Using Sentence Embedding for Cross-Language Plagiarism Detection

Naif Alotaibi(✉) and Mike Joy

Department of Computer Science, University of Warwick, Coventry, UK
{naif.alotaibi,m.s.joy}@warwick.ac.uk

Abstract. The growth of textual content in various languages and the advancement of automatic translation systems has led to an increase of cases of translated plagiarism. When a text is translated into another language, word order will change and words may be substituted by synonyms, and as a result detection will be more challenging. The purpose of this paper is to introduce a new technique for English-Arabic cross-language plagiarism detection. This method combines word embedding, term weighting techniques, and universal sentence encoder models, in order to improve detection of sentence similarity. The proposed model has been evaluated based on English-Arabic cross-lingual datasets, and experimental results show improved performance when compared with other Arabic-English cross-lingual evaluation methods presented at SemEval-2017.

Keywords: Plagiarism detection · Semantic similarity · Cross-Language · English · Arabic

1 Introduction

Development of the Internet and information technology have increased the availability of digital libraries and automatic machine translation tools, where a text easily translates from one language to another language, and these have increased instances of plagiarism. Plagiarism occurs by copying words, phrases or ideas from someone else without giving acknowledgment to original work [8].

Zu Eissen et al. [20] presented a taxonomy of plagiarism, which was enriched by Alzahrani et al. [3], who classified plagiarism into *literal* and *intelligent* plagiarism. Literal plagiarism is word-for-word repetition of a phrase or transcription of a section of someone else's work. There are three types of this form, which are an *exact copy, near copy* and *modified copy*. Whereas, intelligent plagiarism is the changing of content in original text by modifying sentence structure such as paraphrasing or translating text into another language, and is referred to as *cross-language plagiarism.* Identification of translated plagiarism is more challenging than other types of plagiarism since each language has its own structure.

There exist a number of plagiarism detection approaches that are able to capture exact copy and simply modified plagiarism. However, these systems cannot effectively

M. Bramer and R. Ellis (Eds.): SGAI-AI 2020, LNAI 12498, pp. 373–379, 2020.
https://doi.org/10.1007/978-3-030-63799-6_28

detect more extensively disguised cases of plagiarism, including paraphrases and cross-language plagiarism. Eisa et al. [6] noted that existing methods are still struggling with the serious issues in identifying linguistic changes like substituting vocabulary by their synonyms. This paper proposes an English-Arabic cross-lingual plagiarism detection model, that is based on semantic sentence similarity. Effectiveness of word embedding and universal sentence encoding for representing sentence vectors are examined, and a model is proposed based on combining these approaches and using combinations of POS and TFIDF weighting schemes.

Several studies have been conducted on cross-lingual plagiarism detection. For example, a Cross-Lingual Alignment-based Similarity model (CL-ASA), that was presented in [4] used a parallel corpus to create a bilingual statistical dictionary. Another study based on comparable corpora was introduced in [9] and used a Cross-Language Explicit Semantic Analysis (CL-ESA) model that can be applied to corpora that contain texts that are written about similar topics in various languages. There has been little published research on Arabic-English cross-language plagiarism detection. Aljohani and Mohd [2] proposed an Arabic-English cross-lingual detection method based on winnowing algorithm, and used Google Translate to translate the texts. Although this model was able to detect literal plagiarism cases, it could not detect the cases of rewriting words using their synonyms. To overcome this, Hattab [10] employed Latent Semantic Indexing (LSI) to construct a cross-lingual semantic vector space in order to identify context similarity. The author used a parallel corpus to convert the source documents into target text instead of using direct translation, using cosine measurement to calculate degree of similarity. The method gave good results in the cases of replacing words, however the computational procedure of LSI is relatively expensive. Another study [1] introduced a technique based on key phrase extraction from suspect documents and then translated these phrases via machine translation. Thereafter, the similarity between these phrases was measured by a combination of three techniques: cosine similarity, longest common subsequence (lcs) and n-grams. Even though the model worked quite well, computational complexity of using the lcs method has an impact on the overall performance. Recently, Ezzikouri et al. [7] have applied a fuzzy semantic based similarity approach in order to capture cross language plagiarism cases utilizing WordNet and the algorithm that proposed in [18] in order to compute semantic similarity between two words.

