Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases

Salha Alzahrani [a], Hanan Aljuaid [b]

[a] College of Computers and Information Technology, Taif University, Saudi Arabia
[b] College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Saudi Arabia

ARTICLE INFO

Article history:
Received 9 February 2020
Revised 21 March 2020
Accepted 9 April 2020
Available online xxxx

Keywords:
Semantic similarity
Plagiarism detection
Natural language processing
Neural networks

ABSTRACT

The rapid growth in the digital era initiates the need to inculcate and preserve the academic originality of translated texts. Cross-lingual semantic similarity is concerned with identifying the degree of similarity of textual pairs written in two different languages and determining whether they are plagiarized. Unlike existing approaches, which exploit lexical and syntax features for mono-lingual similarity, this work proposed rich semantic features extracted from cross-language textual pairs, including topic similarity, semantic role labeling, spatial role labeling, named entities recognition, bag-of-stop words, bag-of-meanings for all terms, n-most frequent terms, n-least frequent terms, and different sets of their combinations. Knowledge-based semantic networks such as BabelNet and WordNet were used for computing semantic relatedness across different languages. This paper attempts to investigate two tasks, namely, cross-lingual semantic text similarity (CL-STS) and plagiarism detection and judgement (PD) using deep neural networks, which, to the best of our knowledge, have not been implemented before for STS and PD in cross-lingual setting, and using such combination of features. For this purpose, we proposed different neural network architectures to solve the PD task as either binary classification (plagiarism/independently written), or even deeper classification (literally translated/paraphrased/summarized/independently written). Deep neural networks were also used as regressors to predict semantic connotations for CL-STS tasks. Experimental results were performed on a large number of handmade data taken from multiple sources consisting of 71,910 Arabic-English pairs. Overall, experimental results showed that using deep neural networks with rich semantic features achieves encouraging results in comparison to the baselines. The proposed classifiers and regressors tend to show comparable performances when using different architectures of neural networks, but both the binary and multi-class classifiers outperform the regressors. Finally, the evaluation and analysis of using different sets of features reflected the supremacy of deeper semantic features on the classification results.

1. Introduction

Plagiarism is an offensive act of using others' work, which is broadly defined by the Encyclopedia of Applied Linguistics as "an illegitimate and deceptive intertextual relationship" (Pecorari, 2012). Plagiarism is often viewed as academic dishonesty committed intentionally. However, several plagiarism cases in the academic field are inadvertent, especially when writing in a language different from the writer's mother tongue, because of factors such as language barriers and underdeveloped writing skills (Haitch, 2016). As educational research points to the need to "detect and deter" plagiarism, rather than "detect and punish", deep understanding of such human factors with the support of computerized plagiarism checkers would lead to better plagiarism prevention; however, cases of intentional academic dishonesty are frequent, especially when texts are obfuscated to hide plagiarism. Understanding obfuscated plagiarism patterns would lead to better solutions to detect them (Kučečka; Pierce and Zilles, 2017). One obfuscation method consists of translating texts from one language to another without proper credit to the original source. Translated

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

E-mail addresses: s.zahrani@tu.edu.sa (S. Alzahrani), haaljuaid@pnu.edu.sa (H. Aljuaid)

https://doi.org/10.1016/j.jksuci.2020.04.009
1319-1578/

plagiarism has other labels, such as cross-language plagiarism and multi-language plagiarism.

Detecting semantic relations across languages has become possible with the recent advances in lexical knowledgebase resources (Vulic and Moens, 2014; Tomassetti et al., 2014; Batet et al., 2014; Vulic and Moens, 2013; Stoyanova et al.; Sorg and Cimiano, 2010; Shumin et al.; Volk et al., 2002). Models for mono-language plagiarism detection cannot be directly applied to reveal cross-language plagiarism; hence, the need for statistical machine translations is apparent (Barrón-Cedeño, 2010; Barrón-Cedeño et al., 2010). Different research works for cross-language plagiarism been have developed: for example, by using language normalization whereby suspicious and source documents were translated into English as a common language and then used as mono-language plagiarism detection (Corezola Pereira et al., 2010); by exploiting comparable corpora and latent semantic analysis (Potthast et al., 2011); by learning equivalence rules from bilingual names lists, and finding related textual clusters across different languages (Steinberger, 2012); by using bilingual alignment-based semantic similarity (Barrón-Cedeño et al., 2013); by employing machine translation from knowledgebase resources (Danilova, 2013); by extracting key phrases and applying semantic similarity measures (Kothwal and Varma, 2013); by constructing knowledge graphs to recognize contextual semantic similarities in document fragments (Franco-Salvador et al., 2014); and by translating key phrases and implementing linear logistic regression models (Alaa et al., 2016).

However, little research (if any) has investigated the use of numerous semantic features between texts from two different languages and trained them on neural networks regressors and classifiers, in order to judge them as either plagiarized or independently written beyond their contextual similarity. Therefore, this work aims to (i) construct a standard benchmark corpus for evaluation of Arabic-English cross-language plagiarism detection, consisting of thousands of human-translated cases; (ii) extract semantic and syntax features and their combination from multilingual text pairs and study their correlations to plagiarism; and (iii) implement deep learning neural networks for cross-lingual plagiarism detection and evaluate their performances on a dataset created for this purpose.

2. Literature review

2.1. Overview on plagiarism detection

Plagiarism detection has two phases: retrieval of source documents, also known as candidate retrieval, and detailed analysis between these sources and the document under investigation. Giving attention to the recent advances in plagiarism detection in the past five years, different studies have focused on the retrieval of sources and suggested solutions for it, rather than considering the step a simple search. A study (Hussain and Suryani, 2015) showed that using external knowledgebase sources improved semantic similarity and contextual significance when retrieving sources of plagiarism. They used nearest neighbor search and support vector machine to retrieve candidates for various plagiarism types other than cases where the material had been literally copied. A study proposed candidate retrieval for Arabic text-reuse from web documents that used encoded fingerprints to formulate queries and gave the best selection of source documents (Lulu et al., 2016). Apart from previous studies on candidate retrieval from the same language, a study for cross-lingual candidate retrieval using two-level proximity information was proposed (Ehsan and Shakery, 2016). Their study used a keyword-focused approach where the suspicious (i.e., query) document was divided into fragments using a topic-based segmentation algorithm, followed by a proximity-based model to retrieve sources relevant to the query segments. Regarding the second phase of plagiarism detection, review studies showed that the current trends in plagiarism detection research need further improvements and require further extensions to less-sourced languages (Eisa et al., 2015; Alzahrani et al., 2012). A survey for cross-language plagiarism detection (Barrón-Cedeño et al., 2013) highlighted an important direction to enhance cross-lingual performances of existing techniques and investigate more languages and machine learning methods. Therefore, the purpose of the next parts is to review related works from 2015 onward.

Different research works focused on plagiarism detection for obfuscated plagiarism cases (Alzahrani et al., 2015; Schmidt et al., 2016; Vani and Gupta, 2017; Paul and Jamal, 2015). In this research field, Semantic Role Labeling (SRL) was used to improve sentence ranking in plagiarism detection (Paul and Jamal, 2015), and a fuzzy semantic-based similarity model was suggested using WordNet-combined semantic similarity measures to detect highly obfuscated plagiarism cases on handmade paraphrases and artificial plagiarism cases (Alzahrani et al., 2015). Further, visual inspection of highly obfuscated plagiarisms was obtained by adding an intermediate step between candidate retrieval and detailed analysis; this new step implemented an extended Jaccard measure to handle synonyms/hypernyms in text fragments (Schmidt et al., 2016). Apart from content-based obfuscated plagiarism, a study investigated citation-based plagiarism detection in scholar documents and compared them with content-based methods (Pertile et al., 2016). The study showed that citations and references can be used to complement other methods and improve the quality of plagiarism detection results. Recent studies focused on dealing with document plagiarism detection as a binary classification task (Vani and Gupta, 2017). In this regard, suspicious-source document pairs were classified as either containing plagiarism or not using Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) classifiers. Mono-lingual features from text pairs were extracted focusing on minimal but effective syntax-based features: part-of-speech (POS) and chunk features. The proposed classifiers were evaluated on handmade and artificial plagiarism cases in English texts and showed superior results in comparison to classical baselines. Another extension used syntax and semantic features with genetic algorithms (GA) to detect obfuscated plagiarism in the form of summarized texts. GA algorithm was used for sentence-level concept extraction. Two detection phases, namely document-level and passage-level, were incorporated using syntactic and semantic features for synsets extracted from WordNet lexical database (Vani and Gupta, 2017). Further linguistic features including POS tagging, chunking, chunking with POS tagging, Semantic Role Labeling (SRL) and SRL with POS tagging were incorporated to improve the detection of different plagiarism types using a combined syntactic-semantic similarity metric (Vani and Gupta, 2018). The evaluation of various cases of plagiarism also reflected the supremacy of deeper linguistic features for identifying obfuscated plagiarism in a monolingual setting.

2.2. Arabic plagiarism detection

Latent semantic analysis (LSA) and singular value decomposition (SVD) were used to investigate the relation between documents and their n-gram phrases in Arabic documents (Hussein, 2015). N-gram fingerprinting with k-overlapping was used to detect exact or nearly copied texts in Arabic documents (Alzahrani, 2015). Further, paragraph-based retrieval of similar resources was conducted using Arabic Wikipedia, and three approaches were utilized for alignments and detailed analysis: skip-gram based, sentence-based, and common words-based approaches (Magooda et al., 2015).

