcal{I}})\}$
        $\forall\ \langle E,F,L\rangle \in T$ $s.t.$ $c \sqsubseteq E$
          $\forall\ d \in \{D|(\delta_h \in D^{\mathcal{I}}) \wedge (\delta_i,\delta_h) \in (q_p \cup q'_p) \wedge (\nexists(C \sqsubseteq D).\delta_h \in D^{\mathcal{I}})\}$
            if $(d \subseteq F) \wedge (q_p \sqsubseteq^+_{\mathcal{R}} r_k)$ then $l$++;
          if $l = L$ then $\Xi$ = False; $l = 0$; else $l = 0$;
    if $(\mathtt{Inv}(r_k) \cap \mathbf{R_A}) \neq \emptyset$ then
      $(\delta_j,\delta_i) \in r'^{-}_k$;
      $T'$=$\{\langle C,D,l\rangle|((C \triangleq \leq l s_p.D) \vee (C \triangleq = l s_p.D)) \wedge (r^-_k \sqsubseteq^+_{\mathcal{R}} s_p)\}$;
      $\forall\ c \in \{C|(\delta_j \in C^{\mathcal{I}}) \wedge (\nexists(B \sqsubseteq C).\delta_j \in B^{\mathcal{I}})\}$
        $\forall\ \langle E,F,L\rangle \in T'$ $s.t.$ $c \sqsubseteq E$
          $\forall\ d \in \{D|(\delta_h \in D^{\mathcal{I}}) \wedge (\delta_j,\delta_h) \in (q_p \cup q'_p) \wedge (\nexists(C \sqsubseteq D).\delta_h \in D^{\mathcal{I}})\}$
            if $(d \subseteq F) \wedge (q_p \sqsubseteq^+_{\mathcal{R}} r^-_k)$ then $l$++;
          if $l = L$ then $\Xi$ = False; $l = 0$; else $l = 0$;
  $(\delta_i,\delta_j) \in r_k$;
  y++;
```

2.3 Data Assertions

The populations performed by JPoT follow heuristics that guarantee consistencies of the knowledge bases only on the basis of a set of axioms, in other words, without any consideration of a specific domain. The produced ABoxes are somewhere semantically consistent because the population respects the semantic rules of the world present in the TBoxes and RBoxes. While the URIs of the individuals are all based on an integer in a selected range, the concepts they instantiate (concept assertions) and more the relation in which they are involved (role assertions) can represent an artificial but apparently real world. Let's imagine the equivalent of the Turing test for a SDG model in which the ABoxes should appear as real 70% of the time to succeed the test.

Algorithm 3. Data Assertions (DAs)

```
Data: TBox, TBox'', m, $\tau$, $\rho$
Result: $\mathsf{round}(m \cdot (1-\tau) \cdot (1-\rho))$ DAs
z= 0;
while $z \leq \mathsf{round}(m \cdot (1-\tau) \cdot (1-\rho))$ do
    $\Xi$ =FALSE;
    while $\neg\Xi$ do
        $\Xi$ =TRUE;
        $t_k = \mathsf{Draw}(\mathbf{R_T})$;
        $P_\exists = \{(C,D)|C \triangleq \exists t_k.D\}$;
        $(A,B) = \mathsf{Draw}(P_\exists)$ with $P_\exists = \{(C,D)|(C \triangleq \exists s_p.D) \wedge (r_k \sqsubseteq^+_\mathcal{R} s_p)\}$;
        if $(\bigcup_p \mathsf{domain}(s_p)$ $s.t.$ $t_k \sqsubseteq^+_\mathcal{R} s_p) = \emptyset$ then $I = \{A\}$ else $I = \bigcup_p$
          $\mathsf{domain}(s_p)$;
        $\Theta$=FALSE;
        while $\neg\Theta$ do
            $\delta_i = \mathsf{Draw}(\Delta^\mathcal{I})$ s.t. $\exists C.\ \delta_i \in C^\mathcal{I}$;
            if $\{\{C,I\}|\delta_i \in C^\mathcal{I}\} \cap \Omega_T = \emptyset$ then $\Theta$=TRUE;
        $P_\forall = \{(C,D)|(C \triangleq \forall s_p.D) \wedge (t_k \sqsubseteq^+_\mathcal{R} s_p)\}$;
        $\forall\ c \in \{C|\delta_i \in C^\mathcal{I}\}$
            if $\exists(E,F) \in P_\forall$ $s.t.$ $c \sqsubseteq E$ then $F \in W$;
        if $(\bigcup_p \mathsf{range}(s_p)$ $s.t.$ $t_k \sqsubseteq^+_\mathcal{R} s_p) = \emptyset$ then $J = \{B\}$ else $J = \bigcup_p$
          $\mathsf{range}(s_p)$;
        $d_j = \mathsf{Gen}(J, t_k)$;
        $T$=$\{\langle C,D,l\rangle|((C \triangleq\leq ls_p.D) \vee (C \triangleq=ls_p.D)) \wedge (t_k \sqsubseteq^+_\mathcal{R} s_p)\}$;
        $l = 0$;
        $\forall\ c \in \{C|(\delta_i \in C^\mathcal{I}) \wedge (\nexists(B \sqsubseteq C).\delta_i \in B^\mathcal{I})\}$
            $\forall\ \langle E,F,L\rangle \in T$ $s.t.$ $c \sqsubseteq E$
                $\forall\ d \in \{D|(d_h \in D^\mathcal{I}) \wedge (\delta_i, d_h) \in q_p \wedge (\nexists(C \sqsubseteq D).d_h \in D^\mathcal{I})\}$
                    if $(d \subseteq F) \wedge (q_p \sqsubseteq^+_\mathcal{R} t_k)$ then $l$++;
                if $l = L$ then $\Xi$ = FALSE; $l = 0$; else $l = 0$;
    $(\delta_i, d_j) \in t_k$;
    z++;
```

Even if we didn't lead this experiment, we can objectively assume that a JPoT's population implying only concept and role assertions could pass this test. However, the test gets much more complicated to pass when JPoT has to deal with concrete roles due to a set of issues concerning the creation of data values. First, the usage of datatype has to be tackled with caution under a penalty of undecidability. OWL2 solved this problem by recommending datatypes defining a datatype map[3] which lists the datatypes that can be used in the knowledge bases. Even restricting a population to this subset of datatypes, JPoT could fail a Turing test for SDG by strictly following the heuristic described in the Algorithm 3 due to a lack of semantic soundness in the generation of the data values for the data assertions. For example, let's imagine a TBox with the following axioms: each person has exactly one age that is an integer, an author has exactly one number of citations and hindex that are integers and if a person has a name it is always a string. As it is, JPoT could give an inhuman age for a person, assert a number of citations inferior to the hindex squared for the same author and provide an absurd name that would just be a random sequence of characters. We implemented functions $\mathsf{Gen}(D, t)$ generating data values in adequation with a datatype D and with a concrete role t facing with different issues.

Using a Facet Space. The usage of facet spaces is the royal road for the modeler to obtain a knowledge base capable of passing the SDG Turing test. The facet space was introduced as a set of pairs of the form (F, v) where F is a constraining facet and v a constraining value. Each such pair is mapped to a subset of the value space of the datatype. Thus, the data range of concrete roles can be restricted using a *datatypeRestriction* which restricts the value space of a datatype by a constraining facet. In the example of hasAge, the modeler can use a *datatypeRestriction* that would restrict the datatype xsd:nonNegativeInteger by using a singleton facet space i.e. the pair (xsd:maxExclusive, "123"^^xsd:nonNegativeInteger) that corresponds to the limit for an age never reached in the history of the humanity. In addition in JPoT, we used a Gaussian distribution (with an expectation of 42 and a standard deviation of 10) in order to simulate an age distribution of researchers following a pyramidal shape.

Using a Linear Equation. Ensuring the soundness of data assertions restricting dataranges using *datatypeRestrictions* can still produce knowledge bases incapable of passing the SDG Turing test. In fact, one can say that the value of the subject of a data assertion has to be an integer less than 123, but one cannot say that the value of the subject of one data assertion is less than that of another data assertion. In the example of citations and hindex, the modeler doesn't have the ability with the current expressivity of OWL2 to constrain the TBox with the fact that the total number of citations of an author is greater than the hindex-squared. A proposition of extension for OWL2[4] allows the expression of linear equations but not of polynomial equations.

[3] https://www.w3.org/TR/owl2-syntax/#Datatype_Maps.

[4] https://www.w3.org/2007/OWL/wiki/Data_Range_Extension:_Linear_Equations.

Using an API. As we described for a numerical datatype, it is also possible to restrict dataranges using a *datatypeRestriction* for String with a pattern (well known as a regular expression). This expressivity can be useful to generate knowledge bases capable of passing the SDG Turing test. For example, a model that would represent a social security number could use such a pattern. But when the concrete role is hasName, even the usage of space facet doesn't prevent a production of knowledge ineligible for a SDG Turing test. In this case, the only solution is to use another generator or API to produce a soundness value. Then we used the API JaNaG[5] (Java Name Generator) which is a random name generator based on a name fragment database that creates relatively reasonably sounding names from different cultures/influences.

3 Performances of JPOT

We performed an evaluation of the populator JPoT (the version of JPoT based on the OWL-API 4.1.0) using a simplified model of a scholar domain.

3.1 Domain TBox

First, we highlight here that the ontological choices made in this example are intuitive, but arguable. Our aim was to describe the performance of JPoT, not to propose an ontological analysis of a scholar domain.

In the domain described in Fig. 1, Authors (which are Persons with names) write Publications, which can be classified into Papers, Articles, Chapters or

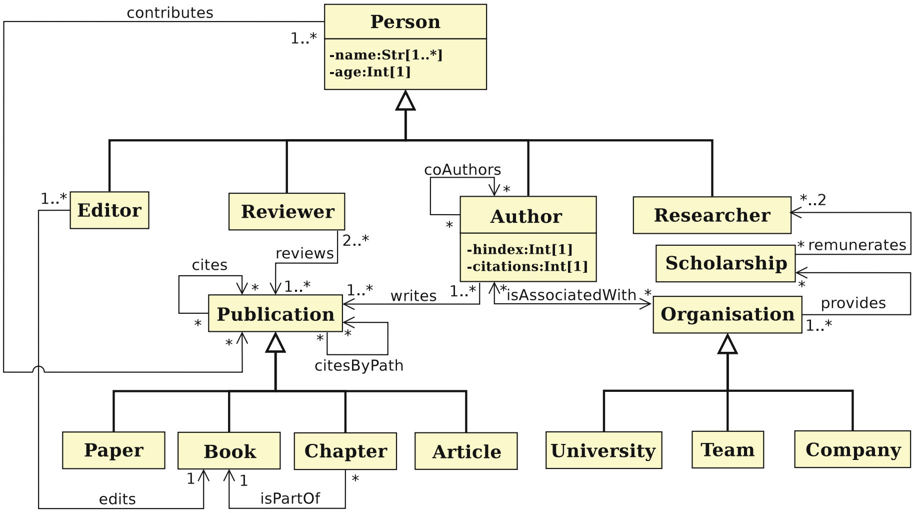


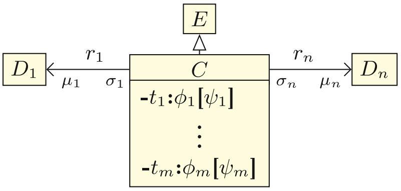
Fig. 1. UML diagram of a scholar domain

[5] https://github.com/mkalus/janag.

Books. A Chapter can be part of maximum one Book. Persons can contribute for Publications. A Publication can be cited by another Publication and cited by path in case of a directed path in the citations graph, Authors can have co-authors and have a total number of citations and an hindex. A Researcher can be remunerated by Scholarships provided by Organisations in such a way that an Organisation can provide several Scholarships, and a Scholarship can remunerate a maximum of two Researchers (e.g. organisations authorizing to change one time the "owner" of a scholarship). Authors and Organisations can be associated. There are two kinds of Organisations: Team and University.

We introduce a UML interpretation in Definition 7 for a subset of the fragment $\mathcal{SHIQ}^{\mathcal{D}}$. We denote $\mathbf{C_A}$ a set of concept names, $\mathbf{R_A}$ a set of abstract role names, $\mathbf{R_T}$ a set of concrete role names, $\mathbf{D_T}$ a set of datatype names and S a set of symbols for Descriptions logics interpreted in first order logic [10], Ω a set of UML cardinalities and a function $\ell(\Omega \to S)$ such that $\ell(*) \mapsto \forall$, $\ell(n) \mapsto \forall_{=n}$, $\ell(0..n) \mapsto \forall_{\leq n}$ and $\ell(n..*) \mapsto \forall_{\geq n}$ (with $n>0$).[6]

Definition 7. *Let* $\{C, D_1, \ldots, D_n, E\} \subseteq \mathbf{C}$, $\{r_1, \ldots, r_n\} \subseteq \mathbf{R}$, $\{t_1, \ldots, t_m\} \subseteq \mathbf{R_T}$, $\{\phi_1, \ldots, \phi_m\} \subseteq \Phi_T$ *and* $\{\mu_1, \ldots, \mu_n, \sigma_1, \ldots, \sigma_n, \psi_1, \ldots, \psi_m\} \subseteq \Omega$*:*



$$\Leftrightarrow \quad \begin{gathered} (C \sqsubseteq E \sqcap \prod_{i=1}^{n} \ell(\mu_i) r_i . D_i \sqcap \prod_{j=1}^{m} \ell(\psi_j) t_j . \phi_j)^{\mathrm{U}} \\ (D_1 \equiv \ell(\sigma_1) r_1^{-1} . C)^{\mathrm{U}} \\ \vdots \\ (D_n \equiv \ell(\sigma_n) r_n^{-1} . C)^{\mathrm{U}} \end{gathered}$$

3.2 Domain Rbox

Our domain RBOX is composed by the following axioms: Inv(cites) = isCited, Inv(citesByPath) = isCitedByPath, Inv(coAuthors) = coAuthors, Inv(contributes) = isContributedBy, Inv(isAcknowledgedBy) = acknowledges, Inv(reviews) = isReviewedBy, Inv(writes) = isWrittenBy, Inv(edits) = isEditedBy, Inv(hasPart) = isPartOf, Inv(provides) = isProvidedBy, Inv(isAssociatedWith) = isAssociatedWith, Inv(remunerates) = isRemuneratedBy, cites $\sqsubseteq$ citesByPath, isAcknowledgedBy $\sqsubseteq$ contributes, reviews $\sqsubseteq$ contributes, writes $\sqsubseteq$ contributes, Tra(citesByPath) = TRUE.

3.3 Generated Aboxes

Our empirical analysis has been performed on a machine equipped with an Intel Core at 3.30 GHz and an Ubuntu 15.04. We ran the Java-based reasoner Pellet 2.4.0 with Sun Java 1.8, and we set the maximum heap space to 7.5 GB. We populated a part of our domain TBOX (added with our domain RBOX in case of a population in $\mathcal{SHIQ}$).

[6] $\forall_{=n} r.C \equiv \forall r.C \sqcap {=}nr.C$, $\forall_{\geq n} r.C \equiv \forall r.C \sqcap {\geq}nr.C$ and $\forall_{\leq n} r.C \equiv \forall r.C \sqcap {\leq}nr.C$.

We performed all the populations with n and m equal and with $\tau = 0.5$ meaning in other words that the half of the assertions were concept assertions. For each pair (n, m), we performed three populations in $\mathcal{ALCQ}^{(\mathcal{D})}$: i- with $\rho = 0$, ii- with $\rho = 0.5$ and iii- with $\rho = 1$ and one population in $\mathcal{SHIQ}$ with $\rho = 1$. We used a range of individuals and assertions going up to around one million.

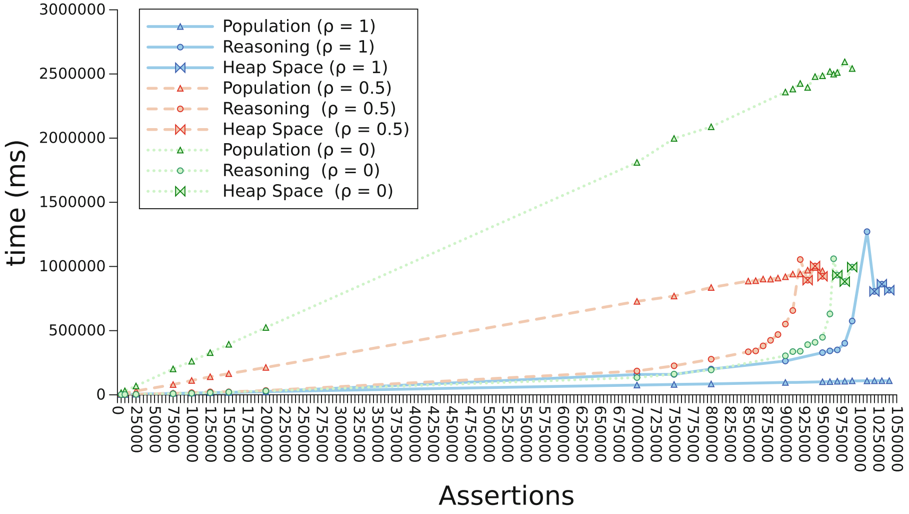


Fig. 2. Some metrics about JPoT's populations

Pellet output systematically that the ABoxes were consistent with the part of the TBox in $\mathcal{ALCQ}^{(\mathcal{D})}$ (and the TBox and RBox in $\mathcal{SHIQ}$) when the consistency checks were possible. We can observe on the Fig. 2, the evolutions of CPU times for the populator JPoT and the heap space limits (marked with bow ties) concerning the consistency checks of Pellet in $\mathcal{ALCQ}^{(\mathcal{D})}$. After a certain amount of assertions, the reasoning tasks were impossible to conclude due to a heap space error thrown whenever the JVM reached the heap size limit. In $\mathcal{ALCQ}^{(\mathcal{D})}$, for $\rho = 0.5$ at $n, m \approx 930.000$, the first limit was detected. The same limit appeared for $\rho = 0$ at $n, m \approx 970.000$ and for $\rho = 1$ at $n, m \approx 1.020.000$. In $\mathcal{SHIQ}$, for $\rho = 1$ the limit is not yet reached at $n, m = 1.000.000$.

4 Related Works

Very few SDGs have been developed in order to perform some stress tests before deploying an ontology in a Semantic Web application. According to the creators, the Lehigh University Benchmark [14] was the first knowledge base generator. The idea of LUBM was to feature a university domain ontology with one statically predefined TBox, and allow different sizes of an artificial generated dataset i.e. an ABox. Due to the growing need to better profile the behaviour of an ontology with regards to differing numbers and complexities of the axioms in the TBox, an extension of LUBM, the so-called University Ontology Benchmark (UOB) has been introduced [15]. Another remarkable tool named OTAGen [16] has the specificity to generate complete knowledge bases providing the capibility of specifying a large range of parameters characterizing them, both on TBox as well as Abox level, the tool can also generate some corresponding queries. Note that after this proposal, another approach proposed a purely TBox generation with different reasoning complexities resulting from the relative proportions of the design patterns of biomedical structures representation [17]. Finally, the only approach that can handle a lack or inaccessibility of data in Aboxes when a TBox is already available is SKTI - a synthetic data generator [18]. This system generates synthetic instances based on a source ontology and user specifications. Note that to mimic the real world scenario the system also allows the insertion of noisy and erroneous instances into the dataset. This last solution is the closest related work with JPoT as a domain independent SDG that can populate TBoxes provided by users. Finally, the reasoners engineers could devote more attention for ontology populations in the future. For example, during the first edition of the OWL Reasoner Evaluation (ORE) in 2012 [19], a method was proposed to generate a benchmark and effectively evaluate semantic reasoners by generating realistic synthetic semantic data [20]. Furthermore, frameworks may need to dispose such steps of population in order to empirically evaluate the reasoning tasks [21]. This situation occurs in the community of conceptual modeling where it is a frequent practice to design model alternatives and perform an empirical comparison between them [22].

5 Conclusion

In this paper, we just presented an extension of another populator of TBox: JPoT[7]. To the best of our knowledge, this is the first domain independent SDG guaranteeing consistency of knowledge bases now founded on TBoxes expressed in $\mathcal{SHIQ}^{(\mathcal{D})}$. Moreover, the issues of the data assertions were tackled and some performances presented. For future work, we intend that JPoT deals with the highest expressiveness to cover the whole fragment $\mathcal{SROIQ}^{(\mathcal{D})}$.

Acknowledgments. This work has been made possible by "la Regione Autonoma della Sardegna e Autorità Portuale di Cagliari con L.R. 7/2007, Tender 16 2011, CRP-49656 con il projeto: Metodi innovativi per il supporto alle decisioni riguardanti l'ottimizzazione delle attività in un terminal container" and by "o EDITAL FAPES/CAPES N°009/2014 (Bolsa de fixacão de doutore N°71047522) com a proposta: Melhor integração de tecnologias de representação de conhecimento e raciocínio nas utilizações local e Web".

[7] JPoT is available at http://bit.ly/2vh5YE4.

