Procesamiento del Lenguaje Natural, Revista nº 62, marzo de 2019, pp. 111-114 recibido 16-11-2018 revisado 07-01-2019 aceptado 05-02-2019

A Cross-domain and Cross-language Knowledge-based Representation of Text and its Meaning*

Una representación translingüe y transdominio del texto y su significado basada en el conocimiento

Marc Franco-Salvador
PRHLT Research Center
Universitat Politècnica de València
Camino de Vera s/n, 46022. Valencia, Spain

Symanto Research
Pretzfelder Str. 15, 90425 Nürnberg
marc.franco@symanto.net

Abstract: Ph.D. thesis (international doctorate mention) in Computer Science written by Marc Franco Salvador under the supervision of Dr. Paolo Rosso at the Universitat Politècnica de València. The author was examined in Valencia in May 2017 by a jury composed of the following doctors: Nicola Ferro (University of Padua), Bernardo Magnini (Fondazone Bruno Kessler), and Simone Paolo Ponzetto (University of Mannheim). The international doctorate mention was granted thanks to the completion of the following research internships: 1 year at the Sapienza University of Rome (Italy) under the supervision of Dr. Roberto Navigli, 2 months at the IIIT of Hyderabad and at Veooz (India) under the supervision of Dr. Vasudeva Varma and Dr. Prasad Pingali, 1 month at the INAOE (Mexico) under the supervision of Dr. Manuel Montes-y-Gómez, and 3 months at Symanto Group (Germany) under the supervision of Dr. Yassine Benajiba. The obtained grade was Excellent with *Cum Laude* distinction.
Keywords: Cross-language, cross-domain, knowledge graphs, plagiarism detection, information retrieval, text classification, sentiment analysis

Resumen: Tesis doctoral (con mención de doctorado internacional) en Informática realizada por Marc Franco Salvador bajo la supervision del Dr. Paolo Rosso en la Universitat Politècnica de València. La lectura de la tesis fue realizada en Valencia en Mayo del 2017 por un jurado compuesto por los siguientes doctores: Nicola Ferro (University of Padua), Bernardo Magnini (Fondazone Bruno Kessler) y Simone Paolo Ponzetto (University of Mannheim). La mención de doctorado internacional fue otorgada gracias a la realización de las siguientes estancias de investigación: 1 año en la Sapienza University of Rome (Italia) bajo la supervisión del Dr. Roberto Navigli, 2 meses en el IIIT de Hyderabad y en Veooz (India) bajo la supervisión del Dr. Vasudeva Varma y el Dr. Prasad Pingali, 1 mes en el INAOE (México) bajo la supervision del Dr. Manuel Montes-y-Gómez y 3 meses en Symanto Group (Alemania) bajo la supervisión del Dr. Yassine Benajiba. La calificación obtenida fue Sobresaliente con mención *Cum Laude*.
Palabras clave: Translingüe, transdominio, grafos de conocimiento, detección de plagio, recuperación de información, clasificación de texto, análisis del sentimiento

* This research has been carried out in the framework of the European Commission project WIQ-EI IRSES (no. 269180), and the national projects DIANA-APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01), Destilado de opiniones desde contenidos generados por usuarios (TIN2011-14726-E), and SomEMBED: SOcial Media language understanding - EMBEDing contexts (TIN2015-71147-C2-1-P).

ISSN 1135-5948. DOI 10.26342/2019-62-15

1 Introduction

One of the most challenging aspects of Natural Language Processing (NLP) involves enabling computers to derive meaning from human natural language. To do so, several meaning or context representations have been proposed with competitive performance. However, these representations still have room for improvement when working in a cross-domain or cross-language scenario.

In this thesis we study the use of knowledge graphs[1] as a cross-domain and cross-language representation of text and its meaning. To do so, we generate the knowledge graphs with BabelNet,[2] the widest-coverage multilingual semantic network. This allows to have graphs that expand and relate the original concepts belonging to a set of words present in a text. This also provides with a language coverage of hundreds of languages and millions human-general and -specific concepts.