On the other hand, word embedding is an approach to provide a distributed representation of vocabularies. There are number of methods which have been introduced to generate word embedding from text data, for example, two methods were offered in [12] to build the words representations model: (i) Continuous bag-of-words (CBOW) and (ii) skip-gram (SKIP-G). The CBOW model predicts the current word based on surrounding words, whereas the second model uses the current word to predict the neighboring words. Furthermore, the Universal Sentence Encoder embeds sentences into vector representations which capture rich semantic information that can be used in variety of natural language processing (NLP) applications, like classification and plagiarism detection [19]. The proposed model for detecting English-Arabic cross-lingual plagiarism is based on analyzing features of sentences using word embedding and multilingual universal sentence encoder (MUSE) models.

2 Proposed Method

The key idea for the proposed plagiarism detection technique for English-Arabic pairs of texts is based on sentence level comparisons. The proposed model is based on word embedding, term weighting, and MUSE. There are two steps in order to represent sentence vectors. Firstly, we combine word embedding and mixing parts of speech (POS) with the Term Frequency Inverse Document Frequency (TFIDF) weighting method, which we name CL-WE-Tw, and is based on machine translation. The second step is to combine the MUSE model with the CL-WE-Tw method in order to enhance detection of sentence similarity. Figure 1 illustrates the framework for the proposed method, and the main processes are shown in the following section.

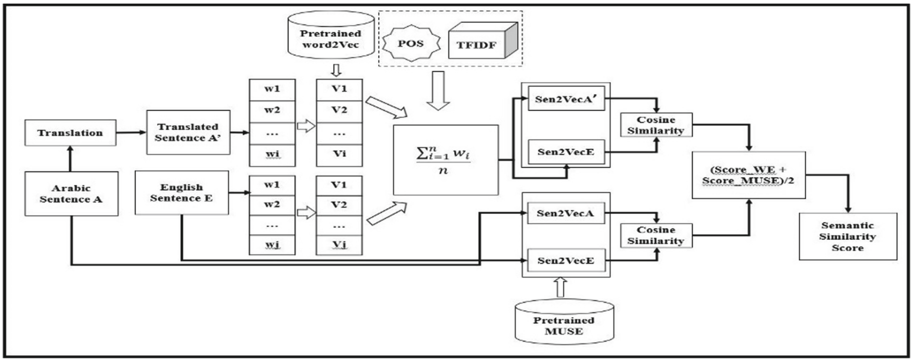

Fig. 1. Proposed framework for English-Arabic cross-language plagiarism detection

2.1 Sentence Embedding Based on CL-WE-Tw Model

Representation of sentences based on word embedding and term weight schemes is a useful way to extract features of each sentence, thus enhancing the ability to compute sentence similarity. Word embedding captures semantic and syntactic features of the language [13]. Term weighting schemes such as TFIDF and POS are methods to assign for each term its weighted contribution in the text. The key idea of TFIDF is to find relations of a word in a document to other documents in the corpus, and it is used to reduce the influence of the most common words such as "is", "the" and "a". According to [14], the TFIDF weighting scheme is integrated traditionally with information retrieval for enhancing textual retrieval performance. In terms of the POS weighting approach, which is beneficial to understand the sentence representation, POS is able to take into account ambiguity problems, such as the word "train" which can be a noun or a verb. The usefulness of combining the POS weighting scheme with information retrieval is to improve retrieval performance [11]. As a result, the proposed method integrates word embedding with both POS and TFIDF in order to represent sentences and then compute the similarity between two sentence vectors. Each word is represented as a vector and is weighted by mixing the TFIDF and POS weights. The weighted average of all word vectors is used to construct the corresponding sentence vector, as shown in Eq. 1.

$$Sv = \frac{1}{n}\sum_{i=1}^{n} vi * (TFIDF * POS(wi)) \tag{1}$$

Where Sv is sentence vector, vi is a function that finds word vector, wi is the i^{th} word of text. After getting each sentence vector, cosine similarity is applied to compute similarity between two texts according to Eq. 2:

$$Ss = \frac{VE.VA'}{\|VE\|.\|VA'\|} \tag{2}$$

Where VE is the first sentence vector (in English sentence) and VA' is the sentence vector (translated from Arabic).