References

1. Melz, E.R., Macgregor, R.M.: Design, implementation, and analysis of a parallel description classifier. Technical report, University of Southern California Marina Del Rey Information Sciences Inst (1995)
2. Guizzardi, G.: Ontological foundations for structural conceptual models. Ph.D. thesis, Centre for Telematics and Information Technology, University of Twente, Enschede, The Netherlands (2005)
3. Bedini, I., Nguyen, B.: Automatic ontology generation: state of the art. PRiSM Laboratory, Technical report. University of Versailles (2007)
4. Guarino, N., Schneider, L.: Ontology-driven conceptual modelling. In: Spaccapietra, S., March, S.T., Kambayashi, Y. (eds.) ER 2002. LNCS, vol. 2503, p. 10. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45816-6_4
5. Baader, F., Horrocks, I., Sattler, U.: Description logics as ontology languages for the semantic web. In: Hutter, D., Stephan, W. (eds.) Mechanizing Mathematical Reasoning. LNCS (LNAI), vol. 2605, pp. 228–248. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-32254-2_14
6. De Giacomo, G., Lenzerini, M.: TBox and ABox reasoning in expressive description logics. In: Aiello, L.C., Doyle, J., Shapiro, S.C. (eds.) Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning (KR 1996), Cambridge, Massachusetts, USA, 5–8 November 1996, pp. 316–327. Morgan Kaufmann, Los Altos (1996)
7. Bourguet, J.R.: JPoT: Just another Populator of TBoxes. In: Ventura, J.A.L., Alatrista-Salas, H. (eds.) Proceedings of the 4th Annual International Symposium on Information Management and Big Data - SIMBig 2017. CEUR Workshop Proceedings. CEUR-WS.org (2017, in press)
8. Schmidt-Schauß, M., Smolka, G.: Attributive concept descriptions with complements. Artif. Intell. **48**, 1–26 (1991)
9. Sattler, U.: A concept language extended with different kinds of transitive roles. In: Görz, G., Hölldobler, S. (eds.) KI 1996. LNCS, vol. 1137, pp. 333–345. Springer, Heidelberg (1996). https://doi.org/10.1007/3-540-61708-6_74
10. Baader, F., Nutt, W.: Basic description logics, pp. 43–95. Cambridge University Press (2003)
11. Horrocks, I., Sattler, U., Tobies, S.: Practical reasoning for expressive description logics. In: Ganzinger, H., McAllester, D., Voronkov, A. (eds.) LPAR 1999. LNCS (LNAI), vol. 1705, pp. 161–180. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-48242-3_11
12. Horrocks, I., Sattler, U., Tobies, S.: Reasoning with individuals for the description logic $\mathcal{SHIQ}$. In: McAllester, D. (ed.) CADE 2000. LNCS (LNAI), vol. 1831, pp. 482–496. Springer, Heidelberg (2000). https://doi.org/10.1007/10721959_39
13. Koopmann, P., Schmidt, R.A.: Uniform interpolation of $\mathcal{ALC}$-ontologies using fixpoints. In: Fontaine, P., Ringeissen, C., Schmidt, R.A. (eds.) FroCoS 2013. LNCS (LNAI), vol. 8152, pp. 87–102. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40885-4_7
14. Guo, Y., Pan, Z., Heflin, J.: LUBM: a benchmark for OWL knowledge base systems. Web Sem. Sci. Serv. Agents World Wide Web **3**(2), 158–182 (2005)
15. Ma, L., Yang, Y., Qiu, Z., Xie, G., Pan, Y., Liu, S.: Towards a complete OWL ontology benchmark. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 125–139. Springer, Heidelberg (2006). https://doi.org/10.1007/11762256_12

16. Ongenae, F., Verstichel, S., De Turck, F., Dhaene, T., Dhoedt, B., Demeester, P.: OTAGen: a tunable ontology generator for benchmarking ontology-based agent collaboration. In: 32nd Annual IEEE International on Computer Software and Applications, pp. 529–530. IEEE (2008)
17. Boeker, M., Hastings, J., Schober, D., Schulz, S.: A T-Box generator for testing scalability of OWL mereotopological patterns. In: Dumontier, M., Courtot, M. (eds.) Proceedings of the 8th International Workshop on OWL: Experiences and Directions. CEUR Workshop Proceedings, vol. 796 (2011)
18. Chowdhury, N.: Ontoevaluator – SKTI synthetic data generator synthetic data generator (2012)
19. Horrocks, I., Yatskevich, M., Jiménez-Ruiz, E. (eds.): Proceedings of the 1st International Workshop on OWL Reasoner Evaluation (ORE-2012), Manchester, UK, 1 July 2012. CEUR Workshop Proceedings, vol. 858. CEUR-WS.org (2012)
20. Li, Y., Yu, Y., Heflin, J.: Evaluating reasoners under realistic semantic web conditions. In: Horrocks, I., Yatskevich, M., Jiménez-Ruiz, E. (eds.) Proceedings of the 1st International Workshop on OWL Reasoner Evaluation. CEUR Workshop Proceedings, vol. 858. CEUR-WS.org (2012)
21. Bourguet, J.R., Pulina, L.: TROvE: a graphical tool to evaluate OWL reasoners. In: Bail, S., Glimm, B., Jiménez-Ruiz, E., Matentzoglu, N., Parsia, B., Steigmiller, A. (eds.) Informal Proceedings of the 3rd International Workshop on OWL Reasoner Evaluation (ORE 2014) Co-Located with the Vienna Summer of Logic (VSL 2014), Vienna, Austria, 13 July 2014. CEUR Workshop Proceedings, vol. 1207, pp. 30–35. CEUR-WS.org (2014)
22. Batsakis, S., Petrakis, E.G.M., Tachmazidis, I., Antoniou, G.: Temporal representation and reasoning in OWL 2. Sem. Web **8**(6), 981–1000 (2017)

Language Identification with Scarce Data: A Case Study from Peru

Alexandra Espichán-Linares and Arturo Oncevay-Marcos(✉)

Research Group on Artificial Intelligence (IA-PUCP),
Pontificia Universidad Católica del Perú, Lima, Peru
a.espichan@pucp.pe, arturo.oncevay@pucp.edu.pe

Abstract. Language identification is an elemental task in natural language processing, where corpus-based methods reign the state-of-the-art results in multi-lingual setups. However, there is a need to extend this application to other scenarios with scarce data and multiple classes to face, analyzing which of the most well-known methods is the best fit. In this way, Peru offers a great challenge as a multi-cultural and linguistic country. Therefore, this study focuses in three steps: (1) to build from scratch a digital annotated corpus for 49 Peruvian indigenous languages and dialects, (2) to fit both standard and deep learning approaches for language identification, and (3) to statistically compare the results obtained. The standard model outperforms the deep learning one as it was expected, with 95.9% in average precision, and both corpus and model will be advantageous inputs for more complex tasks in the future.

1 Introduction

Peru is a diverse cultural country with almost 4 million people who are speakers of an indigenous language among the 47 Peruvian official languages divided by 19 linguistic families. These languages are distributed across the highlands and jungle (Amazon) regions, and most of them are very unique, in spite of their geographical or linguistic closeness [16].

The linguistic diversity calls for equal opportunity across the different indigenous communities, and this could be supported by high-level bilingual education and a deep knowledge about the languages. Computational approaches nowadays provide assisted technology for these applications in different levels, and one of the first required tools is an automatic language detector for written text in different granularity levels, such as a complete document, a paragraph, a sentence or even a word [12].

To develop an automatic language identifier, a basic natural language processing (NLP) task, an annotated textual corpus for the languages is what is needed, as in any supervised learning process. However, not all the languages offer large enough digital and available corpus for computational tasks, so they are known as low-resourced or minority languages from a computer science point of view [6].

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 90–105, 2018.
https://doi.org/10.1007/978-3-319-90596-9_7

Consequently, it is necessary to build a digital repository of textual corpora for these languages as a first step. After that, the language identification could be addressed as a classification task, which includes a deep learning focus that differs from the standard supervised learning algorithms. For that reason, this study will compare the effectiveness of both approaches in the scarce data scenario offered by the Peruvian languages dataset acquired. The corpora is composed by 29 indigenous languages plus 20 dialects, making 49 different classes as targets, and the analysis is performed at sentence level.

In the next section, the Peruvian indigenous languages used in this work are presented. Then, Sect. 3 describes studies related to the language identification task mainly in scarce data scenarios. After that, Sect. 4 presents the corpus building process and the details of the dataset retrieved. Following that, Sect. 5 focuses in the language identification model for both approaches, while Sect. 6 presents the comparative evaluation through statistical tests. The results and discussions are included in Sect. 7. Finally, the conclusions and future work for the study are presented in Sect. 8.

2 Peruvian Indigenous Languages

Among the 47 languages spoken by Peruvian people, 43 are Amazonian (from the jungle) and 4 are Andean (from the highlands). These languages are considered prevailing languages because they have live speakers. Therefore, there are 19 linguistic families (a set of languages related to each other and with a common origin): 2 Andean (Aru and Quechua) and 17 Amazonian [16].

The 47 indigenous languages are highly agglomerative, unlike Spanish (Castillan), the main official language in the country. Even though, most of them presents more than 100 morphemes for the word formation process. For instance, *Quechua del Cusco* contains 130 suffixes [20], meanwhile *Shipibo-konibo* uses 114 suffixes plus 31 prefixes [23].

In this work, the language identification task was performed on 29 languages plus 20 dialects from 15 linguistic families. The dialects sub-group considers 18 other variants of Quechua and 2 additional of Asháninka. The ISO-639-3 codes and the approximate number of speakers of each language and dialect are presented in Table 1.

3 Related Work

Standard Supervised Learning. Given that Peruvian languages can be considered as low-resourced ones, a systematic search for studies focused on automatic language identification for low-resourced languages was performed. The main studies retrieved are described as follows.

Grothe et al. [7] compared the performance of three feature extraction approaches for language identification using the Leipzig Corpora Collection [19] and randomly selected Wikipedia articles. The considered representations for

Table 1. Basic information of the languages and dialects within the scope of the study.

Linguistic family	Language/Dialect	ISO-639-3	Speakers
Arawak	Asháninka	cni	88703
	Ashéninka del Pajonal	cjo	10000
	Ashéninka del Pichis	cpu	8774
	Kakinte	cot	439
	Matsigenka	mcb	11275
	Nomatsigenga	not	8016
	Yanesha	ame	7523
	Yine	pib	3261
Aru	Aymara	aym	443248
Bora	Bora	boa	748
Cahuapana	Shawi	cbt	21650
Huitoto	Murui	huu	1864
Jíbaro	Achuar	acu	11087
	Awajún	agr	55366
	Wampis	hub	10163
Kandozi	Kandozi	cbu	3255
Pano	Amahuaca	amc	301
	Capanahua	kaq	348
	Cashinahua	cbs	2419
	Kakataibo	cbr	1879
	Matses	mcf	1724
	Sharanahua	mcd	486
	Shipibo-konibo	shp	22517
	Yaminahua	yaa	600
Peba-yagua	Yagua	yad	5679
Quechua	Quechua de Ambo-Pasco	qva	90000
	Quechua de Ayacucho	quy	850050
	Quechua de Cajamarca	qvc	30000
	Quechua de Huallaga	qub	40000
	Quechua de Huamalíes	qvh	72500
	Quechua de Lambayeque	quf	21496
	Quechua de Margos	qvm	114000
	Quechua de Panao	qxh	50000
	Quechua de San Martín	qvs	44000
	Quechua de Yauyos	qux	6500
	Quechua del Callejon de Huaylas	qwh	300000
	Quechua del Cusco	quz	1115000
	Quechua del Este de Apurimac	qve	218352
	Quechua del Norte de Conchucos	qxn	250000
	Quechua del Norte de Junín	qvn	60000
	Quechua del Norte de Pastaza	qvz	2100
	Quechua del Sur de Conchucos	qxo	250000
	Quechua del Sur de Pastaza	qup	2200
	Quechua Huaylla Wanca	qvw	298560
Shimaco	Urarina	ura	4854
Tacana	Ese eja	ese	588
Tikuna	Tikuna	tca	6982
Tucano	Secoya	sey	921
Záparo	Arabela	arl	403

features were short words (SW), frequent words (FW) and n-grams (NG). Meanwhile, the employed classification method was Ad-Hoc Ranking. Therefore, the best obtained results for each technique were: FW 25% (99.2%), SW 4 (94.1%) and NG with 3-grams (79.2%).

The potential factors that might affect the performance of text-based language identification was the research goal in [2]. Botha et al. focused in the 11 official languages of South Africa, using NG as language features. In the study, 3 classification methods were tested: SVM (Support Vector Machine), Naive Bayes and N-gram rank ordering on different training and testing text sizes. In this way, it was found that the 6-gram Naive Bayes model had the best performance in general, achieving a 99.4% accuracy value for large training-testing sets and 83% for shorter ones.

In [12], Malmasi et al. presented the first study to distinguish texts between the Persian and Dari languages at the sentence level. As Dari is a low-resourced language, it was developed a 28 thousand sentences corpus for this task, and they used 14 thousand for each language. For the representation phase, characters and sentences n-grams were considered as language features. Finally, using a SVM implementation within a classification ensemble scheme, they discriminated both languages with 96% accuracy.

Selamat et al. [21] proposed a language identification algorithm based on lexical features that works with a minimum amount of training data. For this study, a dataset of 15 mainly low-resourced languages, was extracted from the Universal Declaration of Human Rights. The used technique is based on a spelling checker-based method [17] and the improvement proposed in this research was related to the indexation of the vocabulary words regarding its length. In this way, the average precision of the method was 93% and an improvement of 73% in execution time was achieved.

Deep Supervised Learning. Recent studies have used deep learning techniques to address the language identification problem. In this way, a search for studies that used deep learning for automatic language identification was performed, although is not exclusively related to minority languages. The results are described as follows.

Jaech et al. [9] introduced a hierarchical model for language identification on social networks messages. This model is composed of a Convolutional Neural Network (CNN) followed by a Recurrent Neural Network (RNN) with a bidirectional LSTM layer. The authors tested their model on two Twitter datasets, obtaining 77.1% and 91.2% of average precision respectively.

Bjerva [1] worked on distinguish closer languages using dialects of languages as Spanish, French and Portuguese. The architecture that is proposed uses byte representations on a residual deep network (ResNet). The system developed in this work is called ResIdent, and it was tested on three datasets obtaining 84.88% 68.80% and 69.80% of precision, respectively.

Finally, Kocmi et al. [10] proposed a method for multilingual language identification. If a document is written in two or more languages, the algorithm identifies the sections that belongs to each language and their language labels. This method is based on bidirectional recurrent neural networks (RNN) with Gated Recurrent Unit (GRU). The developed system was trained with a corpus of 131 languages, obtaining 95.5% of average precision.

4 Corpus Development

To build the corpus used in this study, digital documents containing Peruvian indigenous languages texts were retrieved from the web, while other ones were obtained directly from private repositories or books. In that way, it was possible to retrieve documents from 49 different indigenous languages and dialects. It may be considered that these documents must be annotated, which means the target language must be known.

As almost all the documents were in PDF format, the text content was automatically extracted, and some manual corrections were made if it was necessary. Next, a preprocessing program was developed to clean the punctuation and non-alphabet characters from the indigenous languages, to lowercase the text and to split the sentences. After that, Spanish and English sentences were discarded using the resources of a language generic spell-checking library[1], remaining only Peruvian indigenous languages texts.

Table 2 presents information about the textual content, including the total amount of files, the number of sentences/phrases, the number of tokens and the vocabulary size for each Peruvian language and dialect. The preprocessed collection is partially available in a project site, including details of the sources of each language text[2].

The retrieved corpus is analyzed for better description and understanding. Figures 1 and 2 present some statistics regarding the distribution of the total amount of characters per word and per sentence, respectively, in each processed language and dialect.

The first boxplot in Fig. 1 allows the visualization of the rich morphology feature in the Peruvian indigenous languages, as a high number of characters is observed for the word length value in almost all of them. Also, it can be noticed that most of the words are formed by 5 to 10 characters. Nevertheless, there are very large words from *Kakinte* (cot), such as *pimpishivoroquijacoshirentimentanaquempanijite* or *ompitsaraquijacotimentamajatanaquempanijite*, with 46 and 43 characters, respectively. Although, most words from *Kakinte* presents a word-length value between 10 to 20 characters. The terms presented are samples of the reduplication typological feature, which is a common property in some Peruvian linguistic families.

On the other hand, on average, the language with longer words is *Kakinte* (cot), while the language with shorter words is *Kakataibo* (cbr). Moreover, the distribution among languages of the Quechua family is pretty similar.

[1] libenchant: https://github.com/AbiWord/enchant.
[2] chana.inf.pucp.edu.pe/resources/multi-lang-corpus.

Table 2. Retrieved corpus information: $|D|$ = document collection size; $|S|$ = sentences/phrases collection size; $|\mathcal{V}|$ = word vocabulary size, without considering punctuation; $|\mathcal{C}|$ = character vocabulary size; T = number of tokens.

Lang./Dialect	$\|D\|$	$\|S\|$	$\|\mathcal{V}\|$	$\|\mathcal{C}\|$	$\mathcal{T}$
cni	5	22 057	29 350	30	127 342
cjo	3	20 454	26 811	29	129 536
cpu	2	13 198	24 983	31	96 245
cot	3	14 768	32 830	30	130 323
mcb	2	14 259	27 638	30	136 161
not	2	16 835	19 461	28	122 610
ame	2	17 451	24 523	34	190 571
pib	2	14 645	21 508	21	106 284
aym	6	40 179	92 575	38	450 736
boa	3	13 843	19 535	29	123 627
cbt	2	29 569	21 612	28	166 212
huu	1	20 198	9 921	32	149 082
acu	4	24 393	22 718	33	207 439
agr	7	24 849	42 626	32	212 988
hub	1	13 186	20 858	32	139 146
cbu	3	41 746	39 180	32	445 784
amc	1	117	564	24	1 169
kaq	2	14 833	14 241	31	124 927
cbs	4	8 770	17 796	32	149 375
cbr	197	13 397	15 006	37	180 624
mcf	4	21 755	21 803	34	185 367
mcd	2	25 224	21 370	32	228 761
shp	4	11 997	19 645	33	170 226
yaa	1	19 369	15 486	35	161 747
yad	2	11 846	19 828	30	122 554
qva	2	14 050	21 032	38	106 972
quy	2	38 173	69 893	33	389 071
qvc	1	19 263	22 650	28	149 272
qub	1	53 274	58 124	38	360 678
qvh	2	11 347	18 749	38	93 289
quf	9	15 353	22 671	35	141 105
qvm	2	12 042	18 466	38	101 535
qxh	1	11 375	14 052	36	65 992
qvs	1	17 763	18 119	27	133 457
qux	2	11 689	21 703	35	58 491
qwh	5	12 740	26 669	38	111 999
quz	3	1 073	5 754	34	12 891
qve	10	16 505	26 197	33	151 037
qxn	1	11 972	25 202	37	116 058
qvn	2	13 077	24 151	37	111 194
qvz	1	12 837	16 454	32	127 994
qxo	1	14 547	24 815	32	121 475
qup	1	16 585	18 701	30	146 489
qvw	7	9 198	22 058	40	72 932
ura	2	12 928	16 181	26	170 529
ese	1	20 142	6 354	23	211 101
tca	1	14 994	15 496	38	227 054
sey	1	13 634	9 863	33	138 329
arl	4	30 048	26 881	31	263 562

Regarding the sentence length, the longest collected sentences are from *Aymara* (aym) while the shortest are from *Quechua de Yauyos* (qux), as it can be noticed in Fig. 2.

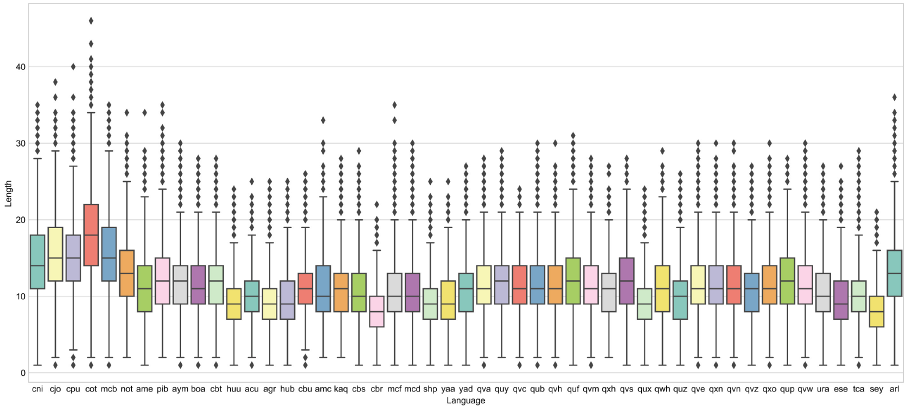


Fig. 1. Boxplots representing the distribution of the word length in number of characters per each language.

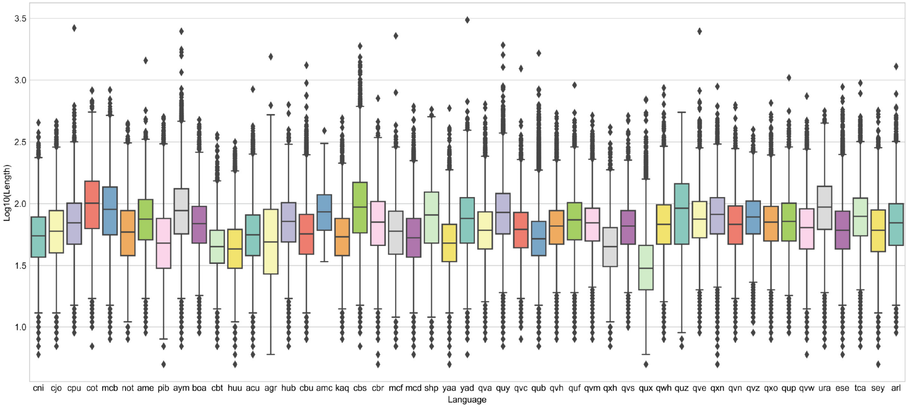


Fig. 2. Boxplots representing the distribution of the sentence length in number of characters per each language. The vertical axe (lenght) is in a log10 scale.