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

Unlike previous works using lexical and contextual features of texts, different researchers have investigated graph-based, machine learning and deep learning methods for Arabic plagiarism detection (Suleiman et al., 2017; Boukhalfa et al.; Nagoudi et al., 2018). Neural networks with one hidden layer on a large Open Source Arabic Corpora (OSAC) dataset were used to construct feature vectors of Arabic words using two models: continuous bag-of-words and continuous skip-gram models (Suleiman et al., 2017); these feature vectors describe each word's context or semantics with high precision. Cosine similarity was used to compute words' vector similarity, and the results of high similarity between the words of text pairs indicate the existence of plagiarism. Another work used a graph-based approach to process Arabic texts and generate stems of their words capable of revealing plagiarism between text pairs (Boukhalfa et al.). Arabic words were represented by a directed weighted graph mapped to a stem from the database; words' stem graphs were then compared to detect plagiarism. Methods for detecting obfuscated plagiarism in Arabic texts were explored based on word embedding, word alignment, and word weighing for the purpose of measuring semantic similarities among textual units (Nagoudi et al., 2018). Lexical, syntactic, and semantic features of sentence alignments were used to train SVM, DT, and Random Forests (RF) evaluated on an AraPlagDet standard corpus.

2.3. Cross-language plagiarism detection methods

In recent years, special attention has been given to plagiarism detection across languages; in this section, we will survey different studies from 2015 to the present. A study used knowledge graph analysis and a new weighing metric between concepts in their graph representations (Franco-Salvador et al., 2016). Their work focused on using different features from multi-lingual graphs, such as word sense disambiguation, vocabulary expansion, and representation by similarities with a collection of concepts. An extension was achieved using continuous space representation (i.e., word embedding) and alignment similarity analysis to complement knowledge graph representation and plagiarism detection (Franco-Salvador et al., 2016). Another study used semantic relatedness metrics between words and concepts from knowledge graphs in WordNet and combined them to measure the similarity of sentences and paragraphs for multilingual plagiarism cases (Hanane et al., 2016). Further, a linear translation was performed to project continuous space word embedding vectors from different languages to English in order to obtain a resource-light semantic similarity of texts (Glavaš et al., 2018). Different research works have investigated cross-language plagiarism of Arabic-English documents (Hattab, 2015; Alaa et al., 2016; Ezzikouri et al., 2018), and latent semantic indexing was used to construct cross-lingual vectors for measuring the contextual similarity of English-Arabic texts (Hattab, 2015). Linear logistic regression was used with English-Arabic texts that were represented by their key phrases projected into a common language (Alaa et al., 2016). A recent study evaluated English-Arabic fuzzy semantic similarity using Apache Hadoop with its distributed file system HDFS for big data (Ezzikouri et al., 2018). Table 1 presented a summary of plagiarism detection methods used for mono-language and cross-language from 2015 until the present; it highlights the methods employed alongside the document/textual feature representation used in each.

2.4. Semantic similarity computations methods

Many studies are concerned with detecting semantic similarity (SS) between words, sentences, and documents through linguistic relations. SS can be computed based on semantic networks (SN), which can be defined as a structured tree representation of lexicons and their relations. Generating SN, also called SN induction, is a well-known problem in natural language processing; WordNet (Miller, 1995); HowNet (Liuling et al., 2008) and BabelNet (Navigli and Ponzetto, 2012) are examples of successful efforts of SN induction. Different studies utilized established knowledge bases such as Wikipedia, Facebook, Google Knowledge Graph, LinkedIn, among others (Qu et al., 2018; Wanjawa and Muchemi, 2018). Sematch is a tool that constructs SN using statistical information from knowledge bases and concepts of trees such as depth, path length, and Least Common Subsumer (Zhu and Iglesias, 2017).

Existing literatures on SS is divided into two parts. The first part comprises semantic word similarity (SWS), which is concerned with the degree of similarity between words using SN. Many SWS metrics have been proposed on the basis of WordNet (Stoyanova et al.; Liu et al., 2012; Qin et al., 2009; Budanitsky and Hirst, 2006; Leacock et al., 1998) and HowNet (Zhang et al., 2014; You et al., 2012; Liuling et al., 2008). Some of the recent advances in this direction include using information content and hybrid features from Wikipedia (Qu et al., 2018), and using set theory and WordNet to calculate the relatedness between the glosses and synsets of two words (Ezzikouri et al., 2019). On the other hand, several methods to investigate not only words but also sentences and documents using knowledge-based metrics and corpora-based statistics have been researched (Pawar and Mago, 2019; Alian and Awajan, 2018; Xu et al., 2016). The second part is semantic text similarity (STS), where the text is defined as a sentence, paragraph, or document. Recent advances in STS include combining SWS metrics, word embedding models, and corpus-based methods to estimate semantic associations and relatedness between texts (Pu et al., 2017; Shajalal and Aono, 2018; Quan et al., 2019), classify tweets for paraphrase identification using support vector regression and maximum entropy trained on different lexical, syntax, and semantic features (Al-Smadi et al., 2017), and investigate parallel algorithms for querying Arabic documents based on translation and SWS approaches on a MapReduce platform (Hadi et al., 2019).

Different studies investigated SWS and STS across languages, which is known as cross-language semantic similarity (CL-SS). Recently, BabelNet achieved state-of-the-art lexical coverage as a multilingual network with a wide coverage of more than 280 languages obtained from the integration of WordNet, Wikipedia, OmegaWiki, Wiktionary, Wikidata, Wikiquote, VerbNet, Microsoft Terminology, GeoNames, ImageNet, FrameNet, WN-Map, and Open Multilingual WordNet (Navigli and Ponzetto, 2012). Concepts and named entities in BabelNet are known as Babel synsets; each represents a given meaning and the synonyms that express that meaning in a range of different languages. One study proposed a joint multilingual method called BabelRelate, aimed at constructing multilingual semantic graphs for two words using BabelNet, and to compute semantic relatedness from graphs intersection (Navigli and Ponzetto). Many research works have been built around BabelNet, such as web query classification using improved visiting probability algorithm (Rashidghalam and Mahmoudi, 2015); text summarization using concept graph and BabelNet knowledge base (Rashidghalam et al., 2016); classification algorithms for verbose queries detection (Gharouit and Nfaoui, 2017); synsets and semantic relations extraction from BabelNet (Ustalov and Panchenko, 2017); mapping senses in BabelNet to Chinese based on word embedding (Meng et al., 2017); Chinese word semantic relation classification (Meng et al., 2017); Babelfy-based extraction of collocations from Turkish hotel reviews (Ekinci and Omurca, 2018); automatic SN induction from unstructured documents (Wanjawa and Muchemi, 2018); and ontology-based topic detection (Gutiérrez-Batista et al., 2018).

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

Table 1
Summary of mono- and cross-lingual plagiarism detection methods from 2015-present.

Type	Stage	L.	Method(s)	Representation/Features	Ref.
Mono-lingual plagiarism detection	Candidates retrieval	En	nearest neighbor search and support vector machine	Documents	(Hussain and Suryani, 2015)
		Ar	encoded fingerprints to formulate queries for best selection of candidates	Documents	(Lulu et al., 2016)
			latent semantic analysis (LSA)	N-gram phrases	(Hussein, 2015)
			paragraph-based retrieval using Arabic Wikipedia		(Magooda et al., 2015)
	Detailed analysis	En	semantic role labeling	Sentences	(Paul and Jamal, 2015)
			fuzzy semantic-based similarity model using WordNet-semantic similarity measures		(Alzahrani et al., 2015)
			extended Jaccard measure	Synonyms/hypernyms in text fragments	(Schmidt et al., 2016)
			Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) classifiers	Syntax-based features; part-of-speech (POS) and chunk features	(Vani and Gupta, 2017)
			Genetic algorithms (GA)	Syntactic and semantic features for synsets extracted from WordNet	(Vani and Gupta, 2017)
			Combined syntactic-semantic similarity metric	POS tagging, chunking, chunking with POS tagging, semantic role labelling (SRL), SRL with POS tagging	(Vani and Gupta, 2018)
			citation-based plagiarism detection	Structural citations	(Pertile et al., 2016)
		Ar	singular value decomposition (SVD)	N-gram phrases, N-gram words, and sentences	(Hussein, 2015)
			skip-gram based, sentence-based, and common words-based approaches		(Magooda et al., 2015)
			n-gram fingerprinting with k-overlapping		(Alzahrani, 2015)
			Deep learning using neural networks for Word2vec model	Continuous bag-of-words and continuous skip-gram feature vectors	(Suleiman et al., 2017)
			Graph-based approach	Word stems	(Boukhalfa et a.)
			Support Vector Machine (SVM), Decision Trees (DT), and Random Forests (RF).	Lexical, syntactic, and semantic features of sentences	(Nagoudi et al., 2018)
Cross-lingual plagiarism detection	Candidates retrieval	En vs. Gr, Sp	Two-level proximity information: topic-based segmentation algorithm + proximity-based model	Documents	(Ehsan and Shakery, 2016)
			Latent Semantic Indexing (LSI) and analysis	Documents	(Hattab, 2015)
	Detailed analysis	En vs. Gr, Sp	knowledge graphs	word sense disambiguation, vocabulary expansion, collection of concepts, continuous word alignments	(Franco-Salvador et al., 2016) (Franco-Salvador et al., 2016)
		En vs. Fr, Ar	WordNet-based semantic relatedness metrics	Words, sentences, paragraphs	(Hanane et al., 2016)
		En vs. Sp, It, Cr	Linear translation and projection of vectors from other languages to English	Continuous word embedding and alignments	(Glavaš et al., 2018)
		En-Ar	Linear Logistic Regression (LLR)	Key phrases	(Alaa et al., 2016)
			Fuzzy semantic similarity using WordNet semantic Wu&Palmer and Lin metrics	Big data (Apache Hadoop with its distributed file system HDFS)	(Ezzikouri et al., 2018)

3. Method

3.1. Problem and general framework

Different studies focus on the first stage, candidate retrieval (CR) (Ehsan and Shakery, 2016; Zhou et al., 2012; Ye et al., 2012; Sorg and Cimiano, 2010; He and Wang, 2009; AbdulJaleel and Larkey; Volk et al., 2002; AlJlayl and Frieder). This study focuses on the next stage for cross-language semantic text similarity (CL-STS) and plagiarism detection (PD). Two research questions were investigated as follows:

- Given a pair of texts t and t' from two languages L and L', respectively, is t' plagiarized from t (i.e., t' is translated from t)?
- Given a pair of texts t and t' from two languages L and L', respectively, find the degree of CL-STS between t' and t.