2.2 Sentence Embedding Based on MUSE

MUSE is a universal sentence encoder which is used to convert sentences into vector representations. The benefit of these vector representation is to extract a high level of descriptive features [19]. Two pre-trained models for semantic text similarity have been released, these models are based on transformer and convolutional neural network (CNN) model architectures [19]. The MUSE model allows the representation of sentences from different languages into a single vector space, where it is possible to find similarities between sentences that are written in different languages directly. This approach therefore proposes to use the MUSE model to detect English-Arabic cross-language plagiarism, performing a direct comparison between English and Arabic sentences and then applying cosine similarity between them to measure the degree of similarity.

2.3 Overall Sentence Similarity

As CL-WE-Tw and MUSE models are two important components for interpreting sentence meaning, the overall sentence similarity is measured by a combination of sentence similarity based on the CL-WE-Tw and MUSE models.

$$Sim_{sentence} = \frac{(s_{we} + s_{muse})}{2} \tag{3}$$

In Eq. 3, s_{we} is the result is obtained by sentence embedding based on CL-WE-Tw while s_{muse} is obtained based on the MUSE model. After obtaining the similarity score between the pair of sentences based on the proposed model, it can be judged whether this pair is plagiarized or non-plagiarized. Namely, if the degree of similarity exceeds a predefined threshold α, the pair of sentences is considered plagiarized.

3 Experiment and Results

In order to assess the proposed model, STS Test and Microsoft Research Video Description Corpus (MSRvid) datasets drawn from the Semantic Textual Similarity (STS) shared task from SemEval-2017 (STS Cross-lingual Arabic-English), released in [5], are used to assess the performance of the model. The total size of these datasets is 985 pairs of sentences, each pair of sentences having been labelled by humans on an integer scale from 0–5 (5 means exactly similar, whereas 0 indicates that the two sentences in the pair are completely different). The Pearson correlation coefficient P between the human rating and the predicted value of the model is used to assess the performance of the proposed model.

The proposed model consists of two components, the first is machine translation and the second is applying monolingual semantic analysis based on word embedding and mixing TFIDF with POS weightings. A pre-processing phase is required for making text is ready for further evaluation, and consists of the following steps. Firstly, all sentences are translated from Arabic into English via the Google Translation API. Secondly, Natural Language ToolKit (NLTK) is used for tokenization, POS tagging, removing punctuation marks, and normalization. After the texts are pre-processed, Mikolov et al.'s [13] pre-trained word2vec model, which is efficient to extract semantic and syntactic features, is used to generate a vector for each word in each sentence. The word2vec model was trained on a Google News dataset of about 100 billion words. On the other hand, the second component of the proposed model, which is based on MUSE model, is not required for the pre-processing steps. After representing sentence vector, the degree of the similarity between the pair of sentences is calculated using cosine similarity.

A number of experiments have been done in order to examine the performance of representing sentence embedding based on the CL-WE-Tw and MUSE models then measuring semantic similarity between two sentence vectors. Table 1 shows the results of the proposed model.

Table 1. Assessment results of proposed model

Datasets Methods	STS Test	MSRvid
Word2vec Model		
Average all vectors	0.6204	0.7269
Average all vectors & TFIDF	0.6693	0.7718
Average all vectors & POS	0.6801	0.7460
CL-WE-Tw	0.6902	0.7732
MUSE Model		
MUSE model	0.78	0.7977
Combination of CL-WE-Tw and MUSE Models		
((CL-WE-Tw) + (MUSE model))**/2**	0.8147	0.837

As displayed in Table 1, integrating the word2vec model with the POS and TFIDF weighting scheme achieves good results on both the STS test dataset and the MSRvid dataset with Pearson correlations of 0.6902 and 0.7732 respectively. Computing similarity between the two sentence vectors based on the MUSE model achieves the highest results for both datasets with correlations of 0.78 and 0.7977 respectively. Interestingly, combining the CL-WE-Tw and MUSE models achieved better performance than using them individually.