5 Language Identification Model

As it is proposed to perform language identification at the sentence level, the main goal was to learn a classifier or classification function (γ) that maps the sentences from the corpus (S) to a target language class (L):

$$\gamma : S \rightarrow L \tag{1}$$

In order to identify which γ classifier is most suited in the task, two approaches were tested: standard supervised learning and deep learning.

For the evaluation of the model performance, the dataset was split in train and test sub-datasets (70%–30%), although not only randomly and proportionally, but also considering, but also considering out-of-vocabulary (OOV) words for the test sub-set. This means that the test sub-set will have the largest possible number of words that are not in the train sub-set for each language.

5.1 Standard Supervised Learning

For this approach, each sentence $s \in S$ will be represented in a feature vector space model: $s_i = \{w_{1,i}, w_{2,i}, ..., w_{t,i}\}$, where t indicates the number of dimensions or terms to be extracted.

Character-level n-grams was one of the most used language features in the revised works for this task [2,7,12]. Hence, the dimensionality of each vector in the space model will be equal to the number of distinct subsequences of n characters in a given sequence from the corpus S [3].

In this experiment, bigrams, trigrams and 4-grams were used to built the vector space model, and a term frequency - inverse document frequency (TF-IDF) matrix from the aforementioned n-grams scheme was calculated [18].

After that, some classification methods identified in the related works [7] were fit using a 5-fold cross-validation schema on the training sub-dataset. The obtained results are shown in Table 3.

Table 3. Results of the 5-fold cross-validation classification on the train sub-dataset

Method	Precision (%)
SVM (linear kernel)	98.70
Multinomial naive bayes	96.7
SGD classifier	97.06
Perceptron	97.89
Passive aggressive classifier	98.61

As the SVM classifier with a linear kernel got the best precision result, this method was selected as the best standard supervised learning model for this task. Then, this model was validated on the test sub-dataset achieving 95.9% of average precision.

5.2 Deep Supervised Learning

For this approach, the classifier will try to learn and represent its own features. As for the preprocessing step, all the existing characters of the dataset were transformed to numbers. In this way, the sentences are processed as numerical vectors. Additionally, the train set was split in train and validation sub-sets (70%–30%).

Although in other studies different architectures were used, in this work the test of a basic deep learning approach is what is wanted, in order to analyze a baseline performance in the context of scarce data. Therefore, it was implemented an architecture composed by a recurrent neural network (RNN) with a Long-Short Term Memory (LSTM) layer [8]. This model was fit with the training sub-set, and the hyper-parameters were adjusted using the validation set in order to get the best possible model. The architecture of the neural network is presented in Fig. 3. Moreover, the hyper-parameters that were adjusted are the character embedding length, the number of units in the LSTM layer, the batch size and the number of epochs.

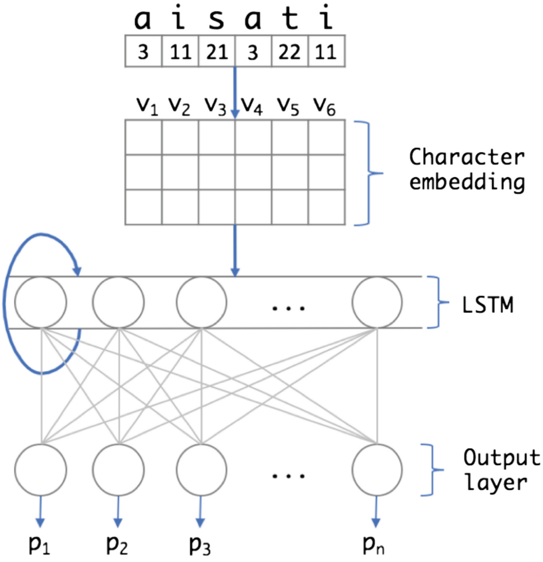


Fig. 3. Architecture of the RNN with a LSTM layer.

Finally, the best hyper-parameters found were an embedding length of 32, 256 units in the LSTM layer, a batch size of 64 and 40 epochs. A model with these hyper-parameters was trained with the entire train sub-dataset and it was validated on the test sub-dataset obtaining 94.8% of average precision.

6 Statistical Comparison Test

This phase is proposed to validate an statistically significant difference between the standard supervised learning model and the deep learning models, and also to know which model performs best in the given scenario.

From the original dataset, 30 sample sets were generated using bootstrap sampling with replacement. Each sample set was split in train and test sub-sets (70%–30%), considering out-of-vocabulary (OOV) words for the test sub-set.

Both models were trained and tested using the sample sets, retrieving the performance in terms of precision for each one. Therefore, two distributions of precision performance were shaped from the 30 sample sets, and the variables for the comparison were defined:

- X: Precision values obtained by the standard supervised learning model.
- Y: Precision values obtained by the deep supervised learning model.

To compare both variables, the first step is to know whether they follow a normal distribution. The Kolmogorov test is required for this evaluation, and the hypothesis for both variables are:

- H_0: The variable follows a normal distribution.
- H_1: The variable does not follow a normal distribution.

In Table 4, the results for the Kolmogorov tests are shown. As for both variables the maximum difference is less than the critical value, then the null hypothesis is accepted, so it can be said that both variables follow a normal distribution.

Table 4. Kolmogorov test for both variables to determine if they follow a normal distribution

	X	Y
Mean	0.950	0.932
Variance	9.75E-08	4.73E-07
Samples	30	
Significance level	0.05	
Maximum difference	0.103	0.107
Critical value	0.242	

As both distributions present a normal shape, an F-test must be evaluated on both variables to know if they have or not similar variances. The F-test hypothesis are as follows:

- H_0: The variances of X and Y are similar.
- H_1: The variances of X and Y are different.

Table 5 presents the results of the F-test. As the P-value (2.88e-05) is lower than the significance level, H_0 is rejected. Thus, it can be said that the variances of X and Y are different.

Finally, a t-student test for distributions with different variance was taken in order to determine which variable has a greater mean. In this way, the hypothesis were defined as follows:

- H_0: The mean of X is lower than the mean of Y
- H_1: The mean of X is greater than the mean of Y

Table 5. F-test for both variables to determine if they have a similar variance

	Results
Significance level	0.05
F	0.206
P(F <= f) one queue	2.88E-05
Critical value for F (one queue)	0.537

In Table 6 the results of the t-student test are shown. The P-value (8.2e-55) is lower than the significance level, H_0 is rejected, and it can be said that the mean of X is greater than the mean of Y. Therefore, the standard supervised learning model has a greater precision than the deep learning approach.

Table 6. t-Student test for both variables to determine which variable have a greater mean

	Results
Significance level	0.05
T	132.626
P(T <= t) one queue	8.2E-55
Critical value for T (one queue)	1.684

7 Results and Discussions

After determining that the standard supervised learning model performs better than the deep learning model, the classification details of the former one (SVM) is shown in Table 7, where the Support column indicates the number of samples that were identified in the test sub-set. In this report, it is shown that the 49 Peruvian languages and dialects used in this work can be distinguishable with 95.9% of precision. This is a promising result for languages that have not previously been worked with. Furthermore, the confusion matrix of this model is presented in Fig. 4 for extended details.

An acceptable overall result was obtained, although there was a great disadvantage to face in the imbalanced corpus, due to the possibility in extracting many more sentences from some languages than from others. For instance, for *Amahuaca* (amc) it was collected only 117 sentences, from which at most 35 ones of them were to part of the test. For that language, perfect precision and recall were obtained. This may indicate an acceptable low-resourced language identification model, but to avoid the possibility of overfitting there must be additional tests when more textual documents can be retrieved.

On the other hand, for closely-related languages like the Quechua family (block 'Q' in Fig. 4), in spite of achieving an acceptable overall precision, a not

Table 7. Detailed classification results for each language (SVM with a linear kernel)

Lang./Dialect	Precision	Recall	F1-score	Support
cni	94.1	87.8	90.9	6617
cjo	95.2	98	96.6	6136
cpu	87.8	98.2	92.7	3959
cot	99.2	96.1	97.6	4430
mcb	97.8	96.5	97.2	4277
not	97.9	97.6	97.8	5050
ame	99.6	99.4	99.5	5235
pib	98.5	98	98.3	4393
aym	99.6	98.4	99	12053
boa	99.6	99.7	99.7	4152
cbt	99.4	99.4	99.4	8870
huu	99.3	99.6	99.4	6059
acu	97.6	95.4	96.5	7317
agr	93.7	92.5	93.1	7454
hub	96	93.5	94.7	3955
cbu	99.4	99.3	99.3	12523
amc	100	100	100	35
kaq	99.4	98.9	99.2	4449
cbs	98.9	99.1	99	2631
cbr	99.3	98.6	99	4019
mcf	99.1	97	98.1	6526
mcd	98.5	99.4	98.9	7567
shp	99.5	94.8	97.1	3599
yaa	99.2	98	98.6	5810
yad	99.4	99.1	99.2	3553
qvw	96.6	96.8	96.7	2759
qva	85.3	87.7	86.5	4215
quy	94.8	98.5	96.6	11451
qvc	99	96.1	97.5	5778
qub	89.3	96.8	92.9	15982
qvh	86.8	83.7	85.2	3404
quf	98.6	94.4	96.5	4605
qvm	87.7	82.3	84.9	3612
qxh	87.7	80.7	84.1	3412
qvs	95.1	93.2	94.1	5328
qux	70.5	91.6	79.7	3506
qwh	97.1	95.7	96.4	3822
quz	81.8	55	65.8	321
qve	95.8	94.1	94.9	4951
qxn	90.3	77.3	83.3	3591
qvn	96	89.1	92.4	3923
qvz	99.2	98.4	98.8	3851
qxo	84.3	84.9	84.6	4364
qup	94.4	95	94.7	4975
ura	99.6	99.5	99.6	3878
ese	99.7	99.7	99.7	6042
tca	99.4	99.7	99.6	4498
sey	99.6	99.5	99.5	4090
arl	99.7	99.6	99.7	9014
Avg/total	95.9	95.8	95.8	262041

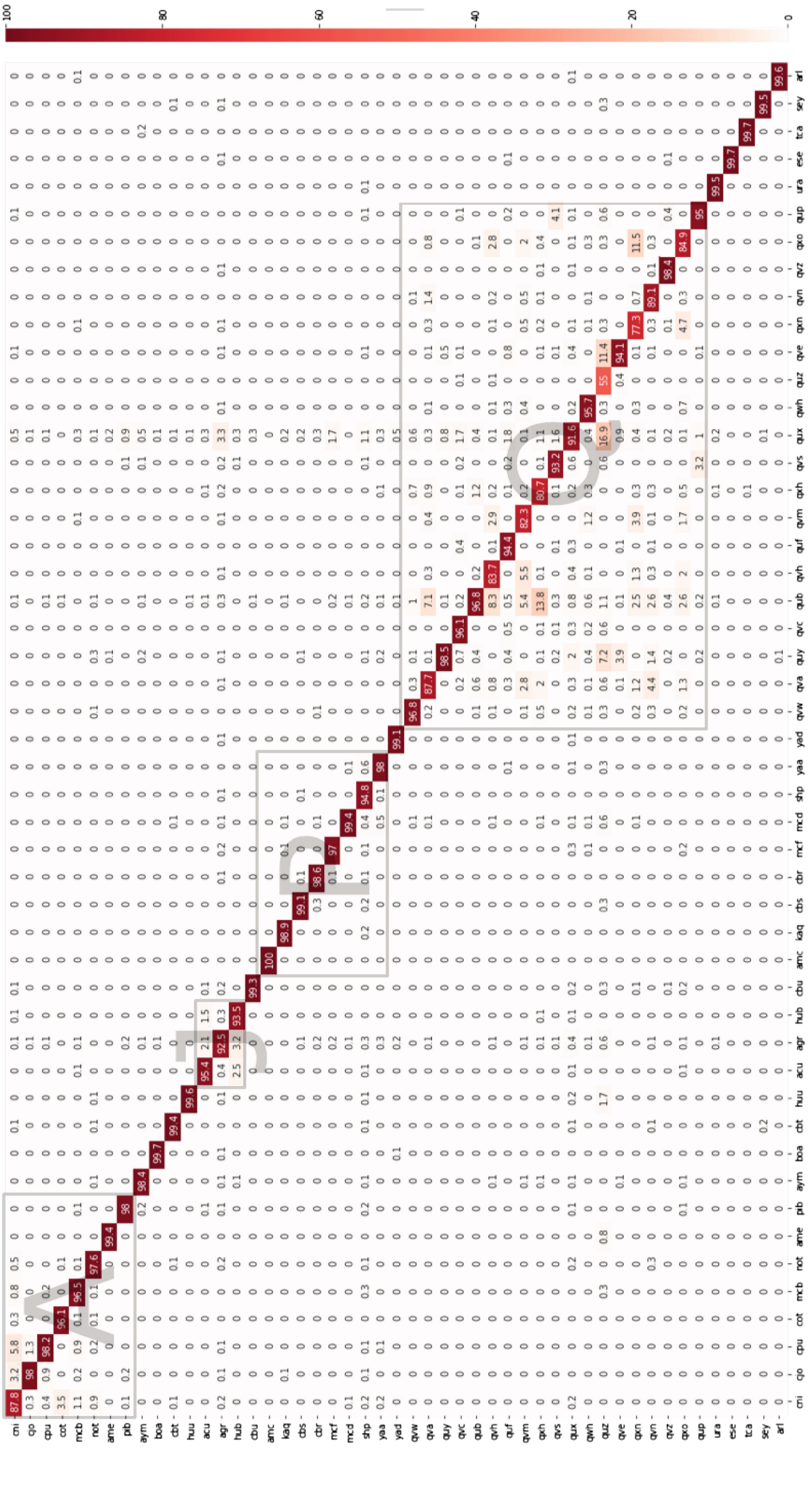


Fig. 4. Confusion matrix obtained by the main language identification model (SVM with lineal kernel). The watermarked letters enhanced the blocks with specific families of languages and dialects (A = Arawak, J = Jíbaro, P = Pano, Q = Quechua).

so good recall value is obtained for those variants with less data. *Quechua del Cuzco* (quz), which is the variety of Quechua with the least amount of extracted sentences, has properly classified only 55% of the test sentences, whereas the model misclassified 16.9% of them as *Quechua de Yauyos* (qux) and 11.4% as *Quechua del Este de Apurímac* (qve). Also, there is a high confusion at discriminating *Quechua de Panao* (qxh) and *Quechua de Huallaga* (qub) since 13.8% of the former language was misidentified as the latter one. Likewise, 11.5% of the sentences of *Quechua del Norte de Conchucos* (qxn) were mislabeled as *Quechua del Sur de Conchucos* (qxo).

In the Arawak family (block 'A' in Fig. 4), the highest ratio of confusion is observed in the three first classes, which are dialects of the Asháninka language. The most affected one is *Asháninka* (cni), which is misclassified 3.2% times as *Asháninka del Pajonal* (cjo), and 5.8% as *Asháninka del Pichis* (cpu).

Regarding other linguistic families, the confusion identified is minimum. All this may indicate the need to go deeper in the representation features used for languages within the same linguistic family, and to greatly consider a hierarchical classifying scheme for future experiments.

8 Conclusions and Future Work

For this study, a corpus for 49 Peruvian indigenous languages and dialects was built through web and private repositories. Besides, it was performed straightforward classification experiments on it, using *n*-grams as features in a tf-idf vector model space and an elemental deep learning architecture. Both approaches were compared through statistical tests, and the best obtained result (95.9% in overall precision in an SVM with lineal kernel) was in the expected range regarding the state-of-the-art for the language identification task in a low-resource scenario.

The fit model may be exploited for other tasks, such as the automatic increasing of the corpus through web and document search [13]. As there are 68 Peruvian indigenous languages and dialects preserved, it is essential to expand the corpus to cover most of them. The Bible will be targeted first, as it is translated in some of the left uncovered languages, and is a very important resource in NLP for minority cases [4].

Also, as the corpus may be growing, other recent methods could be tested on it, such as the bidirectional recurrent neural network proposed in [10] or other similar deep architectures [1,14]. Although it was observed that this kind of algorithms are affected by the low-resourced and unbalanced corpus, those methods could help to decrease the window approach of the classification to a phrase or word-level. Likewise, regarding the confusion presented in languages within the same family, there must be specific considerations in the following experiments with the hierarchical nature in the Peruvian linguistic context [9,11,15].

Finally, it is desired to develop and integrate a way to discriminate languages that are not part of the scheme, in order to not classify out-of-model languages as a Peruvian one. If this latent problem is addressed with positive results, the outcome might offer the possibility to model the origin of a potential new dialect/language among the existing ones.

Acknowledgements. The authors are thankful to J. Rubén Ruiz, bilingual education professor at NOPOKI, for providing access to some private books written in indigenous languages [5,22]. Likewise, it is appreciated the collaboration of Dr. Roberto Zariquiey, linguistic professor at PUCP, for allowing the use of his own corpus for the Panoan family [24].

Furthermore, it is acknowledged the support of the "Concejo Nacional de Ciencia, Tecnología e Innovación Tecnológica" (CONCYTEC Perú) under the contract 225-2015-FONDECYT.

References

1. Bjerva, J.: Byte-based language identification with deep convolutional networks. arXiv preprint arXiv:1609.09004
2. Botha, G.R., Barnard, E.: Factors that affect the accuracy of text-based language identification. Comput. Speech Lang. **26**(5), 307–320 (2012)
3. Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, pp. 161–169 (1994)
4. Christodouloupoulos, C., Steedman, M.: A massively parallel corpus: the bible in 100 languages. Lang. Resour. Eval. **49**(2), 375–395 (2015)
5. Díaz, D.P. (ed.): Relatos de Nopoki. Universidad Católica Sedes Sapientiae (2012)
6. Forcada, M.: Open source machine translation: an opportunity for minor languages. In: Proceedings of the Workshop "Strategies for Developing Machine Translation for Minority Languages", LREC, vol. 6, pp. 1–6. Citeseer (2006)
7. Grothe, L., De Luca, E.W., Nürnberger, A.: A comparative study on language identification methods. In: LREC (2008)
8. Hochreiter, S., Schmidhuber, J.: LSTM can solve hard long time lag problems. In: Advances in Neural Information Processing Systems, pp. 473–479 (1997)
9. Jaech, A., Mulcaire, G., Hathi, S., Ostendorf, M., Smith, N.A.: Hierarchical character-word models for language identification (2016). arXiv preprint arXiv:1608.03030
10. Kocmi, T., Bojar, O.: LanideNN: Multilingual language identification on character window (2017). arXiv preprint arXiv:1701.03338
11. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. Technical report, Stanford InfoLab (1997)
12. Malmasi, S., Dras, M.: Automatic language identification for persian and dari texts. In: Proceedings of PACLING, pp. 59–64 (2015)
13. Martins, B., Silva, M.J.: Language identification in web pages. In: Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 764–768. ACM (2005)
14. Mathur, P., Misra, A., Budur, E.: LIDE: Language identification from text documents (2017). arXiv preprint arXiv:1701.03682
15. McCallum, A., Rosenfeld, R., Mitchell, T.M., Ng, A.Y.: Improving text classification by shrinkage in a hierarchy of classes. In: ICML, vol. 98, pp. 359–367 (1998)
16. Ministerio de Educación, Perú: Documento nacional de lenguas originarias del Perú (2013). http://repositorio.minedu.gob.pe/handle/123456789/3549
17. Pienaar, W., Snyman, D.: Spelling checker-based language identification for the eleven official south african languages. In: Proceedings of the 21st Annual Symposium of Pattern Recognition of SA, Stellenbosch, South Africa, pp. 213–216 (2011)
18. Prager, J.M.: Linguini: Language identification for multilingual documents. J. Manage. Inf. Syst. **16**(3), 71–101 (1999)

19. Quasthoff, U., Richter, M., Biemann, C.: Corpus portal for search in monolingual corpora. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, vol. 17991802, p. 21 (2006)
20. Rios, A.: A Basic Language Technology Toolkit for Quechua (2016)
21. Selamat, A., Akosu, N.: Word-length algorithm for language identification of under-resourced languages. J. King Saud Univ. Comput. Inf. Sci. **28**(4), 457–469 (2016)
22. Universidad Católica Sedes Sapientiae: Relatos Matsigenkas. Universidad Católica Sedes Sapientiae (2015)
23. Valenzuela, P.: Transitivity in shipibo-konibo grammar. Ph.D. thesis, University of Oregon (2003)
24. Zariquiey Biondi, R.: A grammar of Kashibo-Kakataibo. Ph.D. thesis, La Trobe University (2011)

A Multi-modal Data-Set for Systematic Analyses of Linguistic Ambiguities in Situated Contexts

Özge Alaçam(✉), Tobias Staron, and Wolfgang Menzel

Department of Informatics, University of Hamburg, 22527 Hamburg, Germany
alacam@informatik.uni-hamburg.de

Abstract. Human situated language processing involves the interaction of linguistic and visual processing and this cross-modal integration helps to resolve ambiguities and predict what will be revealed next in an unfolding sentence during spoken communication. However, most state-of-the-art parsing approaches rely solely on the language modality. This paper aims to introduce a multi-modal data-set addressing challenging linguistic structures and visual complexities, which state-of-the-art parsers should be able to deal with. It also briefly addresses the multi-modal parsing approach and a proof-of-concept study that shows the contribution of employing visual information during disambiguation.