1.1 Motivation and Objectives

The use of the recent and popular distributed representations enables to accurately model text meaning and produced significant improvements in NLP tasks. However, there is a room of improvement in the cross-language scenario. Most of the approaches need high amounts of data in order to train representative models. In addition, the computational complexity and the amount of training data is proportional to the number of languages employed.

The use of knowledge graphs provided in the past with state-of-the-art results in the task of mono- and cross-language word sense disambiguation. The starting point of this work is the observation that, if these graphs provided with the correct disambiguations of a text, even at cross-language level, they are adequate as representation of the meaning of that text. In addition, we believe that BabelNet, the multilingual semantic network employed to generate the graphs, with a language coverage of hundreds of languages and millions concepts, makes this representation domain- and language-independent. Therefore, its complexity is also independent of the number of languages employed. In consequence, knowledge graphs are adequate for cross-domain and cross-language NLP and Information Retrieval (IR) tasks. Moreover, knowledge graphs have several implicit characteristics (i.e., Word Sense Disambiguation (WSD), vocabulary expansion, and language independence) that have different impact on their performance in NLP similarity analysis tasks. Finally, we consider that this representation may be useful for other non-cross-language or non-cross-domain NLP tasks such as community questions answering, native language identification, and language variety identification.

Considering what aforementioned statements said, this thesis has the following objectives:

- To study the potential of knowledge graph-based features for cross-domain NLP tasks.
- To develop a cross-language similarity analysis model for NLP and IR tasks.
- To study the knowledge graph characteristics for cross-language similarity analysis tasks.
- To evaluate the performance of the developed approaches and compare them with the state-of-the-art models.
- To employ knowledge graphs for other NLP tasks.

2 Thesis Overview

This thesis has six chapters and is presented as a compendium of research articles which were published during the study phase of this PhD. We include two international journal articles and an international conference paper as chapters of this work. The thesis is structured as follows.

In Chapter 1 we present the introduction of this thesis. It includes the related work, motivation and objectives, our research questions, and the contributions of this research.

In Chapter 2 we present our journal article published in Franco-Salvador et al. (2015). In that work we employed knowledge graph-based features, such as WSD and vocabulary expansion-based ones (along with other traditional ones: bag of words and n-grams), for single- and cross-domain polarity classification. The evaluation includes a thorough

[1] A knowledge graph is a subset of a semantic network (also known as knowledge base) focused on the concepts belonging to a text, and the intermediate concepts and relations between them.

[2] `http://babelnet.org`

analysis of the knowledge graph-based features and a comparison with the state of the art in domain adaptation.

In Chapter 3 we present our journal article published in Franco-Salvador, Rosso, and Montes y Gómez (2016). This is the reference work of our cross-language knowledge graph analysis model for cross-language similarity analysis. That method employs knowledge graphs as a cross-language representation of the text and its meaning. We also study the implicit and most relevant characteristics of the knowledge graphs at cross-language level, i.e., WSD, vocabulary expansion, and language independence. The evaluation includes the task of Spanish-English and German-English plagiarism detection and a comparison of the models in cases of plagiarism with paraphrasing.

In Chapter 4 we present our conference article published in Franco-Salvador, Rosso, and Navigli (2014). This publication presents a modified version of our cross-language knowledge graph analysis model. This also includes a vector component to cover shortcomings such as out-of-vocabulary words and verbal tenses. The evaluation in the tasks of cross-language document retrieval and categorization compares this new model with the state of the art using several language pairs.

In Chapter 5 we discuss the results that have been obtained on the previous chapters. Moreover, we complement our study with some further experiments to complete the picture at task level, and analyse the obtained results from a cross-domain and cross-language perspective. In addition, we present our experiments and results with knowledge graphs in other NLP tasks such as community questions answering, native language identification, and language variety identification.