To tackle these questions, and according to the previous methods (Al-Smadi et al., 2017; Xu et al., 2015), our methodology is divided into the following two tasks (it ought to be noted that our work differs from previous works in utilizing cross-language semantic features and deep learning approaches):

- **Input:** text pair (t, t'), where t and t' are from two different languages L and L'. This study investigated the case study of using English and Arabic texts as L and L', respectively.
- **CL-STS Task (Regression):** Given a pair of texts t and t' from two languages L and L', respectively, this task aims to compute the degree of semantic similarity between t and t' in a scale of four [1: dissimilar, 2: similar, 3: highly similar, 4: identical meaning] according to the rating scale of standard semantic similarity datasets (Rubenstein and Goodenough, 1965; Li et al., 2006). We mapped this scale to the range (Pecorari, 2012), as follows:

[0, 0.25] means that t and t' are completely on different topics (t and t' are independently written – *IW*),

[0.25, 0.50] means that t and t' share some details (t' is translated and summarized from t – *ST*),

[0.50, 0.75] means that t and t' share many details (t' is translated and paraphrased from t – *PT*),

[0.75, 1] means that t and t' have identical meaning (t' is literally translated from t – *LT*).

- **PD Task (Classification):** Given a pair of texts t and t' from two languages L and L', respectively, this task aims to classify t and t'

as follows: (i) binary classification: where each pair is classified as plagiarism or not. (ii) 4-type classification: where each pair is classified as IW, ST, PT, and LT.

The general framework of the proposed work is shown in Fig. 1, and each step will be explained in the following subsections.

3.2. Pre-processing

For text cleansing and preparation, such as punctuation removal, sentence segmentation, and word tokenization, we used NLTK, which supports both Arabic and English (Edward and Steven, 2002); diacritics were also removed from Arabic texts. NLTK supports POS tagging and lemmatization of English texts. For Arabic texts, additional modules were used, including Stanford Arabic parser for words lemmatization (Green and Manning) and Stanford POS tagger (Toutanova et al., 2003).

3.3. Feature extraction

Different syntax features based on POS tags were used for mono-language plagiarism detection (Nagoudi et al., 2018; Vani and Gupta, 2017, 2018). Since this study proposed a cross-language similarity evaluation, lexical features such as bag-of-words, character n-grams, word n-grams, and syntax-based features cannot be employed to our advantage. Therefore, we proposed various and rich semantic-based features depending on topic similarity, Named Entity Recognition (NER), Semantic Role Labeling (SRL), Spatial Role Labeling (SpRL), bag-of-meanings, and bag-of-stop words, for text comparisons from two different languages. A total of 18 features were extracted as follows.

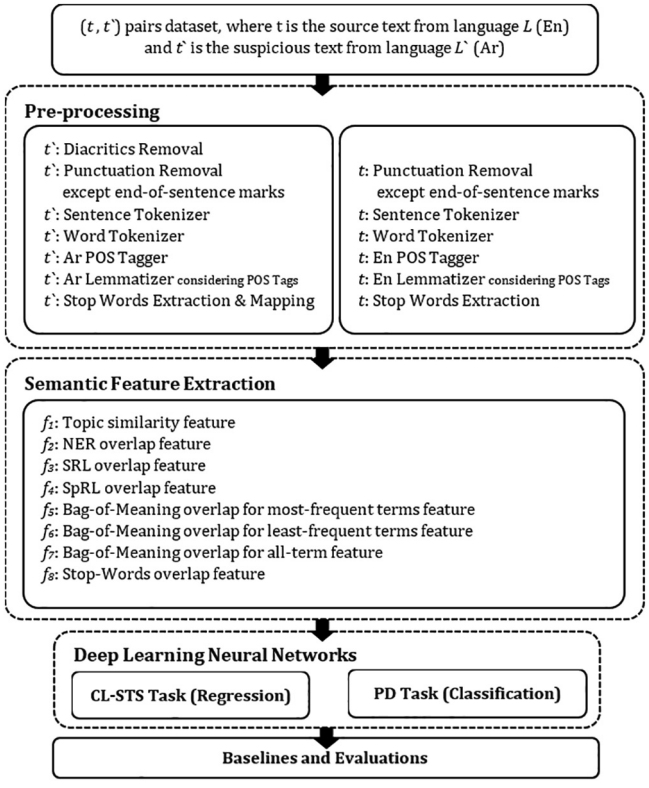

Fig. 1. Framework of the proposed method.

3.3.1. Topic similarity and overlap features (Topic_Sim, Topic_O)

Topic extraction methods are applied to Arabic and English texts. In this regard, we used Latent Dirichlet Allocation (LDA), an unsupervised machine learning model that takes documents as inputs and generates topics as outputs (Blei et al., 2003). Three parameters were employed: the number of topics, the number of words per topic, and the number of topics per document. We proposed two methods to extract these features; in the first method, LDA generated topics from English sources and Arabic texts separately, wherein the number of topics was n = 5 and the length of words vector per topic was wv = 10. Then, each text was matched with the most similar topics from the generated n topics. Further, equivalent English synonyms for the Arabic topics were extracted using BabelNet (Navigli and Ponzetto, 2012), and the averaged semantic similarity between English topics and translated Arabic topics was computed as follows:

$$Topic\_Sim = \frac{\sum_{w \in Topics(t)} \sum_{\acute{w} \in Topics(\acute{t})} wup(w, \acute{w})}{totalnumberoftopics}$$

where wup is WordNet-based word similarity metric (Wu and Palmer, 1994), and LCS is the long common subsequence as follows:

$$wup(w, \acute{w}) = \frac{2 \times depth(LCS(w, \acute{w}))}{depth(w) + depth(\acute{w})}$$

In the second method, LDA generated the common list of topics from all text pairs after converting Arabic texts into English as a common language using Google Translate API. The number of topics was n = 5 and the length of words vector per topic was wv = 5. Lastly, each text was related to the most similar topics from the generated n topics, and the topic overlapping feature of two texts t and t‘ was computed as the Jaccard similarity of their extracted topics vectors as follows:

$$Topic\_O = \frac{|Topics(t) \cap Topics(\acute{t})|}{|Topics(t) \cup Topics(\acute{t})|}$$

3.3.2. Named Entity Recognition overlap feature (NER_O)

Named Entity Recognition (NER) refers to the process of labeling known objects in the text, particularly into three labels. The label *person* is given to the names of people, companies, genes or protein, *organization* is given to universities and companies, and *location* is associated with cities, countries and the like. Stanford NER (Finkel et al., 2005) is a well-known recognizer for English, which we utilized for our research. Various efforts have so far been made toward Arabic NER; however, none has proved stability with contemporary Arabic texts. Thus, to extract this feature, we converted Arabic text into English using Google Translate API, as the direct translation proves efficient in converting the three classes to their equivalent naming; further, we did not regard the coherence of translated texts as a target. The overlapping between named entities was employed as one of the semantic association between two texts (Nagoudi et al., 2018). Using NER, four features were adopted for cross-language plagiarism detection as the Jaccard similarity between the bag-of-named-entities (*BoNE*) of two texts t and t‘—one feature for each class, and the fourth feature as the mean of them according to the equations below.

$$NER\_O_{per} = \frac{|BoNE_{per}(t) \cap BoNE_{per}(\acute{t})|}{|BoNE_{per}(t) \cup BoNE_{per}(\acute{t})|}$$

$$NER\_O_{org} = \frac{|BoNE_{org}(t) \cap BoNE_{org}(\acute{t})|}{|BoNE_{org}(t) \cup BoNE_{org}(\acute{t})|}$$

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

$$NER\_O_{loc} = \frac{|BoNE_{loc}(t) \cap BoNE_{loc}(\acute{t})|}{|BoNE_{loc}(t) \cup BoNE_{loc}(\acute{t})|}$$

$$NER\_O = Mean(NER\_O_{per}, NER\_O_{org}, NER\_O_{loc})$$

3.3.3. Sematic Role Labeling similarity feature (SRL_Sim)

Semantic Role Labeling (SRL), also known as shallow semantic parsing, refers to the process of extracting the predicate (i.e., verb) from a given sentence and its associated semantic arguments with their classification into specific roles (Ye and Baldwin, 2006; Gildea and Jurafskyy, 2000). SRL was used to improve mono-language plagiarism detection (Osman et al., 2012) and was adopted for Arabic (Diab et al.). However, none of the previous works used SRL as a feature for CL-STS and PD. To extract this feature, we used an SRL tool developed by Illinois University (Clarke et al.) and adopted cross-language semantic similarity of the verb predicates from *t* and *t*‘ using BabelNet and *wup* similarity metric (Wu and Palmer, 1994) as follows.

