English-Arabic Cross-language Plagiarism Detection

Naif Alotaibi
Department of Computer Science
University of Warwick
naif.alotaibi@warwick.ac.uk

Mike Joy
Department of Computer Science
University of Warwick
m.s.joy@warwick.ac.uk

Abstract

The advancement of the web and information technology has contributed to the rapid growth of digital libraries and automatic machine translation tools which easily translate texts from one language into another. These have increased the content accessible in different languages, which results in easily performing translated plagiarism, which are referred to as "cross-language plagiarism". Recognition of plagiarism among texts in different languages is more challenging than identifying plagiarism within a corpus written in the same language. This paper proposes a new technique for enhancing English-Arabic cross-language plagiarism detection at the sentence level. This technique is based on semantic and syntactic feature extraction using word order, word embedding and word alignment with multilingual encoders. Those features, and their combination with different machine learning (ML) algorithms, are then used in order to aid the task of classifying sentences as either plagiarized or non-plagiarized. The proposed approach has been deployed and assessed using datasets presented at SemEval-2017. Analysis of experimental data demonstrates that utilizing extracted features and their combinations with various ML classifiers achieves promising results.

1 Introduction

The advancement of the Internet and information technology have expanded rapidly the availability of digital libraries and automatic machine translation tools, which facilitate translating a text from one language to another language. This has caused the number of cases of translated plagiarism, referred to as "cross-language plagiarism", to perform substantially. It is a type of plagiarism that occurs when textual content is translated into another language without giving acknowledgment of original sources. This type of plagiarism is more difficult to detect since each language has its own structure.

Several plagiarism detection techniques have been proposed to address monolingual plagiarism, that identify plagiarism instances written in the same language. However, there have been few studies that concentrate on researching and developing methods for identifying cross-language (and in particular English-Arabic) plagiarism. These techniques cannot effectively detect more extensively disguised cases of cross-language plagiarism. Eisa et al. (2015) observed that existing techniques have difficulty detecting linguistic modifications like replacing words and phrases by synonyms. When a text is translated from Arabic into English, synonyms are introduced after the translation, thus it is difficult to identify plagiarism.

This paper proposes a new approach for enhancing English-Arabic cross-language plagiarism detection at the sentence level. This technique is based on semantic and syntactic feature extraction using word alignment, word order, word embedding and multilingual encoder models. We investigate the effectiveness of using those features and their combination with different machine learning (ML) algorithms for classifying sentences as either plagiarized or non-plagiarized. The rest of this paper is structured as follows: Section 2 presents a review of related work on cross-language plagiarism detection techniques. In Section 3 we illustrate the proposed approach. The experimental results and discussion are provided in

Proceedings of Recent Advances in Natural Language Processing, pages 44–52
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_006

Section 4. Finally, Section 5 concludes and presents future work.

2 Related Work

A number of studies have been scrutinized cross-language plagiarism detection. Potthast et al. (2011) presented a classification of cross-language similarity detection methods which was subsequently developed by Danilova (2013). These approaches were classified on the basis of the mechanism used for detecting similarity as shown in Figure 1.

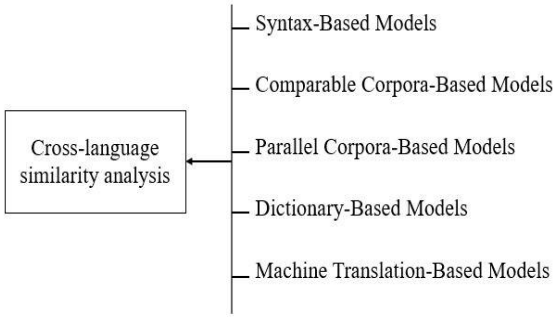

Figure 1: Classification of various techniques for cross-language similarity analysis (Potthast et al., 2011; Danilova, 2013).