In Chapter 6 we draw the main conclusions of the thesis, as well as its contributions and research lines for future work.

3 Thesis Contributions

The main contributions of this thesis are described below.

From the representation viewpoint, we proved that knowledge graphs can be employed as a cross-domain and cross-language representation of text and its meaning. We used several reference datasets to show diverse results and comparisons with the state of the art and to justify the validity and potential of this representation. We supported all our conclusions with standard tests of statistical significance of results. In addition, we studied from a theoretical and practical perspective, the main characteristics that contribute to the knowledge graphs performance.

With respect to the tasks, we showed how to obtain state-of-the-art performance with knowledge graphs in several single- and cross-domain NLP and IR tasks: single- and cross-domain polarity classification (Franco-Salvador et al., 2015; Giménez-Pérez, Franco-Salvador, and Rosso, 2017), cross-language plagiarism detection (Franco-Salvador, Gupta, and Rosso, 2013; Franco-Salvador, Rosso, and Montes y Gómez, 2016; Franco-Salvador et al., 2016), document retrieval and categorization (Franco-Salvador, Rosso, and Navigli, 2014), and community questions answering. In addition, we showed the potentiality of knowledge graphs for native language identification (Franco-Salvador, Kondrak, and Rosso, 2017) and language variety identification (Rangel, Franco-Salvador, and Rosso, 2016).

From the modelling viewpoint, we employed knowledge graphs to obtain state-of-the-art performance in two different ways: (i) as a source of feature extraction for classification and regression, and (ii) as a representation, as part of the proposed cross-language similarity analysis models. With respect to these two models, we proposed one that employs knowledge graphs as representation of the text and its meaning, and we proposed another one that complements that representation with a vector-based representation to cover the graph shortcomings. In addition, we proposed a new embedding-based weighting scheme for the semantic relations between the knowledge graph concepts. This scheme proved to outperform the classical one employed in the BabelNet multilingual semantic network.

Finally, some contributions only partially related to knowledge graphs were achieved during this research. First, we proposed the continuous word alignment-based similarity analysis model that notably improved the performance of distributed representations of words in cross-language plagiarism detection. Next, we proved the relationship between the native language and the language variety identification tasks by solving both with the same approach wit-

hout any task-specific adaptation. The proposed string kernels-based approach obtained state-of-the-art performance in several datasets of the two tasks. Finally, we hypothesised that there is a relationship between knowledge graphs and distributed representations. We studied, with interesting results, how both complement each other for several NLP and IR tasks.

References

Franco-Salvador, M., F. L. Cruz, J. A. Troyano, and P. Rosso. 2015. Cross-domain polarity classification using a knowledge-enhanced meta-classifier. *Knowledge-Based Systems*, 86:46 – 56.

Franco-Salvador, M., P. Gupta, and P. Rosso. 2013. Cross-language plagiarism detection using a multilingual semantic network. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR'13)*, LNCS(7814), pages 710–713. Springer-Verlag.

Franco-Salvador, M., P. Gupta, P. Rosso, and R. E. Banchs. 2016. Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowledge-Based Systems*, 111:87–99.

Franco-Salvador, M., G. Kondrak, and P. Rosso. 2017. Bridging the native language and the language variety identification tasks. In *Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES'17)*.

Franco-Salvador, M., P. Rosso, and M. Montes y Gómez. 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Information Processing & Management*, 52(4):550–570.

Franco-Salvador, M., P. Rosso, and R. Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*, pages 414–423. Association for Computational Linguistics.

Giménez-Pérez, R. M., M. Franco-Salvador, and P. Rosso. 2017. Single and cross-domain polarity classification using string kernels. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*. Association for Computational Linguistics.

Rangel, F., M. Franco-Salvador, and P. Rosso. 2016. A low dimensionality representation for language variety identification. In *Proceedings of the 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'16)*. Springer-Verlag.