Knowledge-Based Systems 000 (2016) 1–13

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language

Marc Franco-Salvador[a,1,*], Parth Gupta[a,1], Paolo Rosso[a], Rafael E. Banchs[b]

[a] Pattern Recognition and Human Language Technology (PRHLT) Research Center, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
[b] Institute for Infocomm Research, 138632, Singapore

ARTICLE INFO

Article history:
Received 13 January 2016
Revised 21 July 2016
Accepted 5 August 2016
Available online xxx

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

ABSTRACT

Cross-language (CL) plagiarism detection aims at detecting plagiarised fragments of text among documents in different languages. The main research question of this work is on whether knowledge graph representations and continuous space representations can complement to each other and improve the state-of-the-art performance in CL plagiarism detection methods. In this sense, we propose and evaluate hybrid models to assess the semantic similarity of two segments of text in different languages. The proposed hybrid models combine knowledge graph representations with continuous space representations aiming at exploiting their complementarity in capturing different aspects of cross-lingual similarity. We also present the continuous word alignment-based similarity analysis, a new model to estimate similarity between text fragments. We compare the aforementioned approaches with several state-of-the-art models in the task of CL plagiarism detection and study their performance in detecting different length and obfuscation types of plagiarism cases. We conduct experiments over Spanish-English and German-English datasets. Experimental results show that continuous representations allow the continuous word alignment-based similarity analysis model to obtain competitive results and the knowledge-based document similarity model to outperform the state-of-the-art in CL plagiarism detection.

1. Introduction

Automatic plagiarism detection refers to the task of automatically identifying which fragment of text is plagiarised. It involves finding plagiarised fragments f_q from a suspicious document d_q along with the source fragments f_s from a collection of source documents D. In the cross-language setting, suspicious and source documents are written in different languages [3,36]. This work aims at studying the combination of knowledge graph and continuous representation-based methods for the task of CL plagiarism detection.

There exist many approaches to CL plagiarism detection (CLPD) [2,13,14,19,39]. Current state-of-the-art approaches to CLPD are based on vector space model representations which operate in high dimensional spaces. Most of these approaches try to establish semantic similarity using external resources such as parallel data, comparable data, or semantic networks. The Knowledge-Based document Similarity (KBSim) model [15] combines relevance cues from knowledge graphs — generated by means of a multilingual semantic network —, and Vector Space Models (VSM) for capturing aspects of text such as out-of-vocabulary words and estimating CL similarity. However, KBSim has not yet been evaluated for CLPD. Compared with the VSM representation, latent semantic — continuous space — models have been shown to offer a higher performance when measuring text similarity [34]. The main research question of this work is on whether knowledge graph representations and continuous space representations can complement to each other and improve the state-of-the-art performance in CLPD. In this sense, we propose and evaluate hybrid models to assess the semantic similarity of two segments of text in different languages. The proposed hybrid models employ KBSim to combine knowledge graph representations with continuous space representations aiming at exploiting their complementarity in capturing different aspects of cross-lingual similarity. We analyse the quality of the KBSim model when it employs continuous models

* Corresponding author.
E-mail addresses: mfranco@prhlt.upv.es (M. Franco-Salvador), pgupta@dsic.upv.es (P. Gupta).

[1] The first two authors are ordered alphabetically and their contributions are the following: Marc Franco-Salvador implemented the state-of-the-art and designed the CL-KGA, KBSim, and CWASA models. He also carried out the evaluation of the models and contributed in the paper writing. Parth Gupta implemented S2Net, designed the BAE and XCNN continuous space models and contributed in the paper writing.

http://dx.doi.org/10.1016/j.knosys.2016.08.004
0950-7051/

such as Siamese Neural Network (S2Net) [44], Bilingual Autoencoder (BAE) [17,25], and eXternal data Composition Neural Network (XCNN) [18] for CL plagiarism detection. In addition, this study aims at filling the research gap of performance analysis of these continuous models for the CLPD task and also evaluates their independent performance. Finally, we investigate an alternative for continuous word composition when measuring similarity between texts. The Continuous Word Alignment-based Similarity Analysis (CWASA) employs directed word alignments on top of the continuous word representations to measure the distance between two texts.

We carry out experiments on standard plagiarism dataset PAN-PC-2011 for two languages Spanish-English (ES-EN) and German-English (DE-EN) in two settings: *i)* entirely plagiarised suspicious-source document linking (Expt. A); and *ii)* plagiarised fragments identification within entire documents (Expt. B). We also present an extensive analysis on performance of these algorithms for different lengths and types of plagiarism cases, and a study of the computational efficiency of the evaluated approaches. Our experiments show that, though continuous models have a small coverage (20k words) and have been trained on a parallel corpus of a limited size, they exhibit robust performance, especially when composed with CWASA, compared to VSM representations that have full coverage. Moreover, when combined together using KBSim, the performance is superior than the one of any other model alone. This points to the fact that knowledge graph and continuous-based models capture different aspects of cross-lingual similarity for CL plagiarism detection.