For an instance, Cross-Language Character n-Grams (CL-CNG), which were presented by McNamee and Mayfield (2004), segmenting texts into n-grams for performing comparisons between pairs of texts and measuring the similarity without translation. Another study was on the basis of comparable corpora and was presented by Potthast et al. (2008), who introduced the Cross-Language Explicit Semantic Analysis (CL-ESA) model that were used Wikipedia for computing the similarity between pairs of documents in different languages. For parallel corpora, Barrón-Cedeño et al. (2008) offered Cross-Lingual Alignment-based Similarity model (CL-ASA), creating a bilingual unigram dictionary for comparing pairs of texts. Gupta et al. (2012) introduced the dictionary-based Cross-Language Conceptual Thesaurus Similarity model (CL-CTS) which detects similarity between texts from different languages. Franco-Salvador et al. (2013) introduced a technique based on knowledge graphs for comparing between documents in different languages. Barrón-Cedeño (2013) presented a machine translation model to convert texts into the common language followed by employing a monolingual analysis.

Some published research has focused on English-Arabic cross-language plagiarism detection. For example, Aljohani and Mohd (2014) proposed an English-Arabic cross-language detection approach based on Google Translate to translate the texts and applying a winnowing algorithm, which proposed by Schleimer et al. (2003). Another study presented by Hattab (2015) proposed a technique based on Latent Semantic Indexing (LSI) and parallel corpora to build a cross-language semantic vector space to compute similarity of the context. Alaa et al. (2016) used a logistic regression classifier based on longest common subsequence and cosine similarity measurements and n-gram features at keyphrase level. A study utilized semantic metrics and WordNet for gauging the degree of semantic similarity between words and used it to calculate the similarity for texts and paragraphs of English-French and English-Arabic plagiarism instances Hanane et al. (2016). Ezzikouri et al. (2018) employed a fuzzy semantic approach to identify cross-language plagiarism cases employing Wu and Palmer's (1994) similarity metrics and WordNet to compute semantic similarity between words.

Based on this review, we have only identified a few studies which have attempted to detect cross-language plagiarism in the English-Arabic domain. Most of these studies have tried to identify plagiarism based on semantic features and key phrases. To the best of our knowledge, none of these studies has tried to detect plagiarism using English-Arabic pairs based on sentence level analysis, nor has any integrated semantic and syntactic features using word embedding and word alignment features with multilingual encoder models.

3 Proposed Method

The key idea of the proposed plagiarism detection technique for English-Arabic pairs of sentence is formulated as a classification task, which classifies each pair of sentences as either plagiarized or non-plagiarized. In order to tackle this problem, it is necessary to analyze texts using different features extracted at syntactic and semantic levels. Thus, we propose methods based on word embedding, word order and word alignment with multilingual

