Information Processing and Management 52 (2016) 550–570

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

A systematic study of knowledge graph analysis for cross-language plagiarism detection

Marc Franco-Salvador[a,*], Paolo Rosso[a], Manuel Montes-y-Gómez[b]

[a] Pattern Recognition and Human Language Technology (PRHLT) Research Center, Universitat Politècnica de València, Camino de Vera s/n, Valencia 46022, Spain
[b] Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro 1, Puebla 72840, Mexico

ARTICLE INFO

Article history:
Received 16 September 2015
Revised 3 December 2015
Accepted 10 December 2015
Available online 15 January 2016

Keywords:
Cross-language
Plagiarism detection
Knowledge graphs
Multilingual semantic network
Distributed representations
Evaluation

ABSTRACT

Cross-language plagiarism detection aims to detect plagiarised fragments of text among documents in different languages. In this paper, we perform a systematic examination of Cross-language Knowledge Graph Analysis; an approach that represents text fragments using knowledge graphs as a language independent content model. We analyse the contributions to cross-language plagiarism detection of the different aspects covered by knowledge graphs: word sense disambiguation, vocabulary expansion, and representation by similarities with a collection of concepts. In addition, we study both the relevance of concepts and their relations when detecting plagiarism. Finally, as a key component of the knowledge graph construction, we present a new weighting scheme of relations between concepts based on distributed representations of concepts. Experimental results in Spanish–English and German–English plagiarism detection show state-of-the-art performance and provide interesting insights on the use of knowledge graphs.

1. Introduction

Given the vastness of the Web, plagiarism, or the deliberate use of someone else's original material without acknowledging its source, has become a serious problem in areas such as Literature, Education, and Science. The ease of access to copyrighted contents has become matter of concern also for researchers. The problem is exacerbated when the source of plagiarism comes from another language, which is known as cross-language (CL) plagiarism. It is not only the additional difficulty of manually detecting the translation performed, but also the people's lack of knowledge about the ethical issues derived from plagiarism. A recent survey about scholar practices and attitudes (Barrón-Cedeño, 2012) reveals that only 36.25% of students believe that translating text fragments and including them in their work is plagiarism.

Although the CL plagiarism detection task could be potentially performed manually, the amount of data, languages, and time required make it impossible to perform in practice. Current approaches to CL plagiarism detection exploit syntactic and lexical properties of the writing, statistical dictionaries or similarities with a multilingual collection of documents. However, most of these techniques are designed for verbatim copies and performance is reduced when dealing with light and especially heavy cases of plagiarism (Clough & Stevenson, 2011), which include paraphrasing.

* Corresponding author. Tel.: +34 644264447.
E-mail address: mfranco@prhlt.upv.es (M. Franco-Salvador).

http://dx.doi.org/10.1016/j.ipm.2015.12.004
0306-4573/

In a previous work, we proposed Cross-Language Knowledge Graph Analysis (CL-KGA) (Franco-Salvador et al., 2013), an approach for CL plagiarism detection aiming at representing context, which employs knowledge graphs both to expand and relate the concepts in a text. Knowledge graphs are generated using BabelNet (Navigli & Ponzetto, 2012a), the most large multilingual semantic network. Thanks to the multilingual representation of concepts available, BabelNet allows for a straightforward comparison of the knowledge graphs obtained in different languages.

In this work, we perform a systematic study of our CL-KGA model. We analyse the impact of the implicit aspects of knowledge graphs on CL plagiarism detection. The research questions we aim to answer are:

- *What is the contribution of the word sense disambiguation (WSD) performed by the knowledge graphs?* These graphs have been explored in the past to perform WSD; our current representation includes disambiguated concepts, which are combined with their intermediate concepts and other disambiguation candidates. We are interested in analysing the performance when the representation is exclusively composed by disambiguated words. This leads us to our next research question.
- *What is the contribution of the vocabulary expansion performed during graph creation?* In our previous work we assumed that the new intermediate concepts that relate the original ones could be a key component in order to obtain a common intersection between related texts. In this work we study this aspect in order to determine if the vocabulary expansion is needed as part of the representation or just as a component during the WSD process itself.
- *What is the relationship between CL-KGA and Cross-Language Explicit Semantic Analysis (CL-ESA)?* These two models represent text by exploiting a collection of multilingual concepts, for instance employing Wikipedia. We are interested in studying the similarities and the differences between the two models. We aim to clarify the particularities that make the two models perform completely different.

In this paper, we also address key aspects such as the language independence of the knowledge graphs. In addition, we study the relevance of the concepts (nodes) and relations (edges) of the knowledge graphs, and the most suitable threshold to consider that their weighted relations are semantically related. Finally, we compare our model with the state of the art according to different scenarios and criteria: (i) we evaluate CL plagiarism detection using a dataset composed by automatic and manually generated paraphrasing cases of plagiarism; (ii) we study the performance of detection using only paraphrasing cases; and (iii) we compare the computational efficiency of the models and the size of the graphs.

The classical weighting scheme used for the relations between the concepts of the knowledge graphs is based on bag of words generated from short concept definitions as representation of WordNet's concepts. Because it is exclusively based on the original wording of the definition, this type of representation is very explicit. In addition to the detailed study of our previous model, in this work we follow the recent and popular trend in the use of distributed representations of words (Mikolov et al., 2013a; Pennington et al., 2014), and present a new weighting scheme for relations between concepts which generates distributed representations of concepts. Our distributed concepts are generated using the continuous Skip-gram model to obtain vector representations of definitions of concepts. In contrast to the classical weighting, our proposed representation measures semantic relatedness modelling not only of the original words in a definition, but also their context. This allows our scheme to successfully measure similarity between definitions which do not share the same words but have the same meaning.

Experimental results show that the vocabulary expansion is more useful when it is only employed to perform the WSD, which is the essential component of our model. The differences between CL-KGA and CL-ESA are proved favouring the first model, which offers a higher performance thanks to the high coverage of BabelNet and the concept relatedness. Our new weighting scheme using distributed representations of concepts achieves state-of-the-art performance compared to the classical weighting and several alternative CL plagiarism detectors. The study with CL paraphrasing cases proved also CL-KGA superiority on this type of plagiarism. Finally, a comparison of the computational efficiency of the models demonstrated that our model is more adequate for systems that only require a fast document similarity and perform the indexing in a preprocessing stage.

The rest of the paper is organised as follows. In Section 2 we provide an overview of the state of the art in CL plagiarism detection and distributed representationsof concepts. In Section 3 we describe the knowledge graphs, their weighting schemes, including our new approach, and their main characteristics. In Section 4 we describe the CL-KGA model for CL plagiarism detection. Finally, in Section 5 we evaluate our approach for Spanish–English and German–English plagiarism detection, comparing our results with several state-of-the-art models. We compare also our new weighting scheme based on distributed representations of concepts with the classical weighting. As part of our analysis, we show the results when detecting only paraphrasing cases and evaluate the computational efficiency of the models.

2. Related work

In this section we first review the approaches of CL similarity analysis that have been used for CL plagiarism detection. Next, we summarise the last advances in the use of distributed representations for conceptual semantic relatedness.