The list of senses (synonyms) for a verb predicate from the Arabic text *t*‘ indicated by *predicate*(*t*‘) was obtained directly in English as the target language using BabelNet queries (Navigli and Ponzetto, 2012). Below is an example of a query to get a sense of a verb predicate, “”يقرأ in Arabic, meaning “read” in English.

https://babelnet.io/v5/getSenses?lemma=□□□&searchLang=AR&pos=VERB&targetLang=EN&source=WORDNET&key=[]

Then, *predicate*(*t*) from the English text *t* was correlated with the senses vector using the following function:

$$SRL_{Sim} = 1 - \prod_{i=1}^{n} \left(1 - wup(predicate(t), predicate(\acute{t}_i)\right)$$

where Π is the product function, *wup* refers to the similarity obtained between the *predicate*(*t*) and *predicate*(*t*‘) (Wu and Palmer, 1994), and *n* is the total number of senses obtained from *predicate*(*t*‘).

3.3.4. Spatial Role Labeling similarity feature (SpRL_Sim)

Similar to SRL, Spatial Role Labeling (SpRL) indicates the process of extracting spatial arguments from a given sentence. The spatial arguments are three labels: the *indicator*, which is given to the prepositions; the *trajector*, which is the precedence object to be located, placed, or moved; and the *landmark*, which is the topological object in the spatial scenario wherein the trajector is located/-placed in, and moved to/from. For English SpRL, we used the methods provided by previous research (Kolomiyets et al., 2013; Bastianelli et al., 2013; Roberts and Harabagiu; Kordjamshidi et al.; Kordjamshidi et al., 2011; Kordjamshidi et al.), while we relied on our previous work for Arabic SpRL (Alzahrani, 2016). Four additional features were adopted for cross-language plagiarism detection from text pairs *t* and *t*‘ as follows.

Three vectors of senses were obtained using BabelNet query (Navigli and Ponzetto, 2012) shown in the previous section from the spatial *indicator*(*t*‘), *trajectory*(*t*‘), and *landmark*(*t*‘), and then the features were computed as follows:

$$SpRL\_Sim_{indicator} = 1 - \prod_{i=1}^{n} \left(1 - wup(indicator(t), indicator(\acute{t}_i)\right)$$

$$SpRL\_Sim_{\text{trajectory}} = 1 - \prod_{i=1}^{m} \left(1 - wup(\text{trajectory}(t), \text{trajectory}(\acute{t}_i)\right)$$

$$SpRL\_Sim_{\text{landmark}} = 1 - \prod_{i=1}^{l} \left(1 - wup(\text{landmark}(t), \text{landmark}(\acute{t}_i)\right)$$

$$SpRL\_Mean = Mean(SpRL\_Sim_{indicator}, SpRL\_Sim_{\text{trajectory}}, SpRL\_Sim_{\text{landmark}})$$

where Π is the product function, *wup* refers to the word semantic similarity, and *n*, *m*, and *l* are the total numbers of senses obtained from extracted labels: *indicator*, *trajectory*, and *landmark*, respectively.

3.3.5. Bag of meanings overlap (BoM_O)

This feature indicates the Jaccard similarity between the bag-of-meanings (*BoM*) for lemmas in two texts *t* and *t*‘, as shown in the equation below.

$$BoM\_O = \frac{|BoM(t) \cap BoM(\acute{t})|}{|BoM(t) \cup BoM(\acute{t})|}$$

Each word was represented by its lemma form and POS tag. The set of senses of each lemma was generated with regard to its POS in the text. To get a sense of lemmas in Arabic and English, we used BabelNet (Navigli and Ponzetto, 2012) to query a word in Arabic, and extracted its senses in English. Another two features were added by limiting this feature to the three most-frequent terms and three least frequent terms, as follows:

$$BoM\_O\_Most = \frac{|BoM(most\_freq\_words(t)) \cap BoM(most\_freq\_words(\acute{t}))|}{|BoM(most\_freq\_words(t)) \cup BoM(most\_freq\_words(\acute{t}))|}$$

$$BoM\_O\_Least = \frac{|BoM(least\_freq\_words(t)) \cap BoM(least\_freq\_words(\acute{t}))|}{|BoM(least\_freq\_words(t)) \cup BoM(least\_freq\_words(\acute{t}))|}$$

3.3.6. Stop words overlap (SW_O)

This feature was extracted from a so-called bag-of-stop-words (*BoSW*) of two texts *t* and *t*‘, where the list of stop words from the Arabic text is generated using a mapping function that transforms extracted stop words in Arabic into one or more of the English stop words. We used NLTK’s lists of Arabic and English stop words; examples of their mapping are shown in Table 2. Four more features were used as shown below.

$$SW\_O = \frac{|BoSW(t) \cap BoSW(\acute{t})|}{|BoSW(t) \cup BoSW(\acute{t})|}$$

$$SW\_O_{En} = \frac{|BoSW(t) \cap BoSW(\acute{t})|}{|BoSW(t)|}$$

$$SW\_O_{Ar} = \frac{|BoSW(t) \cap BoSW(\acute{t})|}{|BoSW(\acute{t})|}$$

$$SW\_O\_Mean = Mean(SW\_O, SW\_O_{En}, SW\_O_{Ar})$$

Table 2
Mapping of Arabic stop words into English stop words.

Ar stop word	Mapped to En stop word	Ar stop word	Mapped to En stop word
إذ	it, it’s	أكثر	more
إذا	if	إلا	but
إذما	then	التي, الذي, الذين, اللاتي, اللائي, اللتان, اللتين , اللذان, اللذين, اللواتي	which, who, whose, that
إذن	so	إلى, إليك, إليكم, إليكما, إليكن	to, into, at
أقل	few	إما	or

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

3.4. Deep neural networks models

Deep neural networks have proved their efficiency in many natural language processing applications (LeCun et al., 2015; Schmidhuber, 2015). We proposed a basic architecture of deep networks with two or more hidden layers and rectified linear unit (ReLU) activation function to better learn the data. Besides its non-linearity, the main advantage of using ReLU function over the other activation functions is that it does not activate all the neurons simultaneously, making the network sparse and very efficient. Additionally, ReLU function was one of the main advancements in the field of deep learning that led to overcoming the vanishing gradient problem. We used Softmax function, which is a type of sigmoid function, in the output layer to handle classification problems and to get the probabilities determining the class of each input. The number of input predictors in the input layer depend on the number of features used from our dataset, whilst the number of output neurons vary based on the task. In the CL-STS task, which is a regression problem, we used one output neuron and a sample of the deep network architecture is demonstrated in Fig. 2. The aim of regression is to estimate to what degree the deep network will evaluate the cross-lingual pairs in comparison to human similarity ground-truth data. In the PD task, which is classification problem, we used two output neurons for the binary classifier (plagiarized or not) and four output neurons for the 4-type classification (LT, PT, ST, IW). Both architectures are shown in Fig. 3.

4. Experimental setup

4.1. Dataset

To conduct our experiments on cross-language, passage-level aligned texts, we constructed our benchmark dataset from different sources and annotated its level of similarity/plagiarism. The first source of the dataset was constructed by a group of students in a computer science course from Princess Nourah bint Abdulrahman University (PNU) in Riyadh, Saudi Arabia. A total number of 114 students within the age group of 18–24 years were instructed by their course tutors to translate different English texts into Arabic; each student was asked to translate three passages from computer science books into Arabic using their own wording. Their responses were collected via a web link and email invitation using a questionnaire created for this purpose. After cleaning the data by excluding missing and weak translations, a total of 220 passage pairs were obtained. About 76% of the contributors have had previous experience in translation from Arabic to English and vice versa, as can be seen in Table 3.

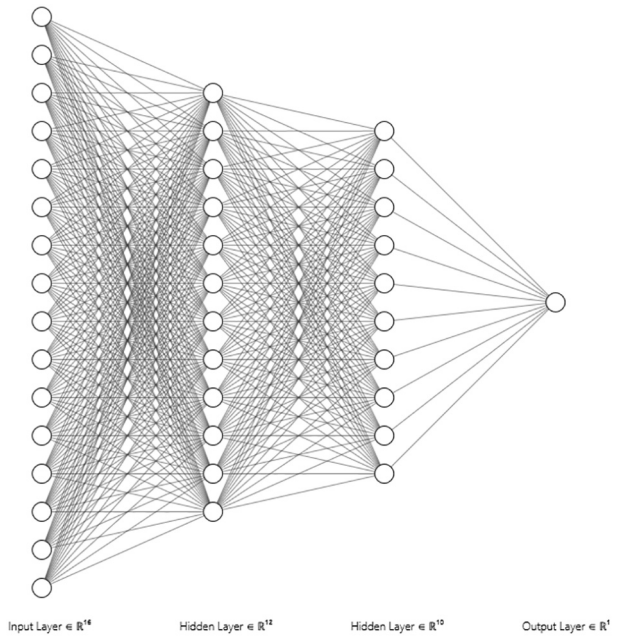

Fig. 2. Deep network architecture for CL-STS regression task.