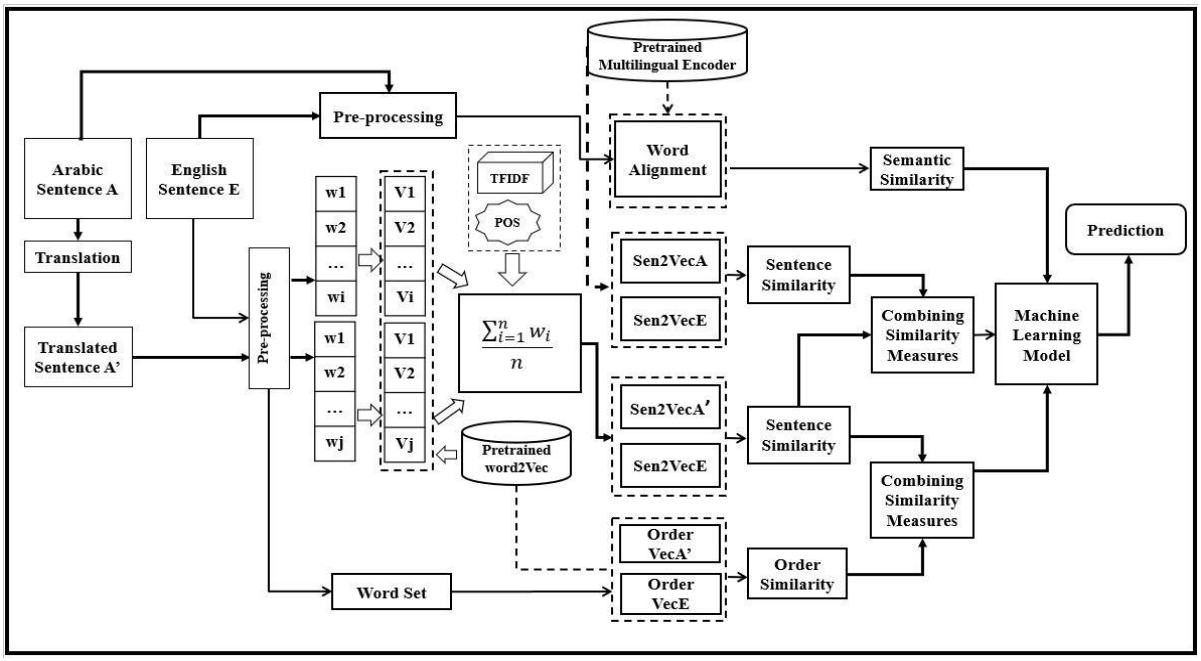

Figure 2: The proposed English-Arabic cross-language detection approach

encoders to extract features and then use them as features for supervised machine learning algorithms. The general framework of the proposed approach is described in Figure 2.

3.1 Feature Extraction

Analysis of semantic and syntactic features forms an essential step for plagiarism detection algorithms. Various sets of extracting features are proposed depending on word embedding, word order and word alignment for pair of sentence comparison. The following subsection describes the extracting features.

3.1.1 Word Order Similarity Features

Syntactic features based on word order are employed in similarity and plagiarism detection algorithms such as those by Li et al. (2006) and Abdi et al. (2015). Therefore, we propose a method that relies on word order features based on machine translation, since word order exhibits beneficial information about the relationships between words. In the case where two sentences have exactly the same words, but in a different order, any approach that measures similarity between texts based on a "bag of words" will show them to be exactly the same. Consequently, the influence of the word order should be taken into consideration when text similarity is computed. Thus, we are motivated by Li et al.'s (2006) approach. However, the proposed method is based on a pre-trained word2vec model released by Mikolov et al. (2013), representing words as vectors that characterize identification of semantic and syntactic features.

In order to gauge word order similarity between pairs of sentences, it is required to convert words into vectors based on a joint word set, which is formed utilizing distinct words from each pair of sentences. For example, given a pair of sentences T1 "A quick brown dog jumps over the lazy fox" and T2 "A quick brown fox jumps over the lazy dog", a joint word set T contains all distinct.
Words from T1 and T2, so T is {A quick brown dog jumps over the lazy fox}. Each word in T1 and T2 has an assigned unique index number, representing the word's location in the sentence. A word order vector is created for each sentence (r1 and r2 respectively), based on word embedding and the joint word set T. Taking T1 as an example, for each word in T, we look for the same or the most similar word in T1 as the following.

1. If the word exists in T1, the value for this word in r1 will take the same index number from T1.

2. If the word does not appear in T1, then we use the pre-trained word2vec model for finding the most similar word using

based on computing cosine similarity between the words. If the similarity score is greater than the predefined threshold (wt), the value of the word in r1 is set to the index number of the word in T1.

3. If the above two processes fail, the value of the index number in r1 is set to 0.

The procedure used for creating r1 will be applied for creating r2, which represents the second sentence. Therefore, word order vectors are constructed as:

r1= [1 2 3 4 5 6 7 8 9]
r2= [1 2 3 9 5 6 7 8 4]

Then, the word order similarity is calculated by using Equation 1.

$$OrderSim(r1, r2) = 1 - \frac{\|r1 - r2\|}{\|r1 + r2\|} \quad (1)$$

3.1.2 Sentence Embedding Features

Due to the detection being based at sentence level, extracting features from a pair of sentence uses the technique proposed by Alotaibi and Joy (2020). They proposed an approach for calculating the degree of semantic similarity between two sentences. The authors leverage models that represent sentences embedding, including universal sentence multilingual encoder (MUSE) and averaging word embedding, for constructing sentence vectors. They represent sentence embedding based on: (i) word embedding and term weight schemes (i.e., term frequency inverse document (TFIDF) and part of speech (POS)), referred to as CL-WE-Tw, and (ii) the MUSE model. Based on the methods, computing the degree of semantic similarity between two sentences is by following these steps:

- **Sentence vector based on CL-WE-Tw**: represents vectors for each sentence by taking the average vectors with their weighting according to Equation 2

$$Sv = \frac{1}{n}\sum_{i=1}^{n} vec * \left(TFIDF * POS(wi)\right) \quad (2)$$

where Sv is sentence embedding, vec is a function that gets word vector, wi is the i^{th} word of text.