1 Disambiguation and Structural Prediction

In order to achieve dynamic human-computer interaction, a better understanding of human perceptual and comprehension processes concerning multi-modal environments is one of the crucial factors that need to be taken into consideration. A considerable amount of empirical research in psycholinguistics indicated that human language processing successfully integrates available information acquired from different modalities to resolve linguistic ambiguities (i.e. syntactic, semantic or discourse) and predict what will be revealed next in the unfolding sentence [1–4]. Online disambiguation and prediction processes allow us to have more accurate and fluent conversations during spoken communication. In contrast, state-of-the-art parsing algorithms are still far away from that accuracy when it comes to challenging linguistic or visual situations. Therefore, by developing a multi-modal parser that integrates contextual (*i.e. visual knowledge*), we expect to enhance syntactic disambiguations (e.g. concerning relative clause attachments and scope ambiguities).

Tanenhaus and his colleagues' study [1] showed that visual information influences incremental thematic role disambiguation by narrowing down the possible interpretations. In their study, they focus on one of the most frequently investigated syntactic ambiguity cases in the literature, namely prepositional phrase (PP) attachment ambiguity, where different semantic interpretations are possible depending on assigning different thematic roles. It can be exemplified as *The*

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 106–122, 2018.
https://doi.org/10.1007/978-3-319-90596-9_8

man saw the woman with a telescope, where the PP *with a telescope* can be interpreted as modifier of the seeing action (as instrument), as marked in sentence 1 below, or as a modifier of the object as possessive relation as in Sentence 2. In a multi-modal setting where the scene contains a man holding a telescope in his hand or a woman with a telescope, the visual information constrains the referential choices as well as the possible interpretations, helping the disambiguation process.

1. The man saw [the woman]$_{\text{obj}}$ [with a telescope]$_{\text{instrument}}$
2. The man saw [the woman with a telescope]$_{\text{obj}}$

Further evidence that supports this conclusion was provided by Knoeferle [3] by addressing relatively more complex scenes containing more agents and relations for both English and German. The results also indicate that the influence of visual information on language processing occurs independent from the experiment language. Furthermore, Altmann and Kamide's work [2] has documented that listeners are able to predict complements of a verb based on its selectional constraints. For example, when people hear the verb 'break', their attention is directed only towards breakable objects in the scene. Some nouns may also produce expectations for certain semantic classes of verbs by activating so-called event schema knowledge [5]. Besides verbs and nouns, Berkum and his colleague's study [6] showed the effect of syntactic gender cues for Dutch in the anticipation of the upcoming words. Similar to German, pre-nominal adjectives as well as nouns are gender-marked in Dutch and the gender of the adjective has to agree with the gender of the noun. Their results showed that the human language processing system uses the gender cue, when it becomes available, to predict the target object if its gender is different than the gender of the other objects in the environment. They interpreted this as evidence for the incremental nature of the human language system, which can predict the upcoming words and immediately begin incremental parsing operations. In a more recent work, Coco and Keller [7] investigated the language - vision interaction and how it influences the interpretation of syntactically ambiguous sentences in a simple real-world setting. Their study provided further evidence that visual and linguistic information influences the interpretation of a sentence at different points during online processing. The aforementioned empirical studies provided insights regarding psycho-linguistically plausible parsing. However, those studies were limited to simple (written) linguistic or visual stimuli where object-action relations could be predicted relatively easily.

Based on the prior research, our project focuses on studying underlying mechanisms of human cross-modal language processing of incrementally revealed utterances with accompanying visual scenes, with the aim of using the empirically gained insights to develop a cross-modal and incremental syntactic parser which can be implemented e.g. on a service robot. A parser that processes only linguistic information is expected to be able to successfully handle syntactically unambiguous cases by using linguistic constraints or statistical methods. However, without external information i.e. from visual modality, neither humans nor parsers can resolve references in syntactically ambiguous cases. They may have only preferences. On the other hand, humans naturally use external information from other modalities for disambiguation when available. Incorporating this feature, cross-modal parsers may also resolve those ambiguities and reach correct interpretations in situated contexts. Furthermore, comparing the performance of the computational model with human performance (e.g. whether ambiguities were resolved correctly, at which point of a spoken utterance a correct resolution was achieved, how many changes were made before reaching the correct thematic role assignment) also provides valuable information about the plausibility and the effectiveness of the proposed parsing architecture. Constructing a data-set that contains challenging linguistic and visual cases and complex multi-modal settings, where state-of-the-art parsers often fail, are fundamental towards achieving this ultimate goal. In this paper, we aim to introduce a multi-modal data-set consisting of fully/temporally syntactically ambiguous sentences provided in situated contexts.

This paper is structured as follows. In Sect. 2, a data-set of various ambiguous linguistic structures and their multi-modal representations are presented. A brief description of our multi-modal parser is presented in Sect. 3, which also addresses a proof-of concept study conducted on fully ambiguous sentence structures. Section 4 summarizes the results of this work and draws conclusions.

2 Linguistic Ambiguities in Situated Contexts

Recently, a corpus of language and vision ambiguities (LAVA) in English has been released [8]. The LAVA corpus contains 237 sentences with linguistic ambiguities that can only be disambiguated using external information provided as short videos or static visual images with real world complexity. It addresses a wide range of syntactic ambiguities including prepositional or verb phrase attachments and ambiguities in the interpretation of conjunctions. However, this corpus does not take linguistically challenging cases like relative clause attachments or scope ambiguities, which may also give valuable insights understanding the underlying mechanisms of cross-modal interactions, into account. To our knowledge, the reference resolution concerning these linguistic cases and the effect of linguistic complexity in visually disambiguated situations have been scarcely investigated. But our multi-modal data-set that we are introducing in this section is not limited to investigating these specific ambiguities. It can be also a useful resource in the investigation of a wide range of tasks such as object detection or relation extraction, which are more related with Computer Vision or more Cognitive Science oriented issues like the effect of perceptual/conceptual saliences.

In addition to language-specific investigations, we plan to benefit from a cross-linguistic comparison which opens up novel opportunities for studying the mechanisms of situated language processing in humans by carrying out experiments with similar stimuli in different languages. The base language of the data-set is German and English. Chinese and Turkish counterparts are being prepared as well for cross-lingual comparison. This will allow us to study the influence of different linguistic phenomena on the process of multi-modal sentence comprehension, i.e. syntactic support and constituent ordering. For example, syntactic support is stronger for indogermanic languages like German or English in contrast to Chinese. If the available visual information facilitates syntactic parsing, the effect should be stronger in German than in Chinese. Moreover, constituent ordering can also be studied for instance by comparing the processing of relative clauses preceding their antecedent (Chinese) or following it (German or English). Because these syntactic patterns induce quite different predictions about the characteristics of the target object, they should have a strong impact on the time course of object identification.

Our main question from the psycholinguistic point of view is whether the presence of linguistic ambiguity and the linguistic complexity affect the processing of multi-modal stimuli. On the other hand, from the computational perspective, we focus on whether and to what extent visual information is useful for the disambiguation and structural prediction processes in order to develop more fluent and accurate computational parsing.

German has three grammatical genders, namely each noun is either feminine*(f)*, masculine*(m)*, or neuter*(n)*. In a sentence that contains a relative clause attachment, the gender of the relative pronoun has to be the same as the gender of its antecedent. Sentence-3 illustrates an example, which contains a relative clause licensing the NP.

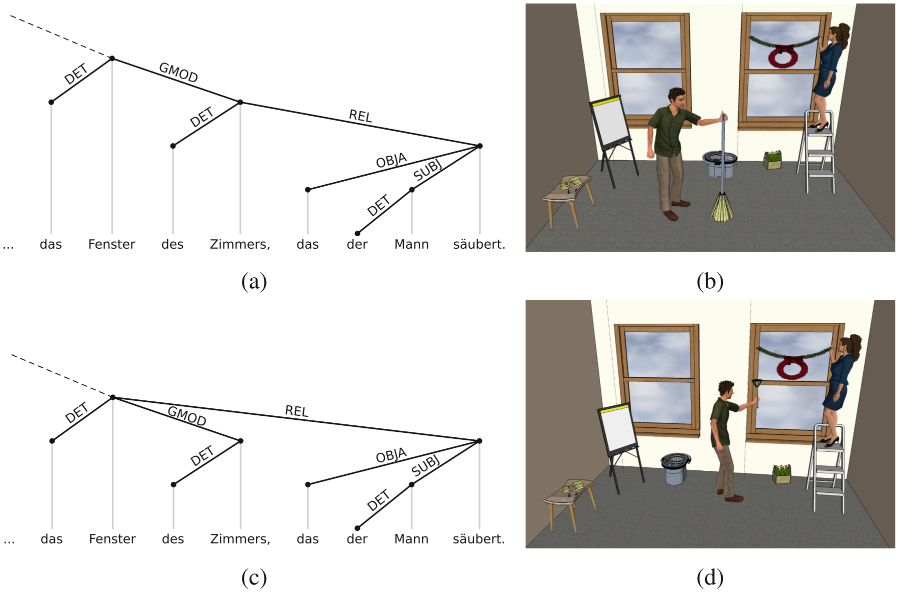
3. Die Frau schmückt das Fenster*(n)*, das*(n)* der Mann säubert.
   The woman decorates the window that the man cleans.

In Sentence-4, the NP is modified by an additional NP, i.e. a genitive object. In this case, since the gender of the relative pronoun matches only the first NP, it is clear that *the window* is being cleaned, not *the car*. However, due to ambiguous German case-marking, if the genders of the nouns of both NPs are the same, as in Sentence-5, both high (*far*) and low (*near*) attachments are possible. Furthermore, the verb is semantically congruent with both NPs as well. In English, since the relative pronoun (*that*) does not have a syntactic marker, both sentences are syntactically ambiguous. Correct reference resolution can not be achieved based on linguistic information alone. On the other hand, having access to visual information eliminates other interpretations and it favors only one assuming there will be no ambiguity in the visual modality (see Fig. 1b and d).

4. Die Frau schmückt das Fenster*(n)* des Wagens*(m)*, das*(n)* der Mann säubert.
   The woman decorates the window of the car that the man cleans.
5. Die Frau schmückt das Fenster*(n)* des Zimmers*(n)*, das*(n)* der Mann säubert.
   The woman decorates the window of the room that the man cleans.

2.1 LASC Data-Set:V1

Our LASC data-set (Linguistic Ambiguities in Situated Context) is currently consisting of 206 situated contexts (*scenarios*) and 599 sentences and addresses 9 linguistically challenging cases (itemized below) concerning relative clause


DET
GMOD
DET
REL
OBJA
SUBJ
DET
... das Fenster des Zimmers, das der Mann säubert.
DET
GMOD
REL
DET
OBJA
SUBJ
DET
... das Fenster des Zimmers, das der Mann säubert.

(a) (b)

(c) (d)

Fig. 1. (a) The first interpretation of syntactically ambiguous Sentence-5: low attachment of relative clause - syntactic gold standard annotation and (b) the corresponding visual scene. (c) the second interpretation of Sentence-5: high attachment of relative clause - syntactic gold standard annotation and (d) the corresponding visual scene.

attachments, Agent/Patient agreement[1], verb/subject agreement, and scope ambiguities for conjunctions and negations. Parsers often have problems with correct reference resolution for such linguistic expressions because they usually attach the relative clause to a nearest option with respect to statistical distributions in their training data or explicitly stated rules. The interpretation of the sentences becomes fully unambiguous in the presence of visual stimuli. The number of scenarios and sentences for each subset is given in Table 1 and the sentences for each structure are generated by using part-of-speech templates exemplified below i.e. for Subset-1A.

- *Template:* $\mathrm{PRO1}_{\mathrm{nom}}$ VP1 $\mathrm{NP1}_{\mathrm{acc}}$ $\mathrm{NP2}_{\mathrm{gen}}$, WDT[2]$_{\mathrm{acc}}$ $\mathrm{PRO2}_{\mathrm{nom}}$ VP2
- *Number of the lexical items:* two pronouns, 48 verbs and 48 nouns

[1] Knoeferle's sentence set [3] was used as baseline since the co-occurrence frequencies between the actions and the Agents in the sentences, as well as between the actions and the Patients, were controlled to single out the effects of semantic associations or preferences during parsing operations. For a syntactic parser, this may seem irrelevant, however, in order to develop a comparable experimental setup for human comprehension, this parameter needs to be taken into account.

[2] Relative Pronoun.

Table 1. The number of scenarios and linguistic structures and the available languages for each subset (*RPA: Relative Pronoun Ambiguity).

	Type	Languages	# of Scenarios	# of Sentences
1	RPA* with a Genitive NP	DE, EN, TR	24	240
2	RP - Scope Amb.	DE, EN	24	24
3	RPA with a Dative PP	DE, EN	48	48
4	RP with an Agent/Patient ambiguity	DE	12	24
5	Negative scope	DE, EN	12	12
6	Agent-Patient agreement	DE	36	72
7	Verb-Subject agreement	DE	6	12
8	Conjunction scope	DE	9	27
9	Constituent (modifier) ordering	DE, EN	35	140
	TOTAL		206	599

Linguistic Structures

Fully Ambiguous Sentence Structures

(1) RPA with a Genitive NP *(DE, EN, TR)*

(A) *Active voice - short sentence.*
(DE) Die Frau schmückt das Fenster*(n)* des Zimmers*(n)*, das*(n)* der Mann säubert.
(EN) The woman decorates the window of the room that the man cleans.
(TR) Adam kadının dekore ettiği odanın penceresini temizliyor.
Int.1[3]*:* The man cleans the room *(low-attachment).*
Int.2: The man cleans the window *(high-attachment).*
(B) *Active voice - long sentence.* The woman with the red hair silently decorates the window of the room that the man cleans in a rush since he needs to go to a meeting soon.
(C) *Passive voice - short sentence.* The woman decorates the window of the room that is cleaned by the man.

(2) RP - Scope Ambiguities *(DE, EN)*

Ich sehe Äpfel*(pl)* und Bananen*(pl)*, die*(pl)* auf dem Tisch liegen.
I see apples and bananas that lie on the table.
Int.1: Only bananas are on the table (low-attachment).
Int.2: Both apples and bananas are on the table.

(3) RPA with a Dative PP *(DE, EN)*

(A) *Syntactically ambiguous.* Da befindet sich ein Becher*(m)* auf einem Tisch*(m)*, den*(m)* sie beschädigt.
It is the mug on the table that she damages.
Int.1: She damages the table *(low-attachment).*
Int.2: She damages the mug *(high-attachment).*

[3] Int.=Interpretation.

(B) *Syntactically unambiguous in German.* Da befindet sich eine Flashe*(f)* auf einem Tisch*(m)*, den*(m)* sie beschädigt.
It is the bottle on the table that she damages. (*She damages the table.)*

(4) RPA with an Agent/Patient ambiguity *(only DE)*
Da ist eine Japanerin*(f)*, die($f, RP_{nom/acc}$) die Putzfrau*(f)* soeben attackiert.
There is a Japanese, who(m) the cleaning lady attacks.
Int.1: The cleaning lady attacks the Japanese woman.
Int.2: The Japanese woman attacks the cleaning lady.

(5) Negative Scope Ambiguities *(DE, EN)*
Die Sängerin kauft die Jacke nicht, weil sie rot ist.
The singer does not buy the coat because it is red.
Int.1: The singer does not buy the coat because of its color.
Int.2: The singer actually buys the coat but not because it is red.

Below, four additional types of temporal ambiguities, which are convenient for the investigation of how/when structural prediction mechanisms are employed during parsing process are presented. It should be noted that, regarding German, all the sentence structures for the fully ambiguous set (except negative scope sentences) presented above can be also transformed to temporally ambiguous sentence structure by changing the noun in either of the NPs (or PPs) with another noun that has an article in different gender.

Temporally Ambiguous Sentence Structures

(6) Agent-Patient Agreement *(only DE)* (following the data-set designed by [3])

- Die Arbeiterin*(f, nom)* kostümiert mal eben den jungen Mann*(m, acc)*.
  The (female) worker just dresses up the young man.
- Die Arbeiterin*(f, acc)* verköstigt mal eben der Astronaut*(m, nom)*.
  The (female) worker is just fed[4] *by the astronaut.*

(7) Verb-Subject Agreement *(only DE)*

- Die Sänger*(f, nom, pl)* waschen*(3rd. Pl.)* den Arzt*(m, acc, sing.)*.
  The singers wash the doctor.
- Die Sänger*(f, acc, pl)* wäscht*(3rd. Sing.)* der Offizier*(m, nom, sing.)*.
  The singers are painted (see footnote 4) *by the officer.*

(8) Conjunction Scope Ambiguities *(only DE)*

- Die Sängerin*(f, nom)* bemalt den Offizier*(m, acc)* und die Ärztin*(f, acc)*.
  The singer paints the (male) officer and the (female) doctor.
- Die Sängerin*(f, nom)* bemalt den Offizier*(m, acc)* und die Ärztin*(f, nom)* wäscht den Radfahrer*(m, acc)*.
  The singer paints the (male) officer and the (female) doctor washes the (male) cyclist.
- Die Sängerin*(f, nom)* bemalt den Offizier*(m, acc)* und die Ärztin*(f, acc)* besprüht der Radfahrer*(m, nom)*.
  The (female) singer paints the (male) officer and the (female) doctor is sprayed (see footnote 4) *by the (male) cyclist.*

[4] The original German sentence is in active voice in OVS word order.

(9) Constituent (Modifier) Ordering *(DE, EN)*

- Bring mir den blauen Becher vom Tresen.
  Bring me the blue mug from the counter.
- Bring mir den Becher, den blauen Becher vom Tresen.
  Bring me the mug, the blue one from the counter.

Image Construction and Semantic Annotations. The multi-modal data-set has been designed to investigate how, when and at which degree does visual complexity affect sentence comprehension and whether visual cues are still strong enough to enforce correct interpretations in such complex linguistic cases.

It should be reminded that for the computational model, we do not need visual scenes, their semantic representations are sufficient, however, the scenes are crucial to conduct comparable experimental studies with human subjects. Furthermore, an automatic extraction of semantic roles from the images is another task that we are aiming for. That is the reason why not just semantic representations but the images themselves are integral part of our multi-modal data-set. The 2D visual scenes were created with the SketchUp Make Software[5] and all 3D objects were exported from the original SketchUp 3D Warehouse. The images were set to 1250 × 840 resolution. Moreover, target objects and agents are located in different parts of the visual scene for each stimulus.

The objects, characters and actions in the images were annotated manually with respect to their semantic roles, similar to McCrae's approach [9], see also [10]. Semantic roles are used to establish a relation between semantic and syntactic levels as an important part of modeling the cross-modal interaction. Semantic roles are linguistic abstractions to distinguish and classify the different functions of the action in an utterance, in other words they are a useful tool to specify *'who did what to whom'*. The most common set of semantic roles includes Agent, Theme, Patient, Instrument, Location, Goal and Path. Figure 2 shows one exemplary semantic annotation for the visual scene displayed in Fig. 1b. There *'die Frau'* is the Agent, who performs the decorating action, *'das Fenster'* is the Patient, the entity undergoing a change of state, caused by the action.

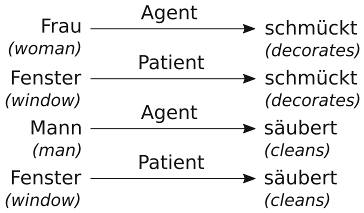


Fig. 2. One exemplary semantic annotation for the visual scene shown in Fig. 1b.

[5] http://www.sketchup.com/ - retrieved on 03.08.2016.

The current version of our LASC multi-modal data-set(v1)[6] that we constructed with the aim of studying disambiguation and structural prediction from both psycholinguistics and computational linguistics perspectives contains following items for each scenario.

- a linguistic form in various languages.
- target interpretation (*gold standard annotations*)
- possible interpretations
- a visual scene of the target interpretation in different visual complexities
- a semantic representation of the scene
- an audio file and a data file with marked onset/offsets (in msec.) of each linguistic entities in the sentence

Visual Complexity. The following figures illustrate how complexity is systematically controlled by giving one subset of the sentences as an example, namely Agent-Patient agreement (Subset-6). In the initial condition, each scenario contains three characters (one Patient, one Agent and one ambiguous Agent/Patient

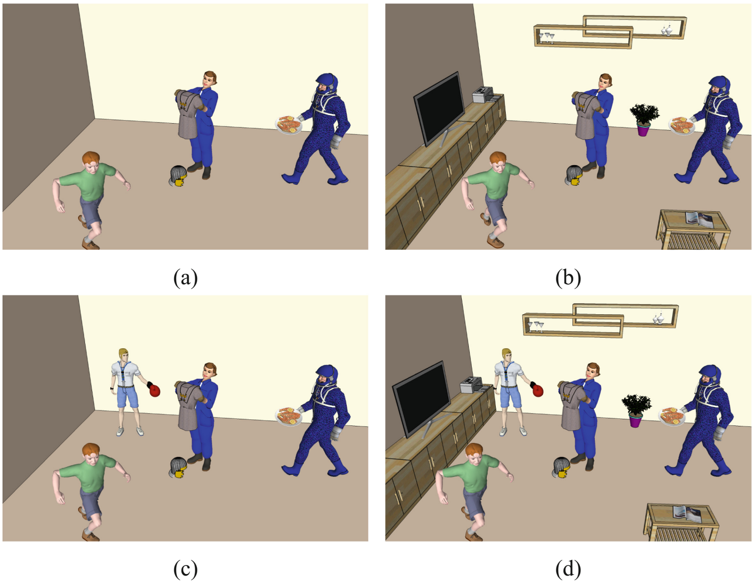
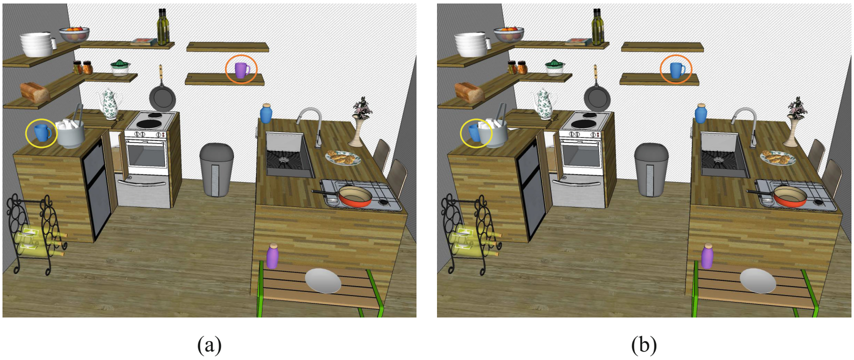

(a) (b) (c) (d)

Fig. 3. (a) 3 agents in an environment with no background objects; *a Patient* (a young boy on the left), *an Agent* (an astronaut on the right) and *an ambiguous Agent/Patient character* (a female worker in the middle) (b) 3 agents in an environment with background objects, (c) 4 agents in an environment without background objects and (d) with background objects.