The second source for the dataset benchmark was taken from different English-Arabic parallel corpora: 547 aligned passages from King Saud University corpus (Alotaibi, 2017), and 58,911 pairs from the United Nations Parallel Corpora (Ziemski et al., 2016) and the OPUS collection of translated texts from the web (Tiedemann, 2012; Tiedemann, 2016), as demonstrated in Table 4.

The third source of the dataset benchmark aims to add independently written texts whereby passage pairs have no semantic similarity. The purpose of adding such pairs is to ensure that our benchmark is free of bias, displaying cases which are both plagiarized or plagiarism-free. A total of 17,550 independently written cases were added to the dataset whereby the sources were taken from United Nations Parallel Corpora (Ziemski et al., 2016) and King Saud University corpus (Alotaibi, 2017) and randomly aligned with Arabic texts from the OPUS (Tiedemann, 2012; Tiedemann, 2016). The resulting benchmark cross-language English-Arabic (CLEA) dataset consists of a total of 71,911 sentence/passage aligned texts tagged manually into *literal translation (LT), paraphrased translation (PT), summarized translation (ST), and independently written (IW)*, using the methodology published for existing benchmarks constructed for other languages (Muneer et al., 2019). In this regard, the texts from the first and second sources described above were tagged as either *literal*, *summarized,* or *paraphrased*. When the text is translated word-by-word and phrase-by-phrase with almost comparable lengths of source and translated texts, it is tagged as *literal translation (LT)*. The translated text was tagged as *paraphrased translation* when restructured with incomparable length to the source but retaining the concept or main idea in the text. When the text is translated, shortened, and summarized from the original text but retains the same idea, it is tagged as *summarized translation*. The texts from the third sources were tagged as *independently written*. The total statistics of the tagged dataset are demonstrated in Table 5.

4.2. Train and test dataset

The data set was divided into train and test datasets using the fraction 0.8; the train dataset contained 57,528 pairs and the test dataset contained 14,382. As seen in Table 6, different types of plagiarism were uniformly distributed between the training and test datasets to ensure the reliability and consistency of our data.

4.3. Baselines and evaluation measures

We adopted two baselines to compare with our methods. The first baseline was the random method (Al-Smadi et al., 2017): for each pair of texts in the test dataset, we assigned a real number between (Pecorari, 2012) as a random semantic similarity degree. We established a cut-off of 0.5 to classify the test dataset as plagiarism or independently written. The second baseline used direct translation to convert Arabic texts into English. Then, we used three mono-lingual analysis and plagiarism detection techniques of word n-gram fingerprinting (or chunking) using word 8-gram chunks (Basile et al., 2009) and word 5-gram chunks (Kasprzak et al., 2009), and statement-based fuzzy semantic similarity as investigated in our work (Alzahrani et al., 2015). It is also noteworthy to compare with simple machine learning algorithms; thus, we

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

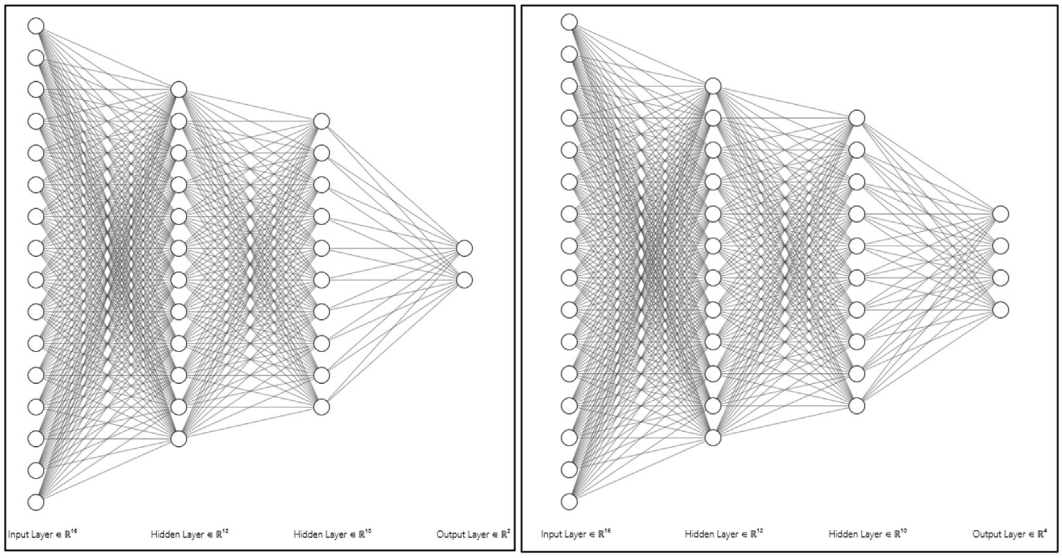

Fig. 3. Deep network architecture for PD classification task.

Table 3
Handmade translated cases by students from Princess Nourah bint Abdulrahman University (PNU) in Riyadh.

Translation experience	Number of Contributors	Domain
No	52	Computer Science
Yes/ Arabic to English	37	Computer Science
Yes/ English to Arabic	121	Computer Science
Yes/ other language	10	Computer Science
Total contributors	220	
Total contributors with translation experience	168 (76% of total contributors)	

Table 4
Details of translated datasets from King Saud University and United Nations Parallel Corpora.

Parallel text source	Number of En-Ar pairs	Domain
King Saud University corpus (Alotaibi, 2017)	547	Arts, Fiction
United Nations Parallel Corpora (Ziemski et al., 2016), OPUS collection of translated texts from the web (Tiedemann, 2012, 2016)	58,911	Political
Total pairs	59,458	
Total pairs (after cleaning short/duplicate phrases)	54,142	

Table 5
Statistics of tagged dataset.

CLEA Cropus	Number of En-Ar pairs	Domain
Literally translated pairs (*LT*)	51,155	Computer Science, Arts, Fiction, Political
Paraphrased translated pairs (*PT*)	2085	
Summarized translated pairs (*ST*)	1120	
Independently written (*IW*)	17,550	–
Total pairs	71,910	

Table 6
Detailed description of training and test dataset.

Type of pair	Training dataset (80%)	Test dataset (20%)
Literally translated pairs (LT)	40,920	10,236
Paraphrased translated pairs (PT)	1679	405
Summarized translated pairs (ST)	889	231
Independently written (IW)	14,040	3510
Total pairs	57,528	14,382

suggested to compare the proposed models with support vector machines (SVM) and linear logistic regression. For evaluating our deep networks, we used optimization score functions called loss functions. The mean squared error (MSE) is the default loss to use for the CL-STS regression problem given by the equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yd_i - y_i)^2$$

where n is the total number of samples for training in one epoch, yd_i is the desired output, and y_i is the actual result; the squared difference is always positive regardless of the sign of yd_i and y_i. This function is optimized to reach its perfect result 0.0.

Cross-entropy is the default loss function to use for PD takes, either as binary classification or k-class classification problems. In both cases, cross-entropy is the measure of the average differences between the desired and estimated probability distributions for predicting the classes in the problem. This function is optimized to reach its perfect result 0.0.

5. Experimental results and discussion

5.1. Sample of pre-processing and feature extraction results

As this paper handles the case of Arabic-English cross-lingual texts, in this section we will demonstrate the results of some samples taken from our dataset to clarify our work. In Table 7, we exemplify the results of pre-processing of one pair. The English text and its translated text in Arabic are taken from Alotaibi (2017). The first row shows the raw texts as taken from the original source. Then, in the following rows, we show the results of a sequence of pre-processing steps, including punctuations removal, diacritics

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

removal for Arabic texts, sentence tokenization, bag-of-words extraction, POS tagging and lemmatization, which were all needed for further feature extraction processes, as will be illustrated shortly.

A part of the feature extraction processes is shown in the tables below. In Table 8, we demonstrated the results from a sample of NER extracted from the first four pairs in the dataset. Their individual feature values and combination show that these features can be good indicators for expressing the semantic connotations of textual pairs, regardless of the fact that the languages are different. Table 9 shows the list of topics extracted by the LDA algorithm from Arabic texts and English texts, respectively. Table 10 shows the results from the two methods we proposed for topic similarity and overlap features, whereby these features can be useful indicators of the semantic connotations of these pairs, regardless of their original languages.

5.2. Classification results

We studied the impact of semantic features and their combinations on CL-STS for Arabic-English text pairs. Different sets of features were formed as follows: bag-of-meaning overlapping (3 features), bag-of-meaning overlapping combined with topic similarity and overlapping (5 features), bag-of-meaning overlapping with stop words overlapping (7 features), NER overlapping combined with SRL and SpRL similarities (9 features), bag-of-meaning overlapping with NER, SRL, SpRL (11 features), and all of the above (18 features). The performance of the classifier was evaluated using binary cross-entropy loss function for binary PD task (plagiarism vs. independently written) and categorical cross-entropy for 4-type PD task (LT, PT, ST, IW). Table 11 shows the classifier results using three deep networks architectures and different sets of features. These classifiers were trained on the train dataset using 100 epochs; generally, the binary classifier outperforms the 4-type classifier as it achieves the highest accuracy (0.9701) using the deep network architecture of two hidden layers, each with 50 neurons. Moreover, different results ought to be highlighted in both classifiers. Using the *BoM_O* features yielded an accuracy of about 0.7; however, combining such features with other semantic connotation features such as *Topic_Sim*, *Topic_O,* and *SW_O* significantly improved the results, increasing the accuracy obtained to 0.9. Similarly, using *SRL_Sim, SpRL_O,* and *NE_O* features obtained an accuracy of 0.7, although these results significantly increased to 0.9 when combined with *BoM_O*. Our classifiers tend to have comparable performances using different architectures of neural networks; however, it ought to be noted that using all feature sets could be pivotal, as the highest accuracy in each architecture was obtained when using all the features extracted from the dataset.