- **Sentence Embedding based on MUSE**: uses a pre-trained model released by Yang et al. (2019) to represent sentence vectors then cosine similarity is employed to measure semantic similarity between pair of vectors as shown in Equation 3.

- **Semantic similarity measure:** after representing vectors for each sentence, cosine similarity is applied to find the degree for pairs of sentence according to Equation 3.

$$Ssim(veE, vcA') = \frac{veE.veA'}{\|veE\|.\|veA'\|} \quad (3)$$

where Ssim is sentence similarity that calculated using cosine similarity on sentence embedding, veE is the sentence vector for the English sentence, and veA' is the sentence vector translated from the Arabic sentence.

- Finally, the authors proposed to integrate semantic similarity features, obtained from the CL-WE-Tw and MUSE methods given by Equation 4.

$$Sim\_sentence = \frac{(S_{CL-We-Tw} + S_{MUSE})}{2} \quad (4)$$

where $S_{CL\text{-}WE\text{-}Tw}$ is the similarity score obtained from CL-WE-Tw method, and S_{MUSE} is obtained from the MUSE model.

3.1.3 Combined Similarity Measures

Since semantic and syntactic features play an important role in interpreting the meaning of a sentence, we propose to combine all sentence similarity measure features, which are described in Section 3.1.1 and CL-WE-Tw, and refer to it as "CL-WET-WO", as shown in Equation 5.

$$S(T1, T2) = \delta Ssim + (1 - \delta)OrderSim \quad (5)$$

Li et al., (2006) suggested that $0.5 < \delta \leq 1$, should be the threshold for weighting significance between components based on word order (OrderSim) and CL-WE_Tw (Ssim).

3.1.4 Word Alignment Features

The word alignment features are employed in different natural language processing tasks like sentence similarity (Sultan et al., 2015) and

paraphrase identification (Mohammad et al., 2017). Therefore, we propose to use semantic based features depending on word alignment. The proposed method is based upon the word alignment approach of Michlase et al. (2006) and Zhou et al. (2019), however, the difference is to use the pre-trained multilingual encoder model, such as that released by Yang et al. (2019) as a bilingual resource. It provides rich semantic information and enables the representation of words from different languages (e.g., English and Arabic) in a single vector space, where it directly determines similarity between words that are written in different languages. Such words are aligned according to their semantic similarity in the model, and cosine similarity is applied to find the similarity between pairs of words. The proposed method consists of two components, that can be used to describe pairs of sentences. The first component finds the similarity score between pairs of sentences as shown in Equation 6, which we call cross-language weighted alignment (CL-WA). This component consists of two processes.

$$S = \frac{1}{2}\left(\frac{\sum_{w\,\epsilon\{T1\}}\left(maxsim(w,T2)*idf(w)\right)}{\sum_{w\,\epsilon\{T1\}} idf(w)} + \frac{\sum_{w\,\epsilon\{T2\}}\left(maxsim(w,T1)*idf(w)\right)}{\sum_{w\,\epsilon\{T2\}} idf(w)}\right) \quad (6)$$

To compute the semantic similarity between two sentences T1 and T2, we use the pre-trained multilingual encoder model instead of a monolingual dictionary, then cosine similarity is employed for measuring similarity between pairs of words. The following steps are used.

1. According to Equation 7, for each term in sentence T1 we determine its aligned word in the sentence T2 which gets the highest semantic similarity and is greater than the threshold (t_1). This threshold is suggested to avoid excessive noise that leads to deterioration of overall performance. For example, when we align word "يضع", which means 'put' in English language, with other words like "dance", "put" and "cook", we find their vectors such [(vector (dance), vector (يضع)), (vector (put), vector (يضع)), (vector (cook), vector (يضع)], then we determine the maximum degree of semantic similarity between their vectors by applying cosine similarity.

$$\max_{\text{sim(wi,T2)}} = \text{maxsim}(wi, T2) \quad (7)$$

2. Determine the importance of words in T1 using inverse document frequency (idf).

The same process is employed to determine the most similar word in T1 beginning with words in T2. Finally, the similarity between the input sentence T1 and T2 is computed using Equation 6.