[6] The data-set can be accessed from https://gitlab.com/natsCML/LASC_v1.

character) and two possible actions. However, the linguistic modality in the data-set addresses only one action and two characters, see sentences in [6]. For each scenario, four different complexity levels were designed. In the first condition, a visual scene contains three characters in an environment without additional background objects, see Fig. 3a. This set-up resemblances Knoeferle's [3] images and provides a baseline to compare our results with previous research. The images in the second condition also contain three characters, but in an environment with noninteracting distractor objects, see Fig. 3b. In the last two conditions, a fourth character in an Agent role, who acts on the ambiguous character is added to the scene. While the images in the third condition do not have additional objects, the images in the fourth condition are in a cluttered environment as in the condition 2 (see Fig. 3c and d). It should be noted that background objects and the fourth character do not have any semantic association with the actions mentioned in the sentences. Besides, visual complexities can be further diversified, e.g. by adding another Patient character to the scene or by adding semantically congruent distractor objects.

Another example of visual complexities represented in the data-set can be given regarding the subset-9 that addresses the effect of constituent ordering. The variations in the scenes of this subset were tailored to investigate how the processing of the linguistic description of the target object (i.e. *the blue mug*) is influenced by the number of the distractor objects which can also be described with the same modifier (i.e. *blue*) or by other visual constraints like saliency and occlusion (i.e. Fig. 4b).


(a)

(b)

Fig. 4. A visual scene for the sentence given as an example in Subset-9. In a simple case (a), there is only one blue mug. In a more complex scenario (b) there are two blue mugs, but the target is occluded. (Color figure online)

3 Multi-modal Parsing

As suggested by the literature discussed in Sect. 1, multi-modal integration is a key to resolve linguistic ambiguities and to anticipate what will be revealed next in an unfolding sentence. However, most state-of-the-art parsing approaches rely solely on the language modality. One of the first systems that integrates contextual knowledge into a grammar-based parser to resolve ambiguities in German (e.g. addressing Genitive-Dative ambiguities of nouns with feminine case markings or PP attachment ambiguities) was proposed by McCrae [11]. Based on this study, Baumgärtner et al. [12] successfully realized a system integrating visual context to improve the processing of German sentences in an incremental manner leading to the only parser so far that is both incremental and cross-modal.

For the proof-of-concept study reported in the following subsection, a parser that utilizes a simple two-stage filtering approach has been developed. In contrast to the rule-based parsers [11,12], we employ a hybrid parsing approach consisting of both data-driven and grammar-based components with the aim to develop a language-independent parser that achieves state-of-the-art results in resolving linguistic ambiguities. The data-driven parser, which searches for the most plausible disambiguation of a given sentence among all possible dependency trees, is interfaced with a grammar-based component, which evaluates the most probable candidates with respect to information derived from visual input, to narrow down the hypotheses towards the most plausible representation for the sentence at hand with respect to given context.

The semantic role annotations of images elaborated in Sect. 2.1 serve as representations of contextual information. The contexts are assumed to be dynamic, i.e. the environments depicted in the visual stimuli may constantly change. So, unseen examples are not guaranteed to be represented by the training instances. Thus, additional modalities, which provide contextual information, cannot be incorporated by extracting features and adding them to the ones derived from the input sentence [17], or by learning a separate model. I the contextual information contains previously unseen information, it might not improve results or might even deteriorate parsing performances. Therefore, instead of using data-driven models for processing the context, we developed a grammar that links semantic roles and the corresponding syntactic structures. The constraints of this linking grammar work with the available information, e.g. Part-of-Speech (PoS) tags, dependency relations or syntactic labels, except for the word forms. Thus, the grammar is not lexicalized and independent of actual words. If there is a match between a semantic role and a syntactic structure, e.g. the Patient of the relative clause verb coincide with one of the possible attachments, then one linking constraint enforces that the relative clause has to be attached to that Patient. This way, it is ensured that the relative clause in Sentence-5 is attached to the *window* as the Patient of the *cleaning* action, and not to the *room*. Instead of developing a full grammar that covers all relations between every semantic role and the syntactic level, in the current version of the data-set, the content of our grammar is limited to the cases relevant with respect to the mentioned actions, agents and objects in the sentence sets. Since the images and the sentences are publicly available, any other relations with respect to user's particular interests can be extracted.

Figure 5 summarizes the parsing process. First, the data-driven RBGParser (RBG) generates the k-best candidates S_1, ..., S_k for the input sentence $x_{linguistic}$. RBG is a language independent, data-driven dependency parser and achieves state-of-the-art results [18,19]. Hill climbing is applied to approximate the optimal dependency tree, i.e. the Maximum Spanning Tree, except for cases an edge-factored model, which consists of first-order features only, is used. Then, the optimal dependency tree is determined by the Chu-Liu-Edmonds (CLE) algorithm [20]. Since we are interested in non-canonical interpretations (namely ambiguous sentences), which are usually not well represented in the training data, the best chosen parse will likely not represent the correct parse, but previous research has shown that the desired parse can more often be recovered if the k-best parses are taken into account [16]. While edge-factored models perform worse compared to higher-order models regarding 1-best parsing, their k-best results are sufficient as input for a second-stage parser, of which model can be arbitrarily complex [21]. Finding the k-best Spanning Trees based on an edge-factored model can be achieved by the algorithm of Camerini et al. [15], which extends the CLE algorithm.

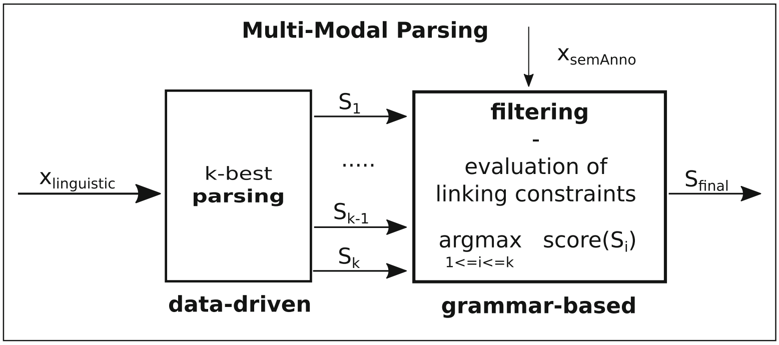


Fig. 5. Multi-modal parsing architecture - The data-driven component performs k-best parsing for input sentence $x_{linguistic}$. The grammar-based component filters all k candidates S_1, ..., S_k by evaluating constraints, which link semantic roles and the corresponding syntactic structures, with respect to the semantic annotation $x_{semAnno}$ of the visual input, the contextual information, to find the most probable solution S_{final}.

Next, all semantic role annotations $x_{semAnno}$ of the corresponding image, in other words the contextual information, are mapped onto the corresponding words of $x_{linguistic}$. For example, the *man* or the *window* from Fig. 2 are mapped onto the respective words of Sentence-5. The scoring mechanism of the grammar-based jwcdg [14], re-evaluates all candidate parses with respect to the semantic annotations using our linking grammar and excludes all candidates that violate constraints. The best remaining candidate S_{final} is the most probable solution. In case, no solution fits the contextual information, the 1-best parse of RBG is taken as a fallback solution. To the best of our knowledge, there exists no comparable system for multi-modal broad-coverage syntactic parsing yet.

3.1 A Proof-of-Concept Study

A proof-of-concept study was conducted in order to evaluate the multi-modal approach to parsing presented in this section and to compare it to the unimodal RBG to show that a simple filtering approach is already sufficient to improve parsing results in case contextual information is provided. The study covered four subsets of fully ambiguous sentences from Sect. 2: (i) RPA with a Genitive NP *(Subset-1A, DE and EN)*[7], (ii) RP - Scope Ambiguities *(Subset-2)*, (iii) RPA with a Dative PP *(Subset-3A)* and (iv) RP with an Agent/Patient ambiguity *(Subset-4)*. For each type of ambiguity, 24 test sentences were chosen. For each sentence, the corresponding contextual information, i.e. the visual stimulus, was manually annotated as described in Subsect. 2.1. The entire test was conducted with German sentences.

In Subset (1) to (3), there are two possible antecedents for each relative clause. On the syntactic level, the relative pronoun agrees with both possible antecedents, i.e. in case, gender and number. On the semantic level, both possibilities could fill the same semantic role. In Subset (4), either the relative pronoun is the Subject and the subsequent NP is the Object or vice versa (due to the free word order of the German language) and both can fill the respective semantic roles. The first interpretation, e.g. low attachment, is the target hypothesis in one half of the test sentences and the second interpretation, e.g. the high attachment, in the other half.

Two RBG models have been trained: a full model, which exploits up to third-order local features as well as global features, and an edge-factored model. Both models have been trained on the first ≈ 98k sentences (without duplicates) of the Hamburg Dependency Treebank (HDT) part A [22]. The dependency relations of all sentences, taken from the German news website Heise Online[8], are manually annotated and the data includes word forms, gold PoS tags (from the Stuttgart-Tübingen Tagset [23]), and gold standard annotations. Instead of using the gold standard PoS tags, TurboTagger [24] is used to predict them. For tagging the training sentences, ten-way jackknifing was performed: the training set is split into ten partitions and each partition is tagged by a model trained on the other nine partitions. The test sentences were tagged by models trained on the entire training set of the respective corpus.

Table 2 shows the results for disambiguating the different types of ambiguities for both the unimodal RBG and the multi-modal filtering approach. For (1) RPA - a Genitive NP (A.), the unimodal RBG always chooses the low attachment of the relative clause, as expected due to the respective statistical distribution in the training data. Thus, it is not able to attach relative clauses

[7] See [25] for a study focused more on the experiments on this Subset regarding all three languages: German, English and Turkish.

[8] https://www.heise.de.

Table 2. The resolution of the ambiguities for the four different types of linguistic ambiguities for unimodal parsing compared to multi-modal parsing.

	low	high	other
low attachment	12	0	0
high attachment	12	0	0

(a) RPA with a Genitive NP unimodal RBG

	low	high	other
low attachment	12	0	0
high attachment	0	12	0

(b) RPA with a Genitive NP multi-modal RBG

	low	high	other
low attachment	12	0	0
high attachment	12	0	0

(c) RP - Scope Ambiguities unimodal RBG

	low	high	other
low attachment	12	0	0
high attachment	0	12	0

(d) RP - Scope Ambiguities multi-modal RBG

	low	high	other
low attachment	11	0	1
high attachment	11	0	1

(e) RPA with a Dative PP unimodal RBG

	low	high	other
low attachment	12	0	0
high attachment	0	12	0

(f) RPA with a Dative PP multi-modal RBG

	s-o	s-s	o-s	o-o	s-other
subject-object	5	6	0	0	1
object-subject	5	6	0	0	1

(g) RP with an Agent/Patient ambiguity unimodal RBG

	s-o	s-s	o-s	o-o	s-other
subject-object	11	0	0	0	1
object-subject	0	10	0	0	2

(h) RP with an Agent/Patient ambiguity multi-modal RBG

correctly in case the high attachment is part of the target hypothesis. In contrast, the multi-modal filtering approach attaches all relative-clauses accordingly with respect to the target hypothesis by utilizing the contextual, i.e. visual, information as described in this section. The same behavior has been observed for (2) RPA - Scope Ambiguities and (3) RPA - a Dative PP (A.). For the latter type, the original, unimodal RBG has attached one supposed low and one supposed high attachment to completely different antecedents. The multi-modal approach avoids these errors.

For (4) RPA with an Agent/Patient ambiguity, the unimodal RBG attached all Agents and Patients respectively subjects and objects correctly, but with correct syntactic labels in only 5 out of 24 cases. Mainly, it is not able to recognize the relative pronoun as object and not able to assign the correct label to the NP following the relative pronoun either. On the other hand, the multi-modal filtering approach improve these results by labeling 11 cases correctly by assigning the correct label to the NP following the relative pronoun in 21 cases. The problem of not recognizing the relative pronoun as object remains. This problem may be overcome by more sophisticated approaches to multi-modal parsing instead of using our simple filtering-based approach.

4 Discussion

Employing a parser that mimics some of the important mechanisms of natural language processing such as prediction and disambiguation is crucial for enabling more fluent and dynamic spoken communication. Which linguistic entity resolves the ambiguities in systematically controlled situated contexts gives us valuable information about the underlying mechanism of human language-vision interaction. In addition to reach an understanding in two endeavors, namely the cognitive aspects of language processing and technical aspects of parsing technology, a multi-modal data-set that pertains challenging ambiguous cases for both areas in a systematic way needs to be designed carefully. This paper addresses this bridging component.

Here we introduce a multi-modal set (see footnote 6) for temporally or fully ambiguous sentences in various languages addressing 9 different linguistic structures and different visual complexities. Furthermore, the contribution of the external information in parsing operations was shown by a proof-of concept study. Further studies will address the comparison between the performance of human subjects and of computational model regarding both disambiguation and structural predictions tasks.

Acknowledgments. This research was funded by the German Research Foundation (DFG) in project 'Crossmodal Learning', TRR-169.

References

1. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C.: Integration of visual and linguistic information in spoken language comprehension. Science **268**(5217), 1632 (1995)
2. Altmann, G.T., Kamide, Y.: Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition **73**(3), 247–264 (1999)
3. Knoeferle, P.S.: The role of visual scenes in spoken language comprehension: evidence from eye-tracking. Ph.D. thesis, Universitätsbibliothek (2005)
4. Ferreira, F., Foucart, A., Engelhardt, P.E.: Language processing in the visual world: effects of preview, visual complexity, and prediction. J. Mem. Lang. **69**(3), 165–182 (2013)
5. McRae, K., Hare, M., Ferretti, T., Elman, J.L.: Activating verbs from typical agents, patients, instruments, and locations via event schemas. In: Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society, Erlbaum Mahwah, NJ, pp. 617–622 (2001)
6. Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P.: Anticipating upcoming words in discourse: evidence from ERPs and reading times. J. Exp. Psychol. Learn. Mem. Cogn. **31**(3), 443 (2005)
7. Coco, M.I., Keller, F.: The interaction of visual and linguistic saliency during syntactic ambiguity resolution. Q. J. Exp. Psychol. **68**(1), 46–74 (2015)

8. Berzak, Y., Barbu, A., Harari, D., Katz, B., Ullman, S.: Do you see what I mean? Visual resolution of linguistic ambiguities. arXiv preprint arXiv:1603.08079 (2016)
9. McCrae, P.: A computational model for the influence of cross-modal context upon syntactic parsing (2010)
10. Mayberry, M.R., Crocker, M.W., Knoeferle, P.: A connectionist model of the coordinated interplay of scene, utterance, and world knowledge. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society, pp. 567–572 (2006)
11. McCrae, P.: A model for the cross-modal influence of visual context upon language processing. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2009), Borovets, Bulgaria, pp. 230–235 (2009)
12. Baumgärtner, C., Beuck, N., Menzel, W.: An architecture for incremental information fusion of cross-modal representations. In: IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Hamburg, Germany, pp. 498–503. IEEE (2012)
13. Beuck, N., Köhn, A., Menzel, W.: Incremental parsing and the evaluation of partial dependency analyses. In: DepLing 2011, Proceedings of the 1st International Conference on Dependency Linguistics (2011)
14. Beuck, N., Köhn, A., Menzel, W.: Predictive incremental parsing and its evaluation. In: Computational Dependency Theory. Frontiers in Artificial Intelligence and Applications, vol. 258, pp. 186–206. IOS Press (2013)
15. Camerini, P.M., Fratta, L., Maffioli, F.: The k best spanning arborescences of a network. Networks **10**(2), 91–109 (1980)
16. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 173–180. Association for Computational Linguistics, June 2005
17. Salama, A.R., Menzel, W.: Multimodal graph-based dependency parsing of natural language. In: Hassanien, A.E., Shaalan, K., Gaber, T., Azar, A.T., Tolba, M.F. (eds.) AISI 2016. AISC, vol. 533, pp. 22–31. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-48308-5_3
18. Zhang, Y., Lei, T., Barzilay, R., Jaakkola T., Globerson, A.: Steps to excellence: simple inference with refined scoring of dependency trees. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, pp. 197–207. Association for Computational Linguistics (2014)
19. Lei, T., Xin, Y., Zhang, Y., Barzilay, R., Jaakkola, T.: Low-rank tensors for scoring dependency structures. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, pp. 1381–1391. Association for Computational Linguistics, June 2014
20. Tarjan, R.E.: Finding optimum branchings. Networks **7**(1), 25–35 (1977)
21. Hall, K.: k-best spanning tree parsing. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp. 392–399 (2007)
22. Foth, K.A., Köhn, A., Beuck, N., Menzel, W.: Because size does matter: the Hamburg dependency treebank. In: Proceedings of the Language Resources and Evaluation Conference 2014, LREC, European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
23. Schiller, A., Teufel, S., Thielen, C.: Guidelines für das tagging deutscher textcorpora mit STTS. Universität Stuttgart und Universität Tübingen (1995)

24. Martins, A.F.T., Almeida, M.B., Smith, N.A.: Turning on the turbo: fast third-order non-projective turbo parsers. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 617–622 (2013)
25. Staron, T., Alacam, O., Menzel, W.: Incorporating contextual information for language-independent, dynamic disambiguation tasks. In: Proceedings of the 11th Language Resources and Evaluation Conference (LREC) (2018)

Community Detection in Bipartite Network: A Modified Coarsening Approach

Alan Valejo(✉), Vinícius Ferreira, Maria C. F. de Oliveira,
and Alneu de Andrade Lopes

Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo
(USP), São Carlos, SP 14560-970, Brazil
alanvalejo@gmail.com, viniciusferreira97@gmail.com,
{cristina,alneu}@icmc.usp.br

Abstract. Interest in algorithms for community detection in networked systems has increased over the last decade, mostly motivated by a search for scalable solutions capable of handling large-scale networks. Multilevel approaches provide a potential solution to scalability, as they reduce the cost of a community detection algorithm by applying it to a coarsened version of the original network. The solution obtained in the small-scale network is then projected back to the original large-scale model to obtain the desired solution. However, standard multilevel methods are not directly applicable to bipartite networks and there is a gap in existing literature on multilevel optimization applied to such networks. This article addresses this gap and introduces a novel multilevel method based on one-mode projection that allows executing traditional multilevel methods in bipartite network models. The approach has been validated with an algorithm for community detection that solves the Barber's modularity problem. We show it can scale a target algorithm to handling larger networks, whilst preserving solution accuracy.

1 Introduction

Complex networks are relational structures that represent many real-world systems formed by a large number of highly interconnected dynamical units. Many such systems exhibit a natural bipartite (or two-layer) structure, in which the set of units (known as vertices) is split into two disjoint subsets (or layers) and connections (known as edges) are established between units placed in different layers. Document-word [25], protein-ligand [14] and actor-movie [36] networks are a few examples of real-world bipartite networks.

Many such networks present community structures. They are defined as groups of vertices densely connected to each other within a group and sparsely connected to vertices in other groups. Such structures are important for characterizing the network behavior, as vertices that belong to the same community may share common properties or play similar roles in a networked system.

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 123–136, 2018.
https://doi.org/10.1007/978-3-319-90596-9_9

Therefore, the identification of community structures in networked systems contributes to a better understanding of their topological structure and dynamical processes [12]. For instance, communities in a biological protein network typically correspond to proteins that share a single specific function [20]. Furthermore, the increasing interest in identifying community structures in bipartite networks [2,5,9,10,17,29] is a strong indicator of this being a relevant research topic.

Community detection algorithms aim at subdividing a set of vertices into k communities while minimizing the number of edges connecting vertices placed in different communities [12,32]. This is a hard combinatorial optimization problem, in which the goal is to optimize a given cost function, such as modularity [13]. Searching for an optimal solution on large-scale networks is unfeasible, since the number of possible solutions can be exponential.

To overcome this problem, researchers resorted to multilevel approaches, in which: an original network is gradually coarsened at multiple levels by joining vertices and edges (coarsening phase); an initial community structure is obtained on the coarsest network (solution finding phase); and the starting solution is successively projected back over the inverse sequence of coarsened networks, until the original network (uncoarsening phase).

Many multilevel community detection algorithms have been developed for unipartite (or one-mode) networks. Some studies introduced multilevel community detection methods targeted at specific types of networks; e.g., Abou-Rjeili and Karypis [1] considered networks with a power-law degree distribution, and Valejo et al. [34,35] explored properties of social networks, such as high transitivity and assortativity. Other contributions focused on applying multilevel optimization to improve a measure of modularity [7,8,19,22,26,27,37]. Furthermore, many authors investigated parallel paradigms to speed up the coarsening and refinement phases [3,4,11,18,19,28,30,31].