Our classifiers surpassed the random baseline, although the latter used all sets of features. In comparison to the traditional baselines, word 5-gram fingerprinting and word 8-gram fingerprinting obtained less results than expected due to the fact that direct translation may not employ the same wording as the English version. The fuzzy sentence semantic similarity approach, although it achieved better results than word n-grams chunks, fell short of the learned patterns obtained by deep neural networks. Classical SVM and logistic regression classifiers perform comparably, and their accuracies almost reached that of the deep networks. Nevertheless, the two hidden layers, each with 50 neurons, 100 epochs achieved the highest accuracy, 0.970, in comparison to SVM and logistic regression classifiers.

Table 7
Results from pre-processing of the English text and its translated text in Arabic taken from Alotaibi (Alotaibi, 2017).

	En	Ar
Raw	In March, NASA used a small electric military drone, the Dragon Eye, to sample and photograph the noxious gas plume spewing from Turrialba Volcano near San Jose in Costa Rica.	«ناسا» ففي مارس الماضي، استخدمت طائرة عسكرية كهربائية صغيرة دون ـ لأخذ عيّنات، «عين التنّين» طيّار ـ تدعى وتصوير عمود الغازات الضارّة التي ينفثها بركان تُرّيالبا قرب سان خوزيه، كوستاريكا.
Punctuations Removal, Diacritics Removal (Ar Only)	In March NASA used a small electric military drone the Dragon Eye to sample and photograph the noxious gas plume spewing from Turrialba Volcano near San Jose in Costa Rica.	ففي مارس الماضي استخدمت ناسا طائرة عسكرية كهربائية صغيرة دون طيار تدعى عين التنين لأخذ عينات وتصوير عمود الغازات الضارة التي ينفثها بركان تريالبا قرب سان خوزيه كوستاريكا.
Sentence Tokenization	[" In March NASA used a small electric military drone the Dragon Eye to sample and photograph the noxious gas plume spewing from Turrialba Volcano near San Jose in Costa Rica."]	[ففي مارس الماضي استخدمت ناسا" طائرة عسكرية كهربائية صغيرة دون طيار تدعى عين التنين لأخذ عينات وتصوير عمود الغازات الضارة التي ينفثها بركان تريالبا قرب سان خوزيه كوستاريكا."]
Bag-of-Words	["In", "March", "NASA", "used", "a", "small", "electric", "military", "drone", "the", "Dragon", "Eye", "to", "sample", "and", "photograph", "the", "noxious", "gas", "plume", "spewing", "from", "Turrialba", "Volcano", "near", "San", "Jose", "in", "Costa", "Rica"]	["ففي", "مارس", "الماضي", "استخدمت", "ناسا", "طائرة", "عسكرية", "كهربائية", "صغيرة", "دون", "طيار", "تدعى", "عين", "التنين", "لأخذ", "عينات", "وتصوير", "عمود", "الغازات", "الضارة", "التي", "ينفثها", "بركان", "تريالبا", "قرب", "سان", "خوزيه", "كوستاريكا"]
POS Tagging	["In", "IN"], ["March", "NNP"], ["NASA", "NNP"], ["used", "VBD"], ["a", "DT"], ["small", "JJ"], ["electric", "JJ"], ["military", "JJ"], ["drone", "NN"], ["the", "DT"], ["Dragon", "NNP"], ["Eye", "NNP"], ["to", "TO"], ["sample", "VB"], ["and", "CC"], ["photograph", "VB"], ["the", "DT"], ["noxious", "JJ"], ["gas", "NN"], ["plume", "NN"], ["spewing", "VBG"], ["from", "IN"], ["Turrialba", "NNP"], ["Volcano", "NNP"], ["near", "IN"], ["San", "NNP"], ["Jose", "NNP"], ["in", "IN"], ["Costa", "NNP"], ["Rica", "NNP"]]	[[["", "ففي/IN"], ["", "مارس/NN"], ["", "الماضي/DTJJ"], ["", "استخدمت/VBN"], ["", "ناسا/NN"], ["", "طائرة/NN"], ["", "عسكرية/JJ"], ["", "كهربائية/JJ"], ["", "صغيرة/JJ"], ["", "دون/NN"], ["", "طيار/NN"], ["", "تدعى/VBN"], ["", "عين/NN"], ["", "التنين/DTNN"], ["", "لأخذ/VBD"], ["", "عينات/NNS"], ["", "وتصوير/NN"], ["", "عمود/NN"], ["", "الغازات/DTNNS"], ["", "الضارة/DTJJ"], ["", "التي/WP"], ["", "ينفثها/VBP"], ["", "بركان/NN"], ["", "تريالبا/NNP"], ["", "قرب/NN"], ["", "سان/NNP"], ["", "خوزيه/NNP"], ["", "كوستاريكا/NNP"]]]
Lemmatization	["In", "March", "NASA", "use", "a", "small", "electric", "military", "drone", "the", "Dragon", "Eye", "to", "sample", "and", "photograph", "the", "noxious", "gas", "plume", "spew", "from", "Turrialba", "Volcano", "near", "San", "Jose", "in", "Costa", "Rica"]	["في", "مارس", "ماضي", "استخدم", "ناسا", "طائرة", "عسكري", "كهربائي", "صغير", "دون", "طيار", "دعا", "عين", "تنين", "أخذ", "عينة", "تصوير", "عمود", "غاز", "ضار", "الذي", "نفث", "بركان", "خوزيه", "سان", "قرب", "تريالبا", "كوستاريكا"]

5.3. Regression results

In this section, we demonstrated different experimental results for CL-STS based on different sets of extracted features. Table 12 shows the regression results using three different architectures of deep neural networks that were overall lower than the classification results (highest regressor accuracy = 0.7113 vs. highest

Table 8
Sample of the individual feature values and their combinations for NER.

Pair#	En			Ar			Extracted Features			
	NER_{per}	NER_{loc}	NER_{org}	NER_{per}	NER_{loc}	NER_{org}	*f_per*	*f_loc*	*f_org*	*f_Mean*
1	[]	['Turrialba', 'Volcano', 'San', 'Jose', 'Costa', 'Rica.']	['NASA']	[]	['Turrialba', 'Volcano', 'San', 'Jose', 'Costa', 'Rica.']	['NASA']	0	1	1	0.6667
2	['Tom', 'McKinnon']	['Boulder']	['InventWorks']	['Tom', 'McKinnon']	['Boulder']	['InventWorks']	1	1	1	1
3	[]	[]	[]	[]	[]	[]	0	0	0	0
4	['Langdon']	[]	['American', 'University', 'of', 'Paris', 'Interpol', 'Langdon']	['Langdon']	[]	['American', 'University', 'of', 'Paris', 'Interpol', 'Langdon']	1	0	1	0.6667

Table 9
List of topics extracted by LDA algorithm.

Ar Topics:
Topic: (0, '0.019*"شخص"*0.010 + "علم"*0.012 + "صحة"*0.013 + "عمل"*0.015 + "طفل"')
Topic: (1, '0.029*"عمل"*0.021 + "نظم"*0.022 + "امم"*0.024 + "حقق"*0.025 + "تحد"')
Topic: (2, '0.056*"عمل"*0.016 + "جلس"*0.022 + "جمع"*0.024 + "لجنة"*0.026 + "قرر"')
Topic: (3, '0.021*"نظم"*0.010 + "اتفق"*0.011 + "عمل"*0.014 + "بلد"*0.014 + "دول"')
Topic: (4, '0.028*"وظف"*0.015 + "علم"*0.017 + "فترة"*0.017 + "عام"*0.025 + "بلغ"')
En Topics:
Topic: (0, '0.009*"country" + 0.009*"international" + 0.008*"cost" + 0.008*"right" + 0.007*"include"')
Topic: (1, '0.011*"estimate" + 0.010*"per" + 0.010*"Lady" + 0.008*"expenditure" + 0.006*"agree"')
Topic: (2, '0.010*"say" + 0.009*"one" + 0.009*"child" + 0.007*"go" + 0.007*"know"')
Topic: (3, '0.018*"report" + 0.018*"Committee" + 0.014*"SecretaryGeneral" + 0.014*"Assembly" + 0.014*"General"')
Topic: (4, '0.039*"United" + 0.033*"Nations" + 0.014*"International" + 0.014*"October" + 0.013*"Law"')

Table 10
Sample of the individual feature values and their combinations for *Topic_O* and *Topics_Sim* features.