The second component is to calculate the semantic similarity for given two sentences T1 and T2 according to Equation 8, which we call "cross-language alignment (CL-A)". The process is as follows.

1- For each term in an Arabic sentence, we try to determine the word in the English sentence that has the highest semantic similarity (i.e., using the pre-trained multilingual encoder model and employing cosine similarity) that is greater than the threshold (t_2). This threshold is suggested to avoid excessive noise which causes deterioration of overall performance.

2- Finally, we take the average score over all the maximal similarity scores as given by Equation 8.

$$ssim(S,T) = \frac{1}{m}\sum_{j=1}^{m} \max_{i=1,\ldots,n} Cos(si, tj) \quad (8)$$

Finally, the overall sentence similarity score is computed based on CL-WA and CL-A, which we name "CL-WA+CL-A", as shown in Equation 9.

$$Sentence_{sim} = th * S + (1 - th)ssim \quad (9)$$

where the value of th value ranges within [0.5,1], and is the threshold for weighting importance between components based on CL-WA (S) and CL-A (ssim).

3.2 Classification Model

Machine learning algorithms have been applied in several fields, such as image processing and natural language processing. We use the extracted features, based on syntactic and semantic computation, along with different ML classification frameworks for detecting whether an English-Arabic pair of sentences is plagiarized or not. We investigate

different ML classifiers such as Logistic Regression (LR), Support Vector Classifier (SVC), Linear Support Vector Classifier (LSVC), Decision Tree (DT), Random Forest (RF), K Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost), using those extracted features.

4 Experiments and Results

To assess the performance of the proposed techniques, we have conducted experiments to examine the impact of both individual and combined features used to train each classifier. For evaluating the performance of the classifiers, we have used 10-fold cross-validation and the F1-measure (F1 score), which is the harmonic average of precision and recall, as shown in Equation 10. The experiments have been carried out using Python with the scikit- learn library to build each classifier, and we have used the Grid Search method to find the best values of hyperparameters for configuration of the ML models.

$$F - meaure = \frac{2 * precision * recall}{precision + recall} \qquad (10)$$

4.1 Dataset

We used SemEval-2017 (Cross-lingual Arabic-English) datasets, released by Cer et al. (2017). The total size of the dataset is 1234 pairs of sentences, which were used for both training and testing data. Humans have labeled each pair of sentences on an integer scale from 0-5 (5 indicates exactly similar, whereas 0 shows that the two sentences in the pair are completely different), which was linearly scaled into the interval [0,1] then each pair of sentences was labeled 1 (means plagiarized) or 0 (means non-plagiarized) if the human similarity score is greater than or equal to a threshold of 0.5. Table 1 illustrates more information about the dataset.

Datasets	Source	Pairs
MSRvid	MSR-Video, Microsoft Research Video Description Corpus	735
SNLI	Stanford Natural Language Inference corpus	250
SMTeu	WMT2008 development	149
MSR-Para	Microsoft Research Paraphrase Corpus	100

Table 1: Evaluation dataset

4.2 Pre-processing Stage

The pre-processing phase is an essential step for preparing the text for further evaluation. As the first two extracted features described in 3.1.1 and CL-WE_Tw are based on machine translation, we used the Google Translation tool to translate Arabic sentences into English sentences. Then, we used the Natural Language Toolkit (NLTK) tool for the following processing: (1) tokenization, (2) part of speech tagging, (3) removing punctuation marks and (4) normalization. On the other hand, the extracted features described in 3.1.4 are based on a pre-trained multilingual encoder model, which represents different languages on the same vector space, therefore, we used NLTK for performing: (1) tokenization, (2) removing punctuation marks and (3) removing stop words for both English (e.g. 'that', 'is' and 'were') and Arabic such as 'في', 'الى' and 'التي', meaning "in", "to" and "that" in English respectively.