Nonetheless, the aforementioned approaches cannot be directly applied to bipartite networks, since standard coarsening algorithms operate by matching and joining pairs of connected vertices. Such solutions presume all vertices are alike, i.e., represent the same type of entity. However, vertices in different layers of bipartite networks are not connected, and typically represent entities of different types. Coarsening such networks requires a different treatment of the vertices, as only vertices of the same type (i.e., in the same layer, and thus unconnected), should be matched.

In a previous study [33] we addressed this gap and introduced a novel one-mode projection-based multilevel method that enables applying standard coarsening algorithms to bipartite networks. In this paper, we extend our previous study in the following manner: **(1)** we describe the proposed model in further detail; **(2)** we evaluate the application of another popular matching algorithm in this context, now contemplating an analysis of three matching algorithms; **(3)** we extend the statistical analysis of the quality of the results yielded by the proposed multilevel solution; **(4)** we expand the analysis of the scalability of the multilevel solution. Our empirical investigation on an expressive set of synthetic network models has shown that our multilevel solution can be combined with a community detection algorithm, yielding expressive speedups with no significant loss in solution quality.

The remainder of the paper is organized as follows: Sect. 2 reviews some basic concepts on networks and provides a brief overview of standard multilevel approaches; Sect. 3 introduces the proposed multilevel formulation for bipartite networks and its implementation; Sect. 4 presents results from an empirical study on a large synthetic test suite; finally, Sect. 5 summarizes the results and discusses future work.

2 Fundamentals

This section describes the terminology and fundamental concepts required to understand the proposed solution.

2.1 Basic Definitions

A unipartite network is given by $G(V, E, \omega)$, where $V = \{v_1, v_2, ..., v_n\}$ is the set of vertices, $E = \{e_1, e_2, ..., e_k\}$ is the set of edges connecting vertices, such that $e_i = (v, u) = \{(u, v) = (v, u) \mid u, v \in V\}$ and $\omega = \{w_1, w_2, ..., w_k\}$ is the set of weights, so that each $w_i \in \mathbb{R}$ is associated with a corresponding edge e_i. Two vertices are said to be neighbors if they are connected by at least one edge.

A bipartite network is given by $G(V, E, \omega)$, where V is partitioned into two sets V_1 and V_2 so that $V_1 \cap V_2 = \emptyset$, $V_1 = \{u_1, u_2, \ldots, u_n\}$ is a set (or layer) of vertices, $V_2 = \{v_1, v_2, \ldots, v_m\}$ is another set of vertices and $E = \{e_1, e_2, ..., e_k\}$ is the set of edges connecting vertices in different layers, i.e. for all $(u, v) \in E$, $u \in V_1$ and $v \in V_2$ and $E \subseteq V_1 \times V_2$. Similarly, $\omega = \{w_1, w_2, ..., w_k\}$ is the set of edge weights. Figure 1 illustrates a bipartite network.

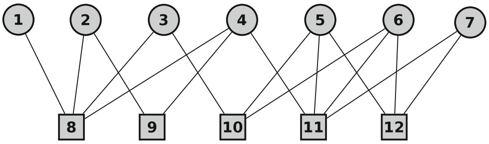
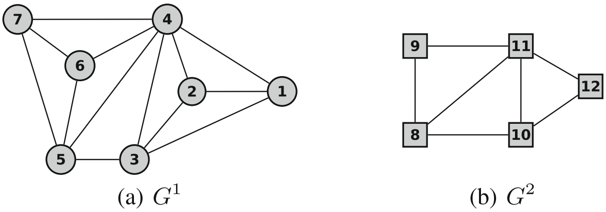


Fig. 1. A bipartite network.

Bipartite networks can be transformed into unipartite networks through a one-mode projection [21,23,24]. Applying a one-mode projection to a bipartite network generates two unipartite networks, one for each layer, G^1 and G^2, so that vertices with common neighbors are connected by edges in their respective projection. Figure 2 illustrates the result of applying such a procedure to the simple bipartite network shown in Fig. 1.



(a) G^1 (b) G^2

Fig. 2. Unipartite networks G^1 and G^2 resulting from a one-mode projection of the bipartite network in Fig. 1.

Figures 2a and b show, respectively, the one-mode projections of V_1 (i.e., G^1) and V_2 (i.e., G^2). If two vertices share multiple common neighbors, their connection in the unipartite projection should reflect this topology. In weighted unipartite projections, edge weights are assigned according to the number of common neighbors between the two connected vertices, as illustrated in Fig. 3 for the particular pair of projected vertices $\{4, 5\}$.

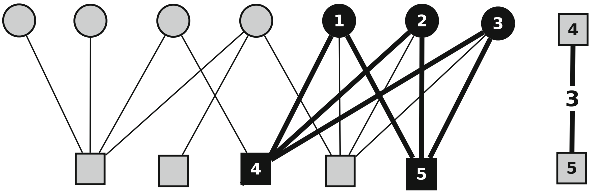


Fig. 3. A weighted projection: vertices 4 and 5 are connected by an edge of weight 3, as they share three neighbors in common.

2.2 Multilevel Optimization

Multilevel optimization reduces the number of operations required to solve a target problem, i.e., the multilevel strategy can execute a complex optimization algorithm, that can not be executed on a very large network, on a coarsened (also called reduced) version of this network which requires a much smaller number of operations. The results obtained in the reduced network are then projected back to obtain the actual solution for the original network.

Let us consider an initial unipartite network $G_0(V_0, E_0, \omega_0)$ and assume its size (in terms of edges and vertices) prevents the execution of a target algorithm. A multilevel approach could be applied as follows [15]:

Coarsening phase: Original network G_0 is transformed into a sequence of *coarsened networks* $G_1, G_2, \cdots, G_m$ (*hierarchy of coarsening*), wherein G_m is the *coarsest network*. The size of the vertex set of each subsequent network is reduced at each iteration, i.e., $|V_0| > |V_1| > |V_2| > ... > |V_m|$.

Solution finding phase: The *target algorithm* is applied to coarsest network G_m, thus generating a *starting solution*. As $|V_m|$ is sufficiently small, the target algorithm can be executed in feasible time. In the present study, we consider a community detection algorithm as the target.

Uncoarsening phase: The starting solution obtained in the network G_m is projected back, through the intermediate levels $G_{m-1}, G_{m-2}, \cdots, G_1$, until the original network G_0, thus generating the *final solution*.

The coarsening phase is an iterative process that constructs a sequence of reduced versions of the initial network G_0. The vertices of a network G_i are joined into super-vertices to obtain a network G_{i+1}. Edges incident to the original vertices are joined to obtain the edges incident to a super-vertex. The coarsening process is split into two phases, namely *matching* and *contracting*.

In the matching phase, edges, or vertex pairs, are selected to be joined. Once an edge in G_i has been selected, its incident vertices are candidates to form a super-vertex. Any vertex from G_i with no incident edge selected is inherited by G_{i+1}. In the present study, we consider three well-known matching methods, introduced by Karypis [15], namely:

Random matching (RM): Vertices are visited in random order. If a vertex v has not been matched yet, one of its unmatched neighbors is selected. If such a vertex u exists, the pair (v, u) is included in the matching set, otherwise v remains unmatched. Although it may yield poor results, RM is very fast and straightforward.

Heavy edge matching (HEM): Similarly to RM, vertices are also visited in random order. However, unlike RM, vertices v and u are matched if edge (v, u) has maximum weight over all valid edges incident to v. Although HEM does not guarantee the matching obtained has maximum weight, it yields better results than RM with equivalent asymptotic complexity.

Greedy Heavy Edge Matching (GHEM): Edges are sorted by descending weight and then visited iteratively. For each visited edge (u, v), if neither u nor v is matched yet, the pair $\{u, v\}$ is included in the matching. The algorithm is near linear time and although slightly more expensive than RM and HEM, it is more robust in practice.

Next, the contracting algorithm builds the coarsened network directly from the matching by joining each pair of matched vertices into a single *super-vertex* (sV). Edges incident to sV, called super-edges, are obtained by joining the edges incident to vertices $\{u, v\} \in V_i$. The weight of the super-edge is given by the sum of the weights of all edges incident to $\{u, v\} \in V_i$.

The target algorithm (community detection, in this case) is then evaluated in the coarsest network G_m to obtain an initial solution. As $|V_M| < |V_0|$, the algorithm converges faster and can generate an initial solution in feasible time.

In the uncoarsening phase, the initial solution is successively projected back to G_0. At each level, each super-vertex $s_v = \{u, v\} \in V_{i+1}$ is expanded to its original vertices in V_i, i.e. u and v, and the solution is projected through the intermediate levels $G_{m-1}, G_{m-2}, ..., G_0$. For each decomposed $s_v \in V_{i+}$, its original vertices $\{u, v\} \in V_i$ are assigned to the same community as their parent $s_v \in G_i$. Figure 4 illustrates this process: super-vertex $s_v = \{4, 5\}$ is expanded to its original vertices 4 and 5, which are assigned to the same community as s_v.

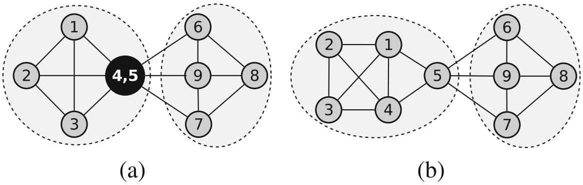


Fig. 4. Super-vertex s_v formed by vertices 4 and 5 is expanded and each single vertex is assigned to the same community as s_v.

3 Community Detection Based on Multilevel Approach by Using One-Mode Projection

This section introduces a multilevel community detection method that handles bipartite networks. Standard multilevel methods do not distinguish between vertex types, whereas the layers in bipartite networks usually represent different entities and should be handled independently. Therefore, usual matching algorithms, such as RM, HEM and GHEM are not directly applicable. Nonetheless, they can be applied to a one-mode projection of G (i.e., G^1 and G^2), since the projection produces two one-mode graphs, in which vertices are all of the same type. Therefore, one-mode projection methods enable applying any standard coarsening algorithm to bipartite networks after a transformation process. This concept was considered to introduce a multilevel community detection method applicable to bipartite networks.

Algorithm 3.1 summarizes the implementation of the proposed ***o****ne-mode* ***p****rojection-based* ***m****ultilevel method for community detection* (OPM). It comprises the phases of coarsening (lines 1–6), community detection (line 7) and uncoarsening (lines 8–10). The inputs are the initial bipartite network $G = (V, E, \sigma, \omega)$, and for each layer a maximal number of levels L and a reduction factor rf.

The bipartite network initially undergoes a one-mode projection transformation, being split into two unipartite networks G^1 and G^2 (line 3). A coarsening process is then applied to each network G^1 and G^2 separately (lines 4–5), level by level, until each one has been reduced by the desired factor. The process comprises a matching algorithm (line 4) and a contracting algorithm (line 5) is applied in G_l from M. In this study, we considered the aforementioned matching algorithms RM, HEM and GHEM [15].

An initial community structure S_L (the starting solution) is then obtained on the coarsest bipartite network G_L, at level $l = L$ (line 7). As G_L and S_L are, respectively, the input and output at this stage (S_L representing the community

Algorithm 3.1. OPM: One-mode projection-based multilevel algorithm for community detection

```
Input:
  bipartite network                  : $\mathsf{G} = (\mathsf{V}, \mathsf{E}, \sigma, \omega)$
  maximal number of levels           : array $\mathsf{L} = \{L_i \mid L_i \in [0, n] \subset \mathbb{Z}\}$ with $1 \leqslant |\mathsf{L}| \leqslant 2$
  reduction factor for each layer:   array $\mathsf{rf} = \{rf_i \mid rf_i \in (0, 0.5] \subset \mathbb{R}\}$ with $1 \leqslant |\mathsf{rf}| \leqslant 2$
Output:
  solution                           : $\mathsf{S} = \{P_1, P_2, \ldots, P_k\}$

 1  for $i \in \{1, 2\}$ do
 2      while ($l \leqslant \mathsf{L}_i$) or (layer is as small as desired) do
 3          $\mathsf{G}_l^i \leftarrow$ one_mode_projection($\mathsf{G}_l$, $i$);
 4          $\mathsf{M} \leftarrow$ matching($\mathsf{G}_l^i$, $\mathsf{rf}_i$);
 5          $\mathsf{G}_{l+1} \leftarrow$ contracting($\mathsf{G}_l$, $\mathsf{M}$);
 6          increment $l$;
 7  $\mathsf{S}_l \leftarrow$ community_detection ($\mathsf{G}_l$);
 8  while $l \neq 0$ do
 9      $\mathsf{S}_{l-1} \leftarrow$ uncoarsening($\mathsf{G}_{l-1}$, $\mathsf{G}_l$, $\mathsf{S}_l$);
10      decrement $l$;
Return: S
```

structure found in network G_L), different algorithms for community detection can be considered to compute this initial solution. Depending on the settings of the coarsening phase, the coarsest bipartite network can be very small, so that computationally expensive algorithms can be considered as possible alternatives (see Sect. 4).

Finally, in the subsequent uncoarsening phase (lines 8–10), solution S_L (as $l = L$) is projected back to G_0 through the intermediate levels $G_{l-1}, G_{l-2}, ..., G_1, G_0$ (line 9). For the specific case of a community detection problem, S_L is described as a partitioning of the vertex set into non-empty partitions P_k with $\cup P_k = S_L$, $P_k \subseteq V_L$. Following previous guidelines proposed for the uncoarsening process [15], a solution S_l is constructed from S_{l+1} simply by assigning vertices $\{u, v\} \in V_l$ to the same community of their parent super-vertex $sV \in V_{l+1}$.

4 Experimental Results and Analysis

We implemented *OPM* and investigated whether it can be employed to scale a costly community detection algorithm, whilst preserving solution quality. We considered the recent *LPAwb+* algorithm introduced by Beckett [5][1]. *LPAwb+* maximizes Barber's modularity through label propagation in weighted bipartite networks. Beckett has shown it has competitive performance compared with state-of-the-art methods. However, it is computationally expensive and thus only feasible on small-scale networks.

[1] https://github.com/sjbeckett/weighted-modularity-LPAwbPLUS.

The *OPM* framework was thus validated by taking Beckett's *LPAwb+* as the target algorithm. Therefore our implementation of *OPM* (Algorithm 3.1) performs the coarsening, runs *LPAwb+* to find the community structure in the coarsest network derived, and projects the solution to obtain the community structure in the original network.

We investigated three alternative instantiations of *OPM* that adopt the matching algorithms GHEM, HEM and RM, hereafter referred to as OPM_{ghem}, OPM_{hem} and OPM_{rm}, respectively. Each instantiation was executed with parameters $rf = 0.5$ and $L = [1, 2, 3]$ in a set of 15 synthetic weighted bipartite networks, identified as R1–R15. Synthetic networks were obtained with the community model described in [5], which creates networks with unbalanced and randomly positioned community structures. Networks of sizes $n = |V_1 + V_2|$ were generated within the range $[1,000; 15,000]$ at increments of $1,000$ and the number of communities was set to $0.01 * n$. Edge weights were randomly assigned from a skewed negative binomial distribution and noise was introduced in the connection patterns by reconnecting a percentage of the edges between and within communities.

Performance was measured by means of the normalized mutual information (NMI), which compares a solution found by a particular algorithm with a reference solution [16], and we also computed execution times. The experiments were executed in a Linux machine with 8 core processor 3.7 GHz CPU and 64 GB main memory. The framework[2] was implemented in Python with the *igraph* library[3]. We report average values obtained from 30 executions for algorithm instances that rely on random choices for matching (OPM_{hem} and OPM_{rm}).

Table 1 shows the accuracy values as measured by NMI on the 15 synthetic networks. The highest values are shown in bold and accuracies equivalent or superior to the baseline solution are highlighted with a gray background.

Table 1. NMI accuracy values of the three *OPM* instantiations and baseline *LPAwb+* in 15 synthetic networks (averages over 30 executions for OPM_{hem} and OPM_{rm}, OPM_{ghem} is deterministic). We highlight the highest accuracy values, shown in bold, and values equivalent or superior to the baseline (*LPAwb+*), shown in shaded cells.

Algorithm		Dataset														
Name	Levels[L]	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15
LPAwb+	0	0.918	0.926	0.983	0.972	0.964	0.990	0.984	**0.999**	**0.999**	0.985	0.989	**0.996**	**0.995**	0.987	0.992
OPM_{ghem}	1	0.992	0.982	0.985	0.990	0.991	0.993	0.990	0.992	0.991	0.995	0.991	0.991	0.994	0.990	0.995
OPM_{ghem}	2	0.981	0.987	0.980	0.986	0.991	0.989	0.988	0.991	0.988	0.989	0.991	0.990	0.991	0.989	0.993
OPM_{ghem}	3	0.875	0.952	0.960	0.964	0.963	0.971	0.972	0.975	0.973	0.975	0.973	0.974	0.976	0.976	0.976
OPM_{hem}	1	0.991	0.987	0.985	0.984	0.991	0.990	0.985	0.992	0.991	0.995	0.990	0.991	0.992	0.991	0.993
OPM_{hem}	2	0.973	0.985	0.981	0.982	0.989	0.987	0.989	0.989	0.988	0.989	0.989	0.988	0.989	0.990	0.991
OPM_{hem}	3	0.873	0.952	0.960	0.966	0.963	0.971	0.972	0.975	0.972	0.975	0.973	0.974	0.975	0.976	0.976
OPM_{rm}	1	0.312	0.358	0.409	0.412	0.407	0.414	0.442	0.462	0.448	0.462	0.468	0.483	0.483	0.502	0.498
OPM_{rm}	2	0.146	0.169	0.147	0.135	0.158	0.171	0.157	0.162	0.150	0.150	0.150	0.161	0.148	0.152	0.161
OPM_{rm}	3	0.100	0.119	0.119	0.098	0.105	0.105	0.090	0.079	0.082	0.069	0.078	0.099	0.084	0.082	0.072

[2] https://github.com/alanvalejo/opm.

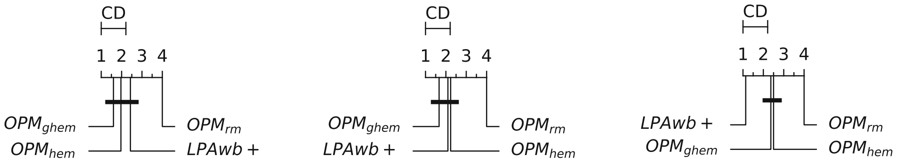
[3] http://igraph.org/python/.

The best performances were achieved by OPM_{ghem} with one level of coarsening ($L = 1$) on 11 out of the 15 networks. The baseline community detection algorithm $LPAwb$+ yielded the best performance in three out of the 15 networks, whereas the worst results were obtained with OPM_{rm} with $L = 3$.

Indeed, the random strategy RM yields extremely poor accuracies, which renders its application unfeasible in real contexts. As opposite, the accuracy values attained by HEM and the greedy strategy GHEM are similar or superior to those of $LPAwb$+. Not surprisingly, limited coarsening levels (mainly $L = 1$) yielded higher accuracy values, and accuracy decreases as the coarsening level increases ($L = 3$). For $L = 3$ the extensive joining of vertices tends to blur the boundaries between adjacent communities. The effect of parameter L depends on network size, i.e., differences in algorithm accuracy are likely to decrease as network sizes increase, which suggests it may be possible to adopt higher values of L on larger networks without such an expressive loss in solution quality.

A Nemenyi post-hoc test [6] was applied to the results shown in Table 1 to detect statistical differences in the performances of the different instances. The results are shown in Fig. 5 for (a) $L = 1$, (b) $L = 2$ and (c) $L = 3$. The critical difference (CD) is indicated at the top of each diagram and the methods' average ranks are placed on the horizontal axes (better ranked on the left). A black line connects algorithms if no significant difference has been detected between them.

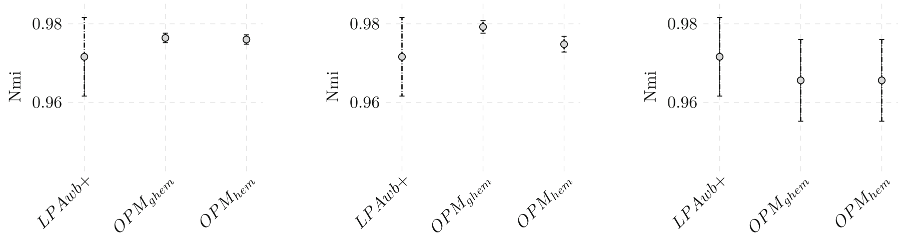
According to the Nemenyi statistics, the critical value for comparing the mean-ranking of two different algorithms at 95 percentile is 1.21 for all diagrams. Mean-rankings differences above this value are significative. When $L = 1$ (Fig. 5(a)), OPM_{ghem} was ranked best, followed by OPM_{hem}, $LPAwb$+ and, finally, OPM_{rm}. Furthermore, no statistically significant difference was observed between OPM_{ghem}, OPM_{hem} and $LPAwb$+. For $L = 2$ (Fig. 5(b)), OPM_{ghem}, OPM_{hem} and $LPAwb$+ were ranked first and no statistically significant difference was observed between them. Finally, for $L = 3$ (Fig. 5(c)) $LPAwb$+ was ranked first with a statistically significant difference between OPM_{ghem}, OPM_{hem} and OPM_{rm}.



(a) $L = 1$ (b) $L = 2$ (c) $L = 3$

Fig. 5. Nemenyi post-hoc test applied to the results from $LPAwb$+ and two OPM variants, for the three settings of parameter L.