The proposed models have been compared with the ECNU [16], BIT [17] and HCTI [15] methods that obtained the best results on the STS Test dataset. Table 2 presents the comparative evaluation.

Table 2. Comparative evaluation

Models	Pearson correlation coefficient
The proposed model	0.8147
ECNU	0.7493
BIT	0.7007
CL-WE-Tw	0.6902
HCTI	0.6836

As shown in Table 2, the proposed model based on sentence embedding obtains the highest performance with a correlation of 0.8147.

4 Conclusion

This paper proposed a technique for detecting cross-lingual plagiarism based on combining word embedding, term weighting and the MUSE models. According to the results of the experiments, the proposed model is competitive when compared against other participating approaches in the SemEval-2017 Arabic-English cross-lingual evaluation task. For future work, we will explore the use of machine learning algorithms (e.g., decision tree and random forest) and neural network architectures (e.g., LSTM and RNN) in order to enhance cross-lingual plagiarism detection.

References

1. Alaa, Z., Tiun, S., Abdulameer, M.: Cross-language plagiarism of Arabic-English documents using linear logistic regression. J. Theor. Appl. Inf. Technol. **83**(1), 20–33 (2016)
2. Aljohani, A., Mohd, M.: Arabic-English cross-language plagiarism detection using winnowing algorithm. Inf. Technol. J. **13**(14), 23–49 (2014)
3. Alzahrani, S.M., Salim, N., Abraham, A.: Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**(2), 133–149 (2011)
4. Barrón-Cedeno, A., Rosso, P., Pinto, D., Juan, A.: On cross-lingual plagiarism analysis using a statistical model. In: PAN, pp. 1–10 (2008)
5. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: Semeval-2017 task 1: semantic textual similarity-multilingual and cross-lingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 1–14 (2017)

6. Eisa, T.A.E., Salim, N., Alzahrani, S.: Existing plagiarism detection techniques: a systematic mapping of the scholarly literature. Online Inf. Rev. **39**(3), 383–400 (2015)
7. Ezzikouri, H., Oukessou, M., Youness, M., Erritali, M.: Fuzzy cross language plagiarism detection (Arabic-English) using WordNet in a big data environment. In: 2nd International Conference on Cloud and Big Data Computing, pp. 22–27. ACM (2018)
8. Gipp, B.: Citation-based Plagiarism Detection, pp. 57–88. Springer, Wiesbaden (2014). https://doi.org/10.1007/978-3-658-06394-8_4
9. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. IJCAI **7**, 1606–1611 (2007)
10. Hattab, E.: Cross-language plagiarism detection method: arabic vs. english. In: 2015 International Conference on Developments of E-Systems Engineering (DeSE), pp. 141–144. IEEE (2015)
11. Lioma, C., Blanco, R.: Part of speech based term weighting for information retrieval. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 412–423. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00958-7_37
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: 26th International Conference on Neural Information Processing Systems, vol. 2, pp. 3111–3119 (2013)
14. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. **24**(5), 513–523 (1988)
15. Shao, Y.: HCTI at SemEval-2017 Task 1: use convolutional neural network to evaluate semantic textual similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, pp. 130–133 (2017)
16. Tian, J., Zhou, Z., Lan, M., Wu, Y.: ECNU at SemEval-2017 task 1: leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, pp. 191–197 (2017)
17. Wu, H., Huang, H.Y., Jian, P., Guo, Y., Su, C.: BIT at SemEval-2017 task 1: using semantic information space to evaluate semantic textual similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, pp. 77–84 (2017)
18. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133–138 (1994)
19. Yang, Y., et al.: Multilingual universal sentence encoder for semantic retrieval. arXiv preprint arXiv:1907.04307 [CS.CL] (2019)
20. Meyer zu Eissen, S., Stein, B., Kulig, M.: Plagiarism detection without reference collections. In: Decker, R., Lenz, H.-J. (eds.) Advances in Data Analysis. SCDAKO, pp. 359–366. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-70981-7_40