4.3 Parameters Setting

The proposed methods of extracting features contain a number of parameters that are required to be tuned parameters for constructing word order vectors, for weighting the importance between syntactic and semantic features, and for word alignment features. Many experiments are performed to determine a suitable value for each parameter. For setting these parameters, we have used pairs of sentences from the Microsoft Research Video Description Corpus dataset, and computed Pearson correlation coefficient between human rates and results obtained from proposed approaches, thus the best Pearson correlation indicates suitable values for these parameters. Therefore, the results acquired from the experiments show the suitable values of the parameters, and we have found the best correlation coefficient values for determining word order similarity is achieved at (wt= 0.54). For weighting importance of syntactic and semantic similarity features, we have obtained the best results at (δ=0.80). In terms of the parameters related to word alignment, the best performance is attained at (t_1= 0.53, t_2= 0.40 and th=0.70). As a result, we

Features \ Classifier	LR	SVC	LSVC	DT	RF	KNN	XGBoost
1- Word order	0.786	0.776	0.783	0.716	0.728	0.756	0.764
2- Word Embedding	0.839	0.845	0.839	0.770	0.770	0.824	0.834
3- CL-WET-WO	0.847	0.842	0.846	0.766	0.766	0.833	0.845
4- CL-WE-Tw+MUSE	0.8561	0.859	0.857	0.812	0.813	0.851	0.865
5- CL-WA	0.812	0.815	0.810	0.743	0.750	0.784	0.817
6- CL-A	0.768	0.772	0.766	0.689	0.711	0.753	0.788
7- CL-WA + CL-A	0.818	0.822	0.826	0.743	0.755	0.804	0.821
8- Features (3 and 7)	0.853	0.850	0.857	0.789	0.814	0.846	0.844
9- Features (3, 4 and 7)	0.871	**0.879**	0.875	0.853	0.861	0.852	0.864

Table 2: Classification results based on set of extracted features.

have used these parameters and values on the rest of the dataset.

4.4 Results

This section shows the contribution of using extracted features, described in Section 3.1, along with a set of classifiers for detecting cross-language plagiarism. Table 2 shows the performance results of utilizing the proposed features along with LR, SVC, LSVC, DT, RF, KNN and XGBoost classifiers, where the first column illustrates the extracted features while the rest of the columns presents the results according to the F1 Score metric.

4.5 Discussion

As presented in Table 2, the performance of the classifiers using sets of extracted features shows encouraging results for classifying pairs of English-Arabic sentences. We can see that the integration of semantic and syntactic features with the classifiers as one feature, based on word embedding and word order features, demonstrates enhancement of the performance through LR, LSVC, KNN and XGBoost. Furthermore, it can be observed that using CL-WA+CL-A features along with the different classifiers obtained better results than CL-WA and CL-A individually. It can be also seen that combinations of features based on CL-WET-WO and CL-WA+CL-A with the classifiers show improvements in the results. Interestingly, using all combined features based on CL-WET-WO, CL-WA+CL-A and CL-WE-Tw+MUSE is efficient in enhancing the performance of most classifiers including LR, SVC, LSVC, DT and RF. We believe this improvement can be ascribed to the word embedding and multilingual encoder models capturing semantic and syntactic features.

5 Conclusion

In this paper we introduced a technique based on analyzing sentences using syntactic and semantic features with ML classifiers to detect English-Arabic cross-lingual plagiarism. The features we used involve word order, word embedding and word alignment with multilingual encoders. We also explored the effects of using extracted features and their combinations along with the different classifiers. The proposed method has been assessed by using a compilation of four datasets. According to the evaluation, the integration of combined extracted features with the classifiers demonstrates improved performance. Overall, the SVC classifier based on combination of all features accomplishes the best results with the F1score of 0.879. In future work, the approach will be expanded to include use of neural network techniques.

References

Aljohani, A., & Mohd, M. (2014). Arabic-English Cross-language Plagiarism Detection using

Winnowing Algorithm. *Information Technology Journal, 13*(14), 2349. https://doi.org/10.3923/itj.2014.2349.2355

Abdi, A., Idris, N., Alguliyev, R.M. & Aliguliyev, R.M. (2015). PDLK: Plagiarism detection using linguistic knowledge. *Expert Systems with Applications, 42*(22), 8936-8946. https://doi.org/10.1016/j.eswa.2015.07.048

Alotaibi, N. & Joy, M. (2020). Using Sentence Embedding for Cross-Language Plagiarism Detection. *In International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 373-379). Springer.