Pair#	En			Ar			Extracted Features	
	Extracted topics	Max topic	Topics list	Extracted topics	Max topic	Topics list	*Topic_O*	*Topics_Sim*
1	[(1, 0.8688575), (2, 0.104606815)]	1	[0, 1, 2, 3]	[(0, 0.10488566), (1, 0.4897528), (2, 0.13927656), (3, 0.2572728)]	1	[0, 1, 2, 3]	0.5	0.608871
2	[(0, 0.10121918), (2, 0.090479225), (4, 0.79161435)]	4	[0, 2, 4]	[(0, 0.2858942), (2, 0.10533588), (4, 0.5932783)]	4	[0, 2, 4]	1	0.5697674
3	[(0, 0.016809892), (1, 0.6563794), (2, 0.29308107), (3, 0.016731204), (4, 0.016998438)]	1	[0, 1, 2, 3, 4]	[(0, 0.029032726), (1, 0.8835211), (2, 0.029053297), (3, 0.028829448), (4, 0.029563425)]	1	[0, 1, 2, 3, 4]	1	0.3579545
4	[(0, 0.011836619), (1, 0.48672372), (2, 0.10500127), (3, 0.011923938), (4, 0.38451448)]	1	[0, 1, 2, 3, 4]	[(0, 0.010070009), (1, 0.010139383), (2, 0.43503538), (3, 0.010076891), (4, 0.53467834)]	4	[0, 1, 2, 3, 4]	1	0.4071661

classifier result = 0.9701). This may be because estimating the degree of similarity as numerical values is more difficult and would require more learning epochs than in the case of classification tasks. When using two hidden layers, each with 10 neurons, the highest accuracy was obtained with *BoM_O* features; on the other hand, when increasing the number of neurons or hidden layers, the highest accuracy was obtained upon using *SRL_Sim*, *SpRL_O,* and *NE_O* features. Unlike classifiers, employing more features for the regressors may not increase accuracy results. As seen from the table below, the random baseline was not a suitable choice for regression, as the neural networks fail to learn the patterns from the features when assigned to random labels. Classical SVM achieved better accuracy, 0.8018, than deep networks regression models which proves that using simple machine learning regression models can be enough for this task.

6. Conclusion and future work

This paper attempts to explain the effect of using deep neural networks and different combinations of semantic connotations from texts in two different languages. In-depth processes of feature extractions were implemented on a considerable number of datasets taken from multiple sources and tagged into either *literally translated (LT), paraphrased (PT), summarized (ST)* or *independently written (IW).* The features we used include topic similarity, NER, SRL, SpRL, bag-of-stop words, and bag-of-meanings. Based on the evaluation results, deep neural networks can effectively learn to classify cross-lingual patterns as either plagiarized or independently written, and further classify the former into translated (LT), paraphrased (PT), and summarized (ST). Using a combination of semantic features can be effective regardless of these texts being written in two different languages.

In the future, we plan to study and analyze more semantic features using different linguistic resources and our own cases. Even though we used our samples from short to middle-length paragraphs, we plan to extend our experiments on cases of document length. Further, advanced types of deep learning including recurrent neural networks are objectives of future works, with the aim to learn the context of textual patterns and styles of writing that plagiarists may use.

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

Table 11
Classification results using three deep networks architectures and different sets of features for PD task.

Classifier	Feature	Predictors/ Inputs	Binary Classifier		4-Type Classifier	
			Binary cross-entropy	Accuracy %	Categorical cross-entropy	Accuracy %
Two hidden layers, each with 10 neurons, 100 epochs	*BoM_O*	(3 features)	0.5446	0.7559	0.6987	0.7210
	BoM_O + Topic_Sim + Topic_O	(5 features)	0.0951	0.9666	0.2594	0.9282
	BoM_O + SW_O	(7 features)	0.0956	0.9660	0.2679	0.9274
	SRL_Sim + SpRL_O + NE_O	(9 features)	0.5321	0.7668	0.7250	0.7227
	BoM_O + SRL_Sim + SpRL_O + NE_O	(11 features)	0.1029	0.9643	0.2813	0.9259
	All	(18 features)	0.0858	**0.9693**	0.2379	**0.9311**
Two hidden layers, each with 50 neurons, 100 epochs	*BoM_O*	(3 features)	0.5219	0.7590	0.6970	0.7229
	BoM_O + Topic_Sim + Topic_O	(5 features)	0.0945	0.9665	0.2566	0.9281
	BoM_O + SW_O	(7 features)	0.0940	0.9665	0.2607	0.9282
	SRL_Sim + SpRL_O + NE_O	(9 features)	0.5317	0.7670	0.7245	0.7230
	BoM_O + SRL_Sim + SpRL_O + NE_O	(11 features)	0.1014	0.9649	0.2788	0.9257
	All	(18 features)	0.0809	**0.9701**	0.2253	**0.9329**
Three hidden layers, 1st and 2nd layers with 10 neurons, and 3rd layer with 5 neurons, 100 epochs	*BoM_O*	(3 features)	0.5446	0.7559	0.7542	0.7113
	BoM_O + Topic_Sim + Topic_O	(5 features)	0.0957	0.9665	0.2610	0.9282
	BoM_O + SW_O	(7 features)	0.0958	0.9658	0.2712	0.9275
	SRL_Sim + SpRL_O + NE_O	(9 features)	0.5320	0.7669	0.7250	0.7227
	BoM_O + SRL_Sim + SpRL_O + NE_O	(11 features)	0.1021	0.9646	0.2812	0.9261
	All	(18 features)	0.0863	**0.9690**	0.2400	**0.9309**
Baseline 1 (random)	*All*	(18 features)	0.6925	0.5049	1.3860	0.2520
Baseline 2 (Word n-grams + fuzzy sentence)	*Word 5-gram fingerprinting*	–	–	0.3266	–	0.1255
	Word 8-gram fingerprinting	–	–	0.3014	–	0.0221
	Fuzzy sentence semantic similarity	–	–	0.6566	–	0.5498
Baseline 3 (SVM)	*All*	(18 features)	MSE: 0.0335	0.9664	MSE: 0.3393	0.9283
Baseline 4 (Logistic Regression)	*All*	(18 features)	MSE: 0.0335	0.9665	MSE: 0.3333	0.9288

Table 12
Regression results using three deep networks architectures and different sets of features for CL-STS task.

Regressor	Feature	Predictors/Inputs	MSE	Accuracy%
Two hidden layers, each with 10 neurons, 100 epochs	*BoM_O*	(3 features)	0.0999	**0.7113**
	BoM_O + Topic_Sim + Topic_O	(5 features)	0.0171	0.6921
	BoM_O + SW_O	(7 features)	0.0175	0.6916
	SRL_Sim + SpRL_O + NE_O	(9 features)	0.0992	0.7049
	BoM_O + SRL_Sim + SpRL_O + NE_O	(11 features)	0.0185	0.6910
	All	(18 features)	0.0159	0.6935
Two hidden layers, each with 50 neurons, 100 epochs	*BoM_O*	(3 features)	0.0190	0.6904
	BoM_O + Topic_Sim + Topic_O	(5 features)	0.0170	0.6917
	BoM_O + SW_O	(7 features)	0.0171	0.6918
	SRL_Sim + SpRL_O + NE_O	(9 features)	0.0993	**0.7108**
	BoM_O + SRL_Sim + SpRL_O + NE_O	(11 features)	0.0183	0.6910
	All	(18 features)	0.0149	0.6938
Three hidden layers, 1st and 2nd layers with 10 neurons, and 3rd layer with 5 neurons, 100 epochs	*BoM_O*	(3 features)	0.0952	0.7106
	BoM_O + Topic_Sim + Topic_O	(5 features)	0.0171	0.6916
	BoM_O + SW_O	(7 features)	0.0173	0.6918
	SRL_Sim + SpRL_O + NE_O	(9 features)	0.1038	**0.7113**
	BoM_O + SRL_Sim + SpRL_O + NE_O	(11 features)	0.0185	0.6905
	All	(18 features)	0.0154	0.6929
Baseline 1 (random)	*All*	(18 features)	9.4527e-07	0.0000
Baseline 2 (SVM)	*All*	(18 features)	0.0205	**0.8018**

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University (Grant #39-S-271).

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

References

AbdulJaleel, N., Larkey, L.S. Statistical transliteration for english-arabic cross language information retrieval. In: Paper presented at the Proceedings of the Twelfth International Conference on Information and Knowledge Management, New Orleans, LA, USA.

Alaa, Z., Tiun, S., Abdulameer, M., 2016. Cross-language plagiarism of Arabic-English documents using linear logistic regression. J. Theor. Appl. Inform. Technol. 83 (1), 20–33.

Alian, M., Awajan, A., 2018. Arabic semantic similarity approaches - review. In: 2018 International Arab Conference on Information Technology (ACIT), 28–30 Nov. 2018, pp. 1–6.

Aljlayl, M., Frieder, O. Effective arabic-english cross-language information retrieval via machine-readable dictionaries and machine translation. In: Paper presented at the Proceedings of the Tenth International Conference on Information and Knowledge Management, Atlanta, Georgia, USA.

Alotaibi, H., 2017. Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching, vol. 8.

Al-Smadi, M., Jaradat, Z., Al-Ayyoub, M., Jararweh, Y., 2017. Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. Inf. Process. Manage. 53 (3), 640–652. https://doi.org/10.1016/j.ipm.2017.01.002.

Alzahrani, S.M., 2016. Spatial role labelling in arabic using probabilistic classifiers. Int. J. Intell. Inform. Process. (ISSN: 2093-1964)

Alzahrani, S.M., Salim, N., Abraham, A., 2012. Understanding plagiarism linguistic patterns, textual features and detection methods. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 42 (2), 133–149.

Alzahrani, S.M., Salim, N., Palade, V., 2015. Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. J. King Saud Univ. – Comp. Inform. Sci. 27 (3), 248–268. https://doi.org/10.1016/j.jksuci.2014.12.001.

Alzahrani, S.M., 2015. Arabic plagiarism detection using word correlation in N-grams with K-overlapping approach, Working Notes for PAN-AraPlagDet at FIRE 2015. In: Paper presented at the Forum for Information Retrieval Evaluation, DAIICT, Gandhinagar, 4–6 December.

Barrón-Cedeño, A., 2010. On the mono- and cross-language detection of text reuse and plagiarism. 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Geneva, Switzerland.

Barrón-Cedeño, A., Rosso, P., Agirre, E., Labaka, G., 2010. Plagiarism detection across distant language pairs. In: 23rd International Conference on Computational Linguistics, Beijing, China, August 23–27 2010, pp. 37–45. Association for Computational Linguistics.

Barrón-Cedeño, A., Gupta, P., Rosso, P., 2013. Methods for cross-language plagiarism detection. Knowl.-Based Syst. 50, 211–217. https://doi.org/10.1016/j.knosys.2013.06.018.

Basile, C., Benedetto, D., Caglioti, E., Cristadoro, G., Esposti, M.D., 2009. A plagiarism detection procedure in three steps: selection, matches and "Squares". In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (Eds.), 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09. Donostia, Spain, pp. 19–23.

Bastianelli, E., Croce, D., Nardi, D., Basili, R., 2013. UNITOR-HMM-TK: structured kernel-based learning for spatial role labeling. In: Paper presented at the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval), Atlanta, Georgia, June 14–15.

Batet, M., Harispe, S., Ranwez, S., Sánchez, D., Ranwez, V., 2014. An information theoretic approach to improve semantic similarity assessments across multiple ontologies. Inf. Sci. 283, 197–210. https://doi.org/10.1016/j.ins.2014.06.039.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Machine Learn. Res. 3, 993–1022.

Boukhalfa, I., Mostefai, S., Chekkai, N. A study of graph based stemmer in arabic extrinsic plagiarism detection. In: Paper presented at the Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence, Rabat, Morocco.

Budanitsky, A., Hirst, G., 2006. Evaluating WordNet-based measures of lexical semantic relatedness. Comput. Linguist. 32 (1), 13–47. https://doi.org/10.1162/coli.2006.32.1.13.

Clarke, J., Srikumar, V., Sammons, M., Roth, D. An NLP Curator (or: How I Learned to Stop Worrying and Love NLP Pipelines). In: Paper presented at the LREC.

Corezola Pereira, R., Moreira, V., Galante, R., 2014. A new approach for cross-language plagiarism analysis. In: Agosti, M., Ferro, N., Peters, C., de Rijke, M., Smeaton, A. (Eds.), Multilingual and Multimodal Information Access Evaluation, vol. 6360. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 15–26.

Danilova, V., 2013. Cross-language plagiarism detection methods. In: Paper presented at the Proceedings of the Student Research Workshop associated with RANLP 2013, Hissar, Bulgaria, 9–11 September.

Diab, M., Moschitti, A., Pighin, D. CUNIT: a semantic role labeling system for modern standard Arabic. In: Paper presented at the Workshop on Semantic Evaluations (SemEval).

Edward, L., Steven, B., 2012. NLTK: the natural language toolkit. In: ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Philadelphia, Pennsylvania 2002, Association for Computational Linguistics, pp. 63–70.

Ehsan, N., Shakery, A., 2016. Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information. Inf. Process. Manage. 52 (6), 1004–1017. https://doi.org/10.1016/j.ipm.2016.04.006.

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

Ekinci, E., Omurca, S.İ., 2018. Babelfy-based extraction of collocations from Turkish Hotel reviews. In: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 28–30 Sept. 2018, pp. 1–5.

Ezzikouri, H., Oukessou, M., Youness, M., Erritali, M., 2018. Fuzzy cross language plagiarism detection (Arabic-English) using WordNet in a Big Data environment. In: Paper presented at the Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing, Barcelona, Spain.

Ezzikouri, H., Madani, Y., Erritali, M., Oukessou, M., 2019. A new approach for calculating semantic similarity between words using WordNet and set theory. Procedia Comput. Sci. 151, 1261–1265. https://doi.org/10.1016/j.procs.2019.04.182.

Finkel, J.R., Grenager, T., Manning, C., 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In: Paper presented at the the 43nd Annual Meeting of the Association for Computational Linguistics (ACL).

Franco-Salvador, M., Gupta, P., Rosso, P., 2014. Knowledge graphs as context models: improving the detection of cross-language plagiarism with paraphrasing. In: Ferro, N. (Ed.), Bridging Between Information Retrieval and Databases, vol. 8173. Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 227–236.

Franco-Salvador, M., Rosso, P., Montes-y-Gómez, M., 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. Inf. Process. Manage. 52 (4), 550–570. https://doi.org/10.1016/j.ipm.2015.12.004.

Franco-Salvador, M., Gupta, P., Rosso, P., Banchs, R.E., 2016. Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. Knowl.-Based Syst. 111, 87–99. https://doi.org/10.1016/j.knosys.2016.08.004.

Gharouit, K., Nfaoui, E.H., 2017. A comparison of classification algorithms for verbose queries detection using BabelNet. In: 2017 Intelligent Systems and Computer Vision (ISCV), 17–19 April 2017, pp. 1–5.

Gildea, D., Jurafskyy, D., 2000. Automatic labeling of semantic roles. In: 38th Annual Conference of the Association for Computational Linguistics (ACL-00), ACL, Hong Kong, pp. 512–520.

Glavaš, G., Franco-Salvador, M., Ponzetto, S.P., Rosso, P., 2018. A resource-light method for cross-lingual semantic textual similarity. Knowl.-Based Syst. 143, 1–9. https://doi.org/10.1016/j.knosys.2017.11.041.

Green, S., Manning, C.D. Better Arabic parsing: baselines, evaluations, and analysis. In: Paper presented at the COLING.

Gutiérrez-Batista, K., Campaña, J.R., Vila, M.-A., Martin-Bautista, M.J., 2018. An ontology-based framework for automatic topic detection in multilingual environments. Int. J. Intell. Syst. 33 (7), 1459–1475. https://doi.org/10.1002/int.21986.

Hadi, A.E.L., Madani, Y., Ayachi, R.E.L., Erritali, M., 2019. A new semantic similarity approach for improving the results of an Arabic search engine. Procedia Comput. Sci. 151, 1170–1175. https://doi.org/10.1016/j.procs.2019.04.167.

Haitch, R., 2016. Stealing or sharing? Cross-cultural issues of plagiarism in an open-source era. Teaching Theol. Religion 19 (3), 264–275. https://doi.org/10.1111/teth.12337.

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), 29 March–1 April 2016, pp. 372–374.

Hattab, E., 2015. Cross-language plagiarism detection method: Arabic vs. English. In: 2015 International Conference on Developments of E-Systems Engineering (DeSE), 13–14 Dec. 2015, pp. 141–144.

He, D., Wang, J., 2009. Cross-language information retrieval. In: Information Retrieval. John Wiley & Sons, Ltd, pp. 233–253.

Hussain, S.F., Suryani, A., 2015. On retrieving intelligently plagiarized documents using semantic similarity. Eng. Appl. Artif. Intell. 45, 246–258. https://doi.org/10.1016/j.engappai.2015.07.011.

Hussein, A.S., 2015. Arabic document similarity analysis using n-grams and singular value decomposition. In: 2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS), 13–15 May 2015, pp. 445–455.

Kasprzak, J., Brandejs, M., Křipač, M., 2009. Finding Plagiarism by Evaluating Document Similarities. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) 25th Conference of the Spanish Society for Natural Language Processing, SEPLN'09, Donostia, Spain 2009, pp. 24-28

Kolomiyets, O., Kordjamshidi, P., Bethard, S., Moens, M.-F., 2013. SemEval-2013 Task 3: spatial role labeling. In: Paper presented at the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, June 14–15.

Kordjamshidi, P., Bethard, S., Moens, M.-F. SemEval-2012 Task 3: spatial role labeling. In: Paper presented at the Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval), Stroudsburg, PA, USA, June 14–15.

Kordjamshidi, P., Otterlo, M.V., Moens, M.-F. Spatial role labeling: task definition and annotation scheme. In: Paper presented at the Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 19–21, May.

Kordjamshidi, P., Otterlo, M.V., Moens, M.-F., 2011. Spatial role labeling: towards extraction of spatial relations from natural language. ACM Trans. Speech Lang. Process. 8 (3), 1–36. https://doi.org/10.1145/2050104.2050105.

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

Kothwal, R., Varma, V., 2013. Cross lingual text reuse detection based on keyphrase extraction and similarity measures. In: Majumder, P., Mitra, M., Bhattacharyya, P., Subramaniam, L.V., Contractor, D., Rosso, P. (Eds.), Multilingual Information Access in South Asian Languages, vol. 7536. Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 71–78.

Kučečka, T. Obfuscating plagiarism detection: vulnerabilities and solutions. In: Paper presented at the Proceedings of the 12th International Conference on Computer Systems and Technologies, Vienna, Austria.

Leacock, C., Chodorow, M., 1998. Combining local context with WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), WordNet: A Lexical Reference System and its Application. MIT Press, Cambridge, MA, pp. 265–283.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K., 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowl. Data Eng. 18 (8), 1138–1150.

Liu, H., Bao, H., Xu, D., 2012. Concept vector for semantic similarity and relatedness