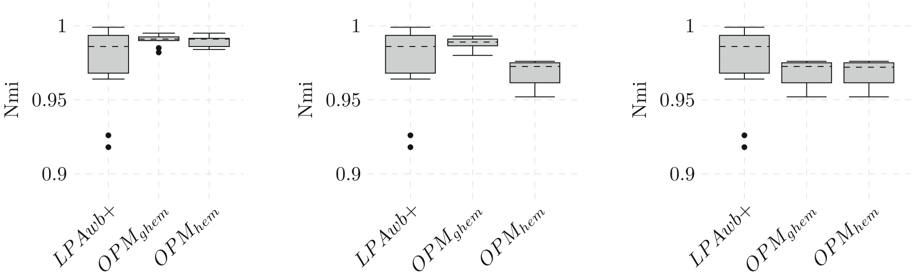
Figure 6 depicts the averages and standard deviations of the accuracy values, whereas Fig. 7 shows their distribution dispersion and outliers, in both cases for the three alternative settings of parameter L. As results from OPM_{rm} were very poor, in order to improve legibility we suppress it from the figures and show only bar and box plots for OPM_{ghem}, OPM_{hem} and $LPAwb$+.



(a) $L = 1$ (b) $L = 2$ (c) $L = 3$

Fig. 6. Averages and standard deviations of the NMI accuracy values obtained with *LPAwb+* and two *OPM* instances in three alternative settings of parameter L (number of levels). (a) $L = 1$, (b) $L = 2$ and (c) $L = 3$.

The bar plots in Fig. 6(a) reveal superior performance and stability of both *OPM* instances when $L = 1$, confirmed by their higher average accuracies and narrower standard deviations. For $L = 3$, *LPAwb+* yielded better results than either *OPM* instance, a consequence of the extensive network reduction. The box plots in Fig. 7(a) reveal that for $L = 1$ all *OPM* instances yielded accuracy values with higher averages and narrower distributions than *LPAwb+*.



(a) $L = 1$ (b) $L = 2$ (c) $L = 3$

Fig. 7. Shape distribution, variability, and averages of the accuracy values yielded by *LPAwb+* and the two instances of *OPM* considering three settings of parameter L. (a) $L = 1$, (b) $L = 2$ and (c) $L = 3$.

We conclude *OPM* instances yielded, in general, higher accuracy and improved stability in comparison to standard *LPAwb+*. In summary, the experimental evidence regarding solution quality (averages, standard deviations and dispersion of the accuracy values) suggests the multilevel framework stabilizes and improves the performance of the algorithm.

The scalability that can be attained with *OPM* was assessed considering the performance of its three instantiations on each individual network. Table 2 shows the absolute execution times (in seconds) of each instance in each network. For OPM_{hem} and OPM_{rm} values refer to average times relative to 30 executions.

Table 2. Absolute runtime (seconds) of *LPAwb+* and four *OPM* instances on each network.

Algorithm		Dataset															sum
Name	Levels[L]	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	
$LPAwb+$	0	14	96	308	904	2,782	2,800	7,146	5,925	9,197	56,119	66,729	75,990	97,392	224,032	302,442	851,876
OPM_{ghem}	1	10	21	48	327	516	540	1,573	2,830	4,289	8,922	11,169	18,214	28,263	41,945	76,439	320,012
OPM_{ghem}	2	8	16	21	201	302	169	661	723	1,984	3,199	8,112	10,219	16,853	19,825	31,591	82,970
OPM_{ghem}	3	3	10	**13**	123	180	142	291	329	528	1,810	2,098	3,991	4,517	5,767	10,892	41,859
OPM_{hem}	1	3	14	43	127	326	540	973	1,933	2,989	6,983	9,628	13,695	18,473	22,875	36,439	115,041
OPM_{hem}	2	2	**6**	21	51	106	139	296	396	854	2,215	3,112	3,776	5,442	6,826	10,538	33,780
OPM_{hem}	3	**1**	**6**	**13**	**33**	60	**55**	80	128	264	818	1,604	1,889	2,351	2,863	3,769	13,934
OPM_{rm}	1	8	57	124	220	419	476	780	1,484	1,991	5,660	9,619	13,127	11,291	17,691	19,202	82,149
OPM_{rm}	2	4	25	40	58	103	150	146	179	345	1,005	1,626	1,873	1,902	1,950	2,041	11,447
OPM_{rm}	3	3	16	28	**33**	**49**	67	**64**	**83**	**194**	**712**	**754**	**849**	**870**	**890**	**903**	5,515
sum		56	267	659	2,077	4,843	5,078	12,010	14,010	22,635	87,443	114,451	143,623	187,354	344,664	494,256	1,558,583

The longest execution time of algorithm *LPAwb+* was 302,442 s (time to process the largest network, R15) and the shortest was 14 s (time to process the smallest one, R1). The most expensive multilevel instance OPM_{ghem} ($L = 1$) consumed 76,439 in R15 and 10 s in R1; hence, OPM_{ghem} run 3.9 to 1.4 times faster than *LAPwb+*, relative to their maximum and minimum execution times, respectively. OPM_{hem} ($L = 1$) consumed 36,439 s in R15 and 3 s in R1. Therefore, OPM_{hem} run 8.3 to 4.6 times faster than *LAPwb+*. The maximum and minimum times of the least expensive instance OPM_{rm} ($L = 3$) were 903 s and 3 s, respectively. Therefore, OPM_{rm} run 335 to 4.6 times faster than *LAPwb+*.

The total time spent running the experiments was 1,558,583 s, or nearly 433 h. In the best case, algorithm implementation OPM_{rm} with $L = 3$ reduced the execution time from 851,876 s (nearly 236 h), required by the standard *LPAwb+*, to 5,515 s (1.5 h), which implies the standard implementation of *LPAwb+* was nearly 154 times slower than this particular implementation of *OPM*.

Executing algorithm *LPAwb+* consumed over 54.6% of the time spent in the experiments, whereas roughly 20.5% of the time was spent running the most expensive instance OPM_{ghem} ($L = 1$) and 0.35% of the time was spent running the least expensive instance OPM_{rm} ($L = 3$).

From this empirical investigation we conclude the proposed *OPM* approach can scale the standard *LPAwb+* whilst yielding more accurate and stable results. Although solution quality degrades as the network is progressively coarsened, runtime drops drastically at each additional coarsening level; hence, a successful solution to the problem requires establishing a suitable trade-off between accuracy and execution time, i.e., identifying a suitable network reduction factor.

5 Conclusions

In this paper we introduced an approach that enables using the multilevel paradigm to scale an expensive community detection algorithm to large-scale bipartite networks. While previous multilevel methods consider only unipartite networks, our one-mode projection-based multilevel method enables handling bipartite networks with standard coarsening algorithms.

Tests on a large suite of synthetic networks have shown that this solution yields results with accuracy comparable to that of the original method and demands considerably shorter execution times. Our tests compared the outcome of the solution implemented with three popular matching strategies for coarsening, namely GHEM, HEM and RM. RM yielded expressive speedups or even improved the asymptotic convergence, but with poor results regarding accuracy, which prevents its practical application. HEM achieved rather good approximation to the standard method in terms of accuracy and acceptable speedups, e.g., execution times over 4.6 times shorter as compared to the standard method. Finally, although GHEM is more costly than RM and HEM, it proved more robust in terms of solution quality, in the test cases considered.

Some issues deserve further attention, such as investigating alternative refinement strategies for the uncoarsening phase and parallel or distributed paradigms to further increase scalability. Another relevant issue is to explore how the choice of rf, the reduction factor parameter, impacts accuracy and speedups on different application scenarios.

Our implementation of the one-mode projection multilevel solution can be downloaded at https://github.com/alanvalejo/opm.

Acknowledgments. Author A. Valejo is supported by a scholarship from the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES). This work has been partially supported by the State of São Paulo Research Foundation (FAPESP) grants 17/05838-3; and the Brazilian Federal Research Council (CNPq) grants 302645/2015-2 and 3056-96/2013-0.

References

1. Abou-Rjeili, A., Karypis, G.: Multilevel algorithms for partitioning power-law graphs. In: Proceedings of the 20th International Parallel and Distributed Processing Symposium, pp. 124–135 (2006)
2. Alzahrani, T., Horadam, K.J.: Community detection in bipartite networks: algorithms and case studies. In: Lü, J., Yu, X., Chen, G., Yu, W. (eds.) Complex Systems and Networks. UCS, pp. 25–50. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-47824-0_2
3. Banos, R., Gil, C., Ortega, J., Montoya, F.G.: Parallel heuristic search in multilevel graph partitioning. In: Proceedings of the 12th Euromicro Conference on Parallel, Distributed and Network-Based Processing, pp. 88–95 (2004)
4. Baños, R., Gil, C., Ortega, J., Montoya, F.G.: A parallel multilevel metaheuristic for graph partitioning. J. Heuristics **10**(3), 315–336 (2004)
5. Beckett, S.J.: Improved community detection in weighted bipartite networks. R. Soc. Open Sci. **3**(1), 18 (2016)
6. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
7. Djidjev, H.N.: A scalable multilevel algorithm for graph clustering and community structure detection. In: Aiello, W., Broder, A., Janssen, J., Milios, E. (eds.) WAW 2006. LNCS, vol. 4936, pp. 117–128. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78808-9_11

8. Djidjev, H.N., Onus, M.: Scalable and accurate graph clustering and community structure detection. IEEE Trans. Parallel Distrib. Syst. **24**(5), 1022–1029 (2013)
9. Dormann, C.F., Strauss, R.: Detecting modules in quantitative bipartite networks: the QuaBiMo algorithm. arXiv preprint 1304.3218 (2013)
10. Dormann, C.F., Strauss, R.: A method for detecting modules in quantitative bipartite networks. Meth. Ecol. Evol. **5**(1), 90–98 (2014)
11. Erciye, K., Alp, A., Marshall, G.: Serial and parallel multilevel graph partitioning using fixed centers. In: Proceedings of the 31st Conference on Current Trends in Theory and Practice of Computer Science, pp. 127–136 (2005)
12. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
13. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**, 7821–7826 (2002)
14. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L.: The large-scale organization of metabolic networks. Nature **407**(6804), 651–654 (2000)
15. Karypis, G., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J. Sci. Comput. **20**(1), 359–392 (1998)
16. Labatut, V.: Generalized measures for the evaluation of community detection methods. CoRR abs/1303.5441 (2013)
17. Larremore, D.B., Clauset, A., Jacobs, A.Z.: Efficiently inferring community structure in bipartite networks. CoRR abs/1403.2933 (2014)
18. LaSalle, D., Karypis, G.: Multi-threaded graph partitioning. In: Proceedings of the 27th IEEE International Parallel and Distributed Processing Symposium, pp. 225–236 (2013)
19. Lasalle, D., Karypis, G.: Multi-threaded modularity based graph clustering using the multilevel paradigm. J. Parallel Distrib. Comput. **76**, 66–80 (2015)
20. Mahmoud, H., Masulli, F., Rovetta, S., Russo, G.: Community detection in protein-protein interaction networks using spectral and graph approaches. In: Formenti, E., Tagliaferri, R., Wit, E. (eds.) CIBB 2013 2013. LNCS, vol. 8452, pp. 62–75. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09042-9_5
21. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA **98**(2), 404–409 (2001)
22. Noack, A., Rotta, R.: Multi-level algorithms for modularity clustering. In: Vahrenhold, J. (ed.) SEA 2009. LNCS, vol. 5526, pp. 257–268. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02011-7_24
23. Opsahl, T.: Triadic closure in two-mode networks: redefining the global and local clustering coefficients. Soc. Netw. **35**, 159–167 (2013)
24. Padrón, B., Nogales, M., Traveset, A.: Alternative approaches of transforming bimodal into unimodal mutualistic networks. The usefulness of preserving weighted information. Basic Appl. Ecol. **12**(8), 713–721 (2011)
25. Rossi, R.G., de Andrade Lopes, A., Rezende, S.O.: Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. Inf. Process. Manage. **52**(2), 217–257 (2016)
26. Rotta, R., Noack, A.: Multilevel local search algorithms for modularity clustering. J. Exp. Algorithmics **16**(2), 2–3 (2011)
27. Schuetz, P., Caflisch, A.: Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. Physical Rev. E Stat. Nonlinear Soft Matter Phys. **77**(4), 1–7 (2008)
28. Schweitz, E.A., Agrawal, D.P.: A parallelization domain oriented multilevel graph partitioner. IEEE Trans. Comput. **51**(12), 1435–1441 (2002)
29. Thébault, E.: Identifying compartments in presence-absence matrices and bipartite networks: insights into modularity measures. J. Biogeogr. **40**(4), 759–768 (2013)

30. Trifunovic, A., Knottenbelt, W.J.: A parallel algorithm for multilevel k-way hypergraph partitioning. In: Proceedings of the Third International Symposium on Parallel and Distributed Computing, pp. 114–121. IEEE (2004)
31. Trifunovic, A., Knottenbelt, W.J.: Parkway 2.0: a parallel multilevel hypergraph partitioning tool. In: Aykanat, C., Dayar, T., Körpeoğlu, İ. (eds.) Proceedings of the 19th International Symposium, Kemer-Antalya, Turkey, 27–29 October 2004
32. Valejo, A., Drury, B., Valverde-Rebaza, J., de Andrade Lopes, A.: Identification of related Brazilian Portuguese verb groups using overlapping community detection. 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. 292–297. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09761-9_35
33. Valejo, A., Ferreira, V., Rocha, G.P., Oliveira, M.C.F., de Andrade Lopes, A.: One-mode projection-based multilevel approach for community detection in bipartite networks. In: Proceedings of the 4th Annual International Symposium on Information Management and Big Data, Track on Social Network and Media Analysis and Mining (SNMAM) (2017)
34. Valejo, A., Rebaza, J.C.V., de Andrade Lopes, A.: A multilevel approach for overlapping community detection. In: Proceedings of the 2014 Brazilian Conference on Intelligent Systems (2014)
35. Valejo, A., Valverde-Rebaza, J., Drury, B., de Andrade Lopes, A.: Multilevel refinement based on neighborhood similarity. In: Proceedings of the 18th International Database Engineering and Applications Symposium, pp. 67–76 (2014)
36. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 409–410 (1998)
37. Ye, Z., Hu, S., Yu, J.: Adaptive clustering algorithm for community detection in complex networks. Phys. Rev. E **78**, 046110 (2008)

Reconstructing Pedestrian Trajectories from Partial Observations in the Urban Context

Ricardo Miguel Puma Alvarez(✉) and Alneu de Andrade Lopes

Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo - Campus de São Carlos,
São Carlos, SP 13560-970, Brazil
rpuma@usp.br, alneu@icmc.usp.br

Abstract. The ever-greater number of technologies providing location-based services has given rise to a deluge of trajectory data. However, most of these trajectories are low-sampling-rate and, consequently, many movement details are lost. Due to that, trajectory reconstruction techniques aim to infer the missing movement details and reduce uncertainty. Nevertheless, most of the effort has been put into reconstructing vehicle trajectories. Here, we study the reconstruction of pedestrian trajectories by using road network information. We compare a simple technique that only uses road network information with a more complex technique that, besides the road network, uses historical trajectory data. Additionally, we use three different trajectory segmentation settings to analyze their influence over reconstruction. Our experiment results show that, with the limited pedestrian trajectory data available, a simple technique that does not use historical data performs considerably better than a more complex technique that does use it. Furthermore, our results also show that trajectories segmented in such a way as to allow a greater distance and time span between border points of pairs of consecutive trajectories obtain better reconstruction results in the majority of the cases, regardless of the technique used.

1 Introduction

Currently, there are many technologies providing location-based services. Some of them are the Global Positioning Systems (GPS), Radio Frequency Identification (RFID), smartphone sensors, ultrasonic and infrared systems, etc. [12]. All these technologies allow a large-scale generation of trajectory data of moving objects, which can be used to perform several data mining tasks. This scenario paved the way for the rise of the trajectory data mining field. There are several trajectory data mining applications such as path discovery, location prediction, behavior analysis, urban services improvement, etc. [12]. However, there are still some important challenges to be addressed regarding storage, computation and trajectory data mining [13]. Due to storage, energy consumption and transmission issues, these trajectories are generally collected at low sampling rates,

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 137–148, 2018.
https://doi.org/10.1007/978-3-319-90596-9_10

consequently, they have long time intervals between location updates. In general, these trajectories provide a very limited representation of the real paths. This type of trajectories are called uncertain trajectories [14].

Most often, trajectories can be tracked very accurately with GPS-embedded devices like smartphones or automotive navigation systems. Nevertheless, a recent study has demonstrated that, aiming at reducing energy consumption, the majority of taxis of big cities use sampling intervals of two minutes [15].

The high energy consumption of GPS impairs its use in smartphones for long periods of time. Furthermore, most social networks provide check-in services, which allow user location sharing. Thus, it is possible to create trajectories by sorting these check-ins chronologically. In a similar way, trajectories can be generated from geo-tagged photos in photo sharing sites like Flickr[1]. In spite of that, the location updates generated through these sites are low-sampling-rate.

Addressing this issue is very important for several different trajectory data mining applications. For instance, trajectories generated from geo-tagged photos could be reconstructed and used in itinerary recommendations applications. Additionally, other tasks like indexing and querying processing efficiency can be affected [11].

Motivated by this problem, many works on trajectory reconstruction have been published. Most of them use road network information through a graph whose nodes represent intersections and terminal points, and the edges depict road segments. On the other hand, there are also some works that do not take into account this kind of information [15]. These works aim to reconstruct trajectories in rural areas where there is no road network, and, also, trajectories of animals and certain natural phenomena like hurricanes. However, here we are focused on pedestrian trajectory reconstruction in urban areas.

An example of a method of reconstruction that uses road network information is InferTra [10]. This technique, instead of predicting the most likely route of a vehicle, returns an edge-weighted graph that summarizes all probable routes. The trajectory reconstruction process employs Gibss sampling by learning a Network Mobility Model (NMM) from a database of historical trajectories. Other works that also use road information are [11,16–18], to cite a few.

Nevertheless, as InferTra, most works are focused on reconstructing vehicle trajectories. This is mainly due to the fact that some pedestrian routes comprise small alleys and trails that are so narrowed to be traversed by other transportation mode different from walk. Despite of that, free collaborative maps like OpenStreetMap[2] allow the addition of these type of routes exclusively traversed by pedestrians to the road network. This way, it would be possible to reconstruct pedestrian trajectories using road network information.

Considering that, the main contribution of this work is to be one of the first ones aimed to study exclusively the reconstruction of pedestrian trajectories using road network information. We compare two main techniques originally used for vehicle trajectory reconstruction, determine the one that performs the

[1] https://www.flickr.com/.

[2] https://www.openstreetmap.org.

best and its settings to make it possible. Thus, we demonstrate that it is possible to reconstruct pedestrian trajectories effectively by using information from the road network. To achieve this, we depict a framework to reconstruct pedestrian trajectories composed by three phases. Firstly, we segment trajectories by using three different settings in order to study their influence over the quality of the reconstruction. Secondly, we perform a map matching task on these segmented trajectories using a free tool, thereby generating a set of network-constrained trajectories. Thirdly, we apply two different trajectory reconstruction techniques on this new trajectory set. We compare these two techniques, one of them a simple technique that only takes into account the road network information, and a more complex one that besides the road network structure uses historical trajectory data. We show that, under limited data conditions, the simpler technique greatly outperforms the more complex technique in pedestrian trajectory reconstruction. Furthermore, our results also demonstrate that trajectories segmented in such a way as to allow a greater distance and time span between consecutive sub-trajectories obtain better reconstruction results in the majority of the cases, regardless of the technique used.

The remainder of this paper is organized as follows: Sect. 2 presents the basic concepts and related work on trajectory reconstruction. In Sect. 3, we describe how these concepts and methods are adapted for the case of pedestrian trajectories. Section 4 presents the experiments performed over two data sets and the respective analysis of the results obtained. Finally, some conclusions are presented in Sect. 5.

2 Background Concepts and Related Work

A general process of trajectory reconstruction illustrated in [10] can be seen in Fig. 1. This process is composed by three phases: segmentation, map matching and inference. During the segmentation phase, the users logs (commonly stored as a set of flat text files where each file belongs to a unique user) are transformed into trajectories taking into account some specific criteria. Once the logs are transformed, the resulting trajectories are grouped and the user information is discarded. After segmentation, in the map-matching phase, this group of trajectories changes its original format (a sequence of spatial-temporal points) and each trajectory is represented in terms of road network vertices. Finally, these network-constrained trajectories in conjunction with road network information are used in the inference phase to reconstruct a trajectory from a set of observations received as input.

2.1 Trajectory Segmentation

Many trajectory data mining tasks as trajectory clustering and classification need to divide a trajectory into segments. Segmentation reduces complexity and allows mining other interesting patterns, such as sub-trajectories [20]. In the context of trajectory reconstruction, these sub-trajectories could be thought as pedestrian trips.

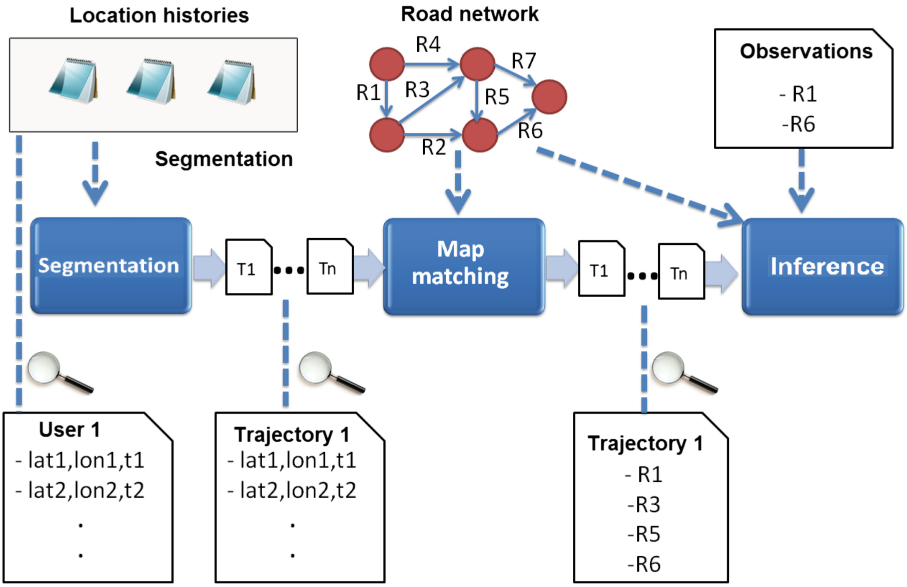


Fig. 1. Trajectory reconstruction framework.