The rest of the paper is structured as follows: in Section 2 we present related work on cross-language plagiarism detection and continuous models for cross-language similarity estimation. We detail state-of-the-art methods for CL plagiarism detection and continuous representation-based models in Sections 3 and 4, respectively. Section 5 covers the details about the KBSim model. We present our experimental framework with results and analysis in Section 6. Finally, in Section 7 we draw some conclusions.

2. Related work

The Cross-Language Character *n*-Gram (CL-CNG) model [28] follows the architecture of some monolingual models for plagiarism [7,27]. It employs vectors of character *n*-grams to represent texts and uses a measure of similarity between vectors such as the cosine similarity to compare them. This model proved to be effective for Romance and Germanic languages that share lexical and syntactic similarities.

There exist several approaches designed to measure CL similarity between distant languages. The Cross-Language Explicit Semantic Analysis (CL-ESA) [39] model adapts the well-known ESA [16] architecture to represent texts by their similarities with a multilingual collection of documents. The use of a multilingual collection such as Wikipedia, with comparable documents across languages, allows to directly compare vectors generated from distinct languages.

Models based on parallel corpora have also been proposed. This type of corpora allows to create statistical bilingual dictionaries. The Cross-Language Alignment-based Similarity Analysis (CL-ASA) model [2,4,33] uses them to translate and align words. The alignments are based on the translation probabilities and also account for the difference in length of equivalent texts in distinct languages.

The use of multilingual knowledge resources or semantic networks have been explored too. The Cross-Language Conceptual Thesaurus based Similarity (CL-CTS) model [19] uses the Eurovoc conceptual thesaurus[2] to measure the similarity between texts in terms of the number of concepts and named entities that they share. With respect to CL-CNG and CL-ASA, it provided with an average performance and excelled for Spanish-English. On the other hand, the Cross-Language Knowledge Graph Analysis (CL-KGA) model [13,14] employs a multilingual semantic network to represent the context of documents by means of knowledge graphs. This representation includes characteristics such as word sense disambiguation, concept relatedness, or vocabulary expansion. This model excels even in cases with paraphrasing and represents the state of the art in CL plagiarism detection. In recent years, several improvements over the CL-KGA architecture made the Knowledge-Based document Similarity (KBSim) model [15] an interesting alternative for CL document retrieval and categorisation. This model complements knowledge graphs with a vector component to cover knowledge graph shortcomings such as out-of-vocabulary words and verbal tenses. However, KBSim has not yet been evaluated for CL plagiarism detection neither combined with continuous representations.

Recently, the Conference and Labs of the Evaluation Forum (CLEF) actively covered the plagiarism detection task in the framework of the evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN).[3] The shared task in plagiarism detection [38] provides with a corpus of plagiarism and allows participants to create systems in order to compete at detecting its plagiarism cases. The datasets of the 2010 and 2011 editions [35,37] included also German-English and Spanish-English CL partitions. The most popular methods to detect CL plagiarism at PAN followed [8] and used machine translation techniques to convert the problem into a monolingual one. However, this puts forward a heavy dependence on availability of Machine Translation (MT) systems in the involved languages and their quality. In addition, we believe that the nature of those methods is not purely cross-lingual and can be considered monolingual with a MT preprocessing. Therefore, all the models employed in this work for CL plagiarism detection do not depend on full MT systems. Nevertheless, in [3] authors show a comparison of the CL-ASA and CL-CNG models with an approach (T+MA) employing MT to analyse the similarities at monolingual level. That study shows that T+MA is superior in short cases of plagiarism but similar to CL-ASA, that always obtains a higher precision and better results for long cases of plagiarism. Considering that [3] included an evaluation setting very similar to ours (see Section 6), we decided to not include T+MA again in this work.

In [14] CL-KGA was compared with CL-CNG, CL-ESA and CL-ASA obtaining the highest results in Spanish-English and German-English plagiarism detection. In addition, a comparison of the CL-CNG, CL-ESA and CL-ASA models for CL plagiarism detection has been provided in [36]. Different performances were observed depending on the languages, and the dataset employed. For instance, CL-ESA and CL-CNG were more stable across datasets, obtaining a higher performance on the comparable Wikipedia dataset. In contrast, CL-ASA obtained better results on the parallel JRC-Acquis dataset. Finally, CL-CNG reduced the performance for language pairs without lexical and syntactic similarities. Therefore, for the sake of completeness, in this work we decided to compare our KBSim model based on continuous representations against the CL-CNG, CL-ESA, CL-ASA, and CL-KGA models.

With respect to the continuous space representations of texts, often referred to as embeddings, the advancement in the area has been quite limited. The high dimensional representation of text in vector space is projected into a low dimensional space

[2] http://eurovoc.europa.eu/.
[3] pan.webis.de/.

Please cite this article as: M. Franco-Salvador et al., Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.08.004