Alaa, Z., Tiun, S. & Abdulameer, M. (2016). Cross-language plagiarism of Arabic-English documents using linear logistic regression. *Journal of Theoretical & Applied Information Technology, 83*(1), 20-33.

Barrón-Cedeño, A., Gupta, P. & Rosso, P. (2013). Methods for cross-language plagiarism detection. *Knowledge-Based Systems, 50*, 211-217. https://doi.org/10.1016/j.knosys.2013.06.018

Barrón-Cedeno, A., Rosso, P., Pinto, D. & Juan, A. (2008). On Cross-lingual Plagiarism Analysis using a Statistical Model. *PAN, 212,* 1-10. https://webis.de/events/pan-07/pan07-web/

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I. & Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Danilova, V. (2013). Cross-language plagiarism detection methods. In *Proceedings of the Student Research Workshop associated with RANLP 2013* (pp. 51-57).

Eisa, T.A.E., Salim, N. & Alzahrani, S. (2015). Existing plagiarism detection techniques: A systematic mapping of the scholarly literature. *Online Information Review, 39* (3), 383-400. https://doi.org/10.1108/OIR-12-2014-0315

Ezzikouri, H., Oukessou, M., Youness, M. & Erritali, M. (2018). Fuzzy Cross Language Plagiarism Detection (Arabic-English) using WordNet in a Big Data environment. In *Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing* (pp. 22-27). ACM. https://doi.org/10.1145/3264560.3264562

Franco-Salvador, M., Gupta, P., & Rosso, P. (2013). Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing. *In PROMISE Winter School* (pp. 227-236). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-54798-0_12

Schleimer, S., Wilkerson, D. S., & Aiken, A. (2003). Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data* (pp. 76-85). https://doi.org/10.1145/872757.872770

Gupta, P., Barrón-Cedeno, A., & Rosso, P. (2012). Cross-language high similarity search using a conceptual thesaurus. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 67-75). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33247-0_8

Hanane, E., Erritali, M., & Oukessou, M. (2016). Semantic Similarity/Relatedness for Cross language plagiarism detection. *In 2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)* (pp. 372-374). IEEE.

Hattab, E. (2015). Cross-language plagiarism detection method: Arabic vs. English. In *2015 International Conference on Developments of E-Systems Engineering (DeSE),* (pp. 141-144). IEEE.

Li, Y., McLean, D., Bandar, Z.A. & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering, 18*(8), 1138-1150.

Mohammad, A. S., Jaradat, Z., Mahmoud, A. A., & Jararweh, Y. (2017). Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. *Information Processing & Management*, *53*(3), 640-652. https://doi.org/10.1016/j.ipm.2017.01.002

McNamee, P., & Mayfield, J. (2004). Character n-gram tokenization for European language text retrieval. *Information retrieval*, *7*(1), 73-97. https://doi.org/10.1023/B:INRT.0000009441.78971.be

Mihalcea, R., Corley, C. & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai* (Vol. 6, No. 2006, pp. 775-780).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Potthast, M., Barrón-Cedeno, A., Stein, B. & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation, 45*(1), 45-62. https://doi.org/10.1007/s10579-009-9114-z

Potthast, M., Stein, B., & Anderka, M. (2008). A wikipedia-based multilingual retrieval model. In *European conference on information retrieval* (pp. 522-530). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78646-7_51

Sultan, M.A., Bethard, S. & Sumner, T. (2015). Dls@ cu: Sentence similarity from word alignment and

semantic vector composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 148-153). https://doi.org/10.18653/v1/S15-2027

Wu, Z. & Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* (pp. 133-138). https://doi.org/10.3115/981732.981751

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y.H. & Strope, B. (2019). Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307.*

Zhou, Y. & Bollegala, D. (2019). Unsupervised evaluation of human translation quality. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (pp. 55-64). SCITEPRESS-Science and Technology Publications.