GPS logs generally record people's movement for long periods (e.g., days, weeks, months and even years), in which the person could make multiple trips. When the person stops for a relatively long time, this could indicate the end of a trajectory and the start of the next one. However, if we took each user log as a long and unique trajectory, these trips could be lost. Therefore, in order to reflect the real pedestrian intention as well as possible, we must use a segmentation method.

There are three main categories of segmentation methods based on different criteria, namely, time interval, shape of a trajectory and the semantic meanings of points in a trajectory [20].

In the first category, as showed in Fig. 2(A), if the time interval between a pair of consecutive sampling points overpasses a given threshold, the trajectory is split into two parts at these points. On the other hand, the methods based on the shape of a trajectory search for turning points whose changes of direction are regulated with a threshold. Figures 2(B) and (C) illustrate this kind of methods. Finally, the category of methods based on the semantics of the trajectory points aim to leverage the underlying meaning in the trajectory. For example, a common segmentation method is to divide a trajectory into sub-trajectories of different transportation modes such as walking or driving. Another method of this category leverages the information provided by stay points, as depicted in Fig. 2(D). A stay point is a spatial point that denotes locations where people have stayed for a while, such as restaurants, libraries or home [20]. So, this kind of method aims to detect stay points and use them as boundaries to split a trajectory into sub-trajectories.

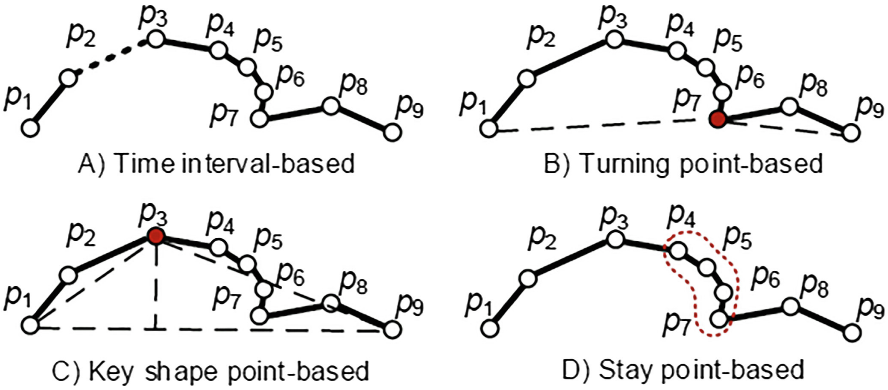


Fig. 2. Methods of trajectory segmentation [20].

2.2 Map Matching

The second phase of this general framework is the map matching process, which aims to transform our set of GPS trajectories into network-constrained trajectories by matching each GPS point to an edge of the road network of a certain city. Map matching is an important research topic and there are many works focused on it [7–9].

Using a correct map matching method to align GPS points onto the road segments is relevant because the GPS points do not reflect their true position due to the GPS measurement error.

2.3 Inference

The task of inferring a preprocessed (network-constrained) trajectory from partial observations is called trajectory inference and it is the last phase of the framework. It is important to notice that we make a distinction between the terms "inference" and "reconstruction". We use the term "reconstruction" to refer to the entire process (segmentation, map matching and inference).

One of the best techniques of trajectory inference using road network information is InferTra [10]. This method outperforms other state-of-the-art techniques by a large margin. InferTra is composed by two phases. Firstly, in the offline processing, this technique uses the historical network-constrained trajectories and a road network to create a generative model called Network Mobility Model (NMM), which is a weighted directed graph whose edge weights denote the probability of the corresponding road segment being traversed. Hence, NMM learns the mobility patterns in a road network from a database of historical trajectories. Secondly, in the online processing, given an uncertain trajectory (a trajectory with low-sample-rate location updates), NMM is used to generate a weighted subgraph that depicts the probabilities associated to each possible trajectory arising from the uncertain trajectory location updates. The entire pipeline of the InferTra algorithm is depicted in Fig. 3.

3 Reconstructing Uncertain Pedestrian Trajectories

In this section, we describe the settings and methods used in each phase of the framework presented in the previous section for pedestrian trajectory reconstruction.

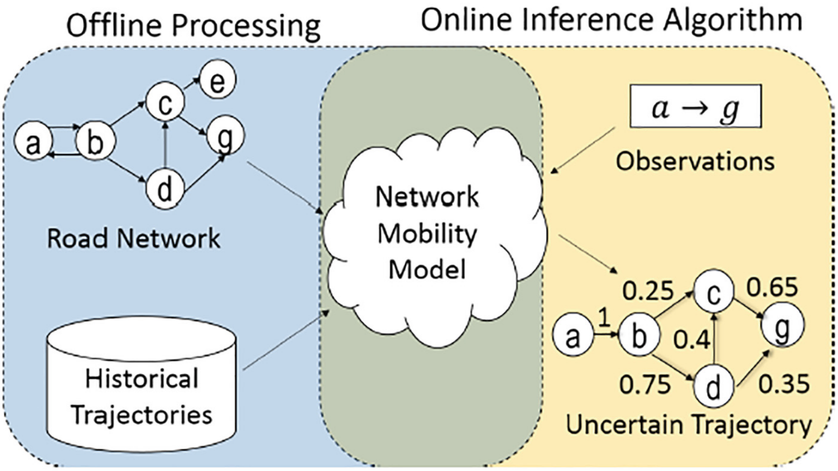


Fig. 3. Pipeline of the InferTra algorithm [10].

3.1 Trajectory Segmentation

In this work, we used the method of stay points which is based on the semantic meanings of points. Thus, we segment the GPS logs into effective trajectories, which have specific source and destination stay points.

In the Table 1, we observe three different segmentation settings based on the degree of tolerance, which is based on two criteria, the time between one GPS point to the next one and the distance between them. A GPS point is the representation of a location update in terms of space and time by means of geographic coordinates (latitude and longitude) and a timestamp. These criteria were already used by [22] to build a stay point detection algorithm. However, in our case, since we work under data limited conditions, instead of detecting groups of points, we detect single point locations in order to preserve as much information as possible to perform the reconstruction.

The rationale behind this arrangement of values lies mainly in the medium tolerance setting. This setting states that if the distance between a location update and the next one is greater than the combined lengths of three average city blocks (100 m [1–3]), the previous location update is considered as the end of a trajectory and the last one as the start of the next trajectory. Likewise, if the time between two location updates is more than 10 min, the first location

Table 1. Trajectory segmentation settings

	Medium	High	Low
Distance (m)	300	600	150
Time (mins)	10	20	5

update and its successor are considered as the end and the start, respectively, of two different and successive trajectories. The idea behind the period of 10 min is to assume that a pedestrian can make some small stops due to external factors such as a quick conversation with some unexpected acquaintance on the way or waiting for the traffic light to change to cross a street, which far from meaning a source or destination, are just trip interruptions. Thus, finally, the half and the double of the values of these time and distance thresholds are allocated to the low and high tolerance settings respectively.

3.2 Map Matching

As mentioned in Sect. 1, thanks to the growing use of free collaborative maps, now, road networks can include the trails walked exclusively by pedestrians as new edges. Thus, it becomes possible to apply map matching techniques on pedestrian trajectories.

Additionally, there are free tools available that perform map matching tasks as Graphhopper[3]. Thus, in this work, we used the Graphhopper tool to create a set of network-constrained trajectories that are used by the inference methods in the next phase. This tool is an open-sourced project supported by an extensive community of developers. In addition to that, Graphhopper counts with a comprehensive documentation, successful use cases and forum.

3.3 Inference

In this last phase, we use the method described in the previous section, InferTra. On the other hand, we also used the Shortest Path technique to establish contrast with InferTra. The Shortest Path is a much simpler technique compared to InferTra, so this comparison can reveal whether a simple or more complex approach performs better when it comes to reconstructing pedestrian trajectories using road network information.

4 Experiments

We use two data sets to perform the reconstruction of pedestrian trajectories. In each data set, the three trajectory segmentation settings previously depicted are used, low, medium and high tolerance, in order to study their influences over the performance of the reconstruction. Finally, we compare the performance of InferTra [10] and Shortest Path for different sampling intervals.

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

4.1 Data Sets

The data sets considered in our experiments are (i) RadrPlus and (ii) Geolife [4–6]. RardPlus is a location-based social network developed in the University of São Paulo, campus of São Carlos, Brazil. This social network has the unique characteristic of being focused on communities. Therefore, RadrPlus provides functionalities not just for individual users like traditional social networks, but for groups of users as well, within a geolocated environment. The RadrPlus data set comprises trajectories of a group of 15 users in a period of 9 months. These trajectories were recorded in different parts of the city of São Carlos, but mainly around the campus of the university and its surroundings. Additionally, RadrPlus data set trajectories were labeled with two transportation modes, car and walk. The second data set is provided by the Geolife project, a location-based social network developed by Microsoft Research Asia. The Geolife data set contains trajectories of 182 users in a period of over five years. These trajectories were recorded in 30 cities of China and some cities in USA and Europe; however, most trajectories were recorded in the city of Beijing, China. In addition to that, a group of 73 users labeled their trajectories with transportation modes like walk, bike, bus, car, train, airplane and others. In spite of the variety of trajectory data of both data sets, we only select for our experiments the trajectories made by pedestrians, *i.e,* trajectories labeled with walk as transportation mode.

4.2 Experimental Results

We present and discuss the results of using the InferTra and Shortest Path techniques in the reconstruction of pedestrian trajectories. As mentioned before, we compare these two techniques to evaluate if a simple or a more complex approach becomes more suitable for the case of pedestrian trajectory reconstruction. Each technique was implemented in Java. We segment trajectories from RadrPlus and Geolife data sets by using the low, medium and high tolerance settings and for each resulting trajectory set, we use the Graphhopper tool to transform these trajectories into network-constrained trajectories. Finally, we apply the aforementioned reconstruction techniques on the six different data sets generated as

Table 2. Main characteristics of each resulting trajectory data set after applying segmentation settings Low Tolerance (LT), Medium Tolerance (MT) and High Tolerance (HT) on RadrPlus and Geolife data sets.

	NT	AL (pp)	AD (secs)
RadrPlus (LT)	76	14.12	487.48
RadrPlus (MT)	116	17.31	604.15
RadrPlus (HT)	155	21.06	1045.61
Geolife (LT)	3895	13.06	721.74
Geolife (MT)	3601	13.71	819.57
Geolife (HT)	3134	13.79	877.2

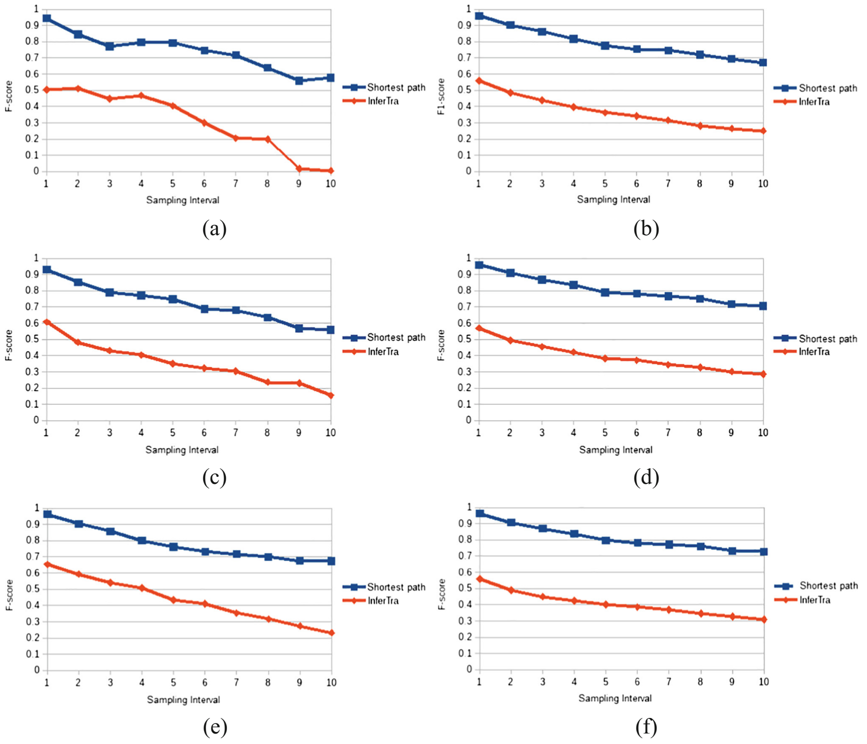
Table 3. Results of trajectory reconstruction on RadrPlus and Geolife data sets under trajectory segmentation configurations High Tolerance (HT), Medium Tolerance (MT) and Low Tolerance (LT) for different sampling intervals.

	Sampling interval	LT	MT	HT
RadrPlus (SP)	1	0.944	0.930	**0.963**
	2	0.846	0.852	**0.903**
	3	0.770	0.790	**0.857**
	4	0.796	0.771	**0.800**
	5	**0.793**	0.748	0.762
	6	**0.746**	0.689	0.732
	7	**0.718**	0.680	0.717
	8	0.639	0.636	**0.700**
	9	0.558	0.567	**0.676**
	10	0.578	0.558	**0.673**
Geolife (SP)	1	0.959	0.960	**0.962**
	2	0.902	**0.909**	0.907
	3	0.862	0.868	**0.870**
	4	0.818	0.835	**0.837**
	5	0.775	0.789	**0.797**
	6	0.754	0.782	**0.783**
	7	0.746	0.766	**0.771**
	8	0.721	0.751	**0.761**
	9	0.694	0.717	**0.732**
	10	0.671	0.705	**0.727**
RadrPlus (IT)	1	0.502	0.608	**0.654**
	2	0.510	0.481	**0.593**
	3	0.446	0.430	**0.539**
	4	0.467	0.404	**0.507**
	5	0.404	0.350	**0.434**
	6	0.298	0.323	**0.411**
	7	0.205	0.303	**0.353**
	8	0.197	0.237	**0.315**
	9	0.018	0.230	**0.273**
	10	0.005	0.154	**0.231**
Geolife (IT)	1	0.558	**0.569**	0.558
	2	0.484	**0.494**	0.490
	3	0.437	**0.455**	0.450
	4	0.395	0.421	**0.424**
	5	0.363	0.382	**0.401**
	6	0.342	0.371	**0.386**
	7	0.314	0.344	**0.369**
	8	0.281	0.326	**0.346**
	9	0.264	0.300	**0.326**
	10	0.250	0.285	**0.309**

showed in Table 2. The main characteristics depicted in that table are the Number of Trajectories (NT), Average Length (AL), which is the average number of points per trajectory (pp), and the Average Duration (AD) of the trajectories.

We notice that, in the case of Geolife, the lower the tolerance, the greater the number of trajectories, and the shorter the trajectory length and duration. However, in the case of RadrPlus, the lower the tolerance, the lesser the number of trajectories, and the shorter the trajectory length and duration. These results are interesting due to the fact that Geolife's data was collected with a fixed, short sampling interval (Static Duty Cycle [19]) while RadrPlus' data collection process used a adaptive strategy that allocates short and long sampling intervals depending on the context (Dynamic Duty Cycle [19]). Therefore, we observe how the election of the data collection method affects the main characteristics of the resulting segmented trajectories.

On the other hand, since InferTra reconstructs a trajectory as a weighted graph, we use the adapted F-score measure described in [10] to evaluate InferTra performance, whereas the standard F-score was used in the case of Shortest Path.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4. Results of the trajectory reconstruction on the RadrPlus (left column), and Geolife (right column) data sets using InferTra and Shortest Path algorithms. Each point in the figure indicates the value of the F-score for a certain sampling interval, expressed in minutes, under the segmentation settings (a,b) High Tolerance, (c,d) Medium Tolerance, and (e,f) Low Tolerance.

These two techniques are evaluated for different sampling rates expressed in minutes.

From Fig. 4, we can easily observe that Shortest Path greatly outperforms InferTra, regardless of the data set, trajectory segmentation setting and sampling interval used. Additionally, as expected, we also observe that the best results correspond to the shortest sampling intervals. Nevertheless, it is not clear if there is a difference among trajectory segmentation settings. Thus, to analyze the impact of these settings over the reconstruction, we organize the data so that we can easily compare the obtained results for each setting. This way, in Table 3, we observe a clear evidence that the High Tolerance (HT) setting obtains better results for the majority of the cases. This setting only presents lower values of F-score in the 17.5% of the cases.

5 Conclusion

We studied the reconstruction of pedestrian trajectories using road network information. Three different segmentation settings were proposed to study their influence over the reconstruction. These settings were established based on the concept of tolerance which was defined based on two criteria that comprise the distance and time span between points in a trajectory. Moreover, two state-of-the-art techniques were tested. One of these techniques uses both road network information and historical trajectories, whereas the simpler one only uses the road network structure. Empirical analysis of these two techniques on two data sets shows that the simpler technique performs better under limited data conditions than the more complex one when it comes to pedestrian trajectories. Additionally, the high tolerance segmentation setting proposed obtains better reconstruction results in a majority of the cases for both techniques.

Acknowledgments. This work was supported by FAPESP grant 2015/14228-9, CNPq grants 302645/2015-2, 162262/2014-0 and CAPES grant PROEX-9152559/M.

References

1. Yeang, L.D., et al.: Urban Design Compendium. English Partnerships/Housing Corporation, London (2000)
2. Transportation Studies Group of the German Aerospace Center and others: Urban Block Design guideline/manual to best practice - Project METRASYS (2012)
3. DMJM HARRIS and AECOM and CLARKE CATON HINTZ: New Jersey Long Range Transportation Plan 2030. Technical Memorandum. Task 11: Local Street Connectivity Redefined. The New Jersey Department of Transportation (2008)
4. Zheng, Y., Zhang, L., Xie, X., Ma, W.Y.: Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th international conference on World wide web, pp. 791–800 (2009)
5. Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W.Y.: Understanding mobility based on GPS data. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 312–321 (2008)

6. Zheng, Y., Xie, X., Ma, W.Y.: GeoLife: a collaborative social networking service among user, location and trajectory. IEEE Data Eng. Bull. **33**(2), 32–39 (2010)
7. Lou, Y., Zhang, C., Zheng, Y., Xie X., Wang W., Huang, Y.: Map-matching for low-sampling-rate GPS trajectories. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 352–361 (2009)
8. Greenfeld, J.S.: Matching GPS observations to locations on a digital map. In: 81th Annual Meeting of the Transportation Research Board, vol. 1(3), pp. 164–173 (2002)
9. Yuan, J., Zheng, Y., Zhang, C., Xie, X., Sun, G.Z.: An interactive-voting based map matching algorithm. In: 2010 Eleventh International Conference on Mobile Data Management (MDM), pp. 43–52 (2010)
10. Banerjee, P., Ranu, S., Raghavan, S.: Inferring uncertain trajectories from partial observations. In: 2014 IEEE International Conference on Data Mining (ICDM), pp. 30–39 (2014)
11. Zheng, K., Zheng, Y., Xie, X., Zhou, X.: Reducing uncertainty of low-sampling-rate trajectories. In: 2012 IEEE 28th International Conference on Data Engineering (ICDE), pp. 1144–1155 (2012)
12. Feng, Z., Zhu, Y.: A survey on trajectory data mining: techniques and applications. IEEE Access **4**, 2056–2067 (2016)
13. Baraniuk, R.G.: More is less: signal processing and the data deluge. Science **331**(6018), 717–719 (2011)
14. Zheng, Y., Zhou, X.: Computing with Spatial Trajectories. Springer, New York (2011). https://doi.org/10.1007/978-1-4614-1629-6
15. Wei, L.Y., Zheng, Y., Peng, W.C.: Constructing popular routes from uncertain trajectories. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 195–203 (2012)
16. Hunter, T., Abbeel, P., Bayen, A.M.: The path inference filter: model-based low-latency map matching of probe vehicle data. In: Frazzoli, E., Lozano-Perez, T., Roy, N., Rus, D. (eds.) Algorithmic Foundations of Robotics X. STAR, vol. 86, pp. 591–607. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36279-8_36
17. Li, M., Ahmed, A., Smola, A.J.: Inferring movement trajectories from GPS snippets. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 325–334 (2015)
18. Chiang, M.F., Lin, Y.H., Peng, W.C., Yu, P.S.: Inferring distant-time location in low-sampling-rate trajectories. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1454–1457 (2013)
19. Wu, C.L., Huang, Y.T., Wu, C.L., Chu, H.H., Huang, P., Chen, L.J.: An adaptive duty-cycle scheme for GPS scheduling in mobile location sensing applications. In: Proceedings of PhoneSense (2011)
20. Zheng, Y.: Trajectory data mining: an overview. ACM Trans. Intell. Syst. Technol. (TIST) **6**(3), 29 (2015)
21. Ochieng, W.Y., Quddus, M., Noland, R.B.: Map-matching in complex urban road networks. Revista Brasileira de Cartografia **2**(55), 1–14 (2003)
22. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.Y.: Mining user similarity based on location history. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 34 (2008)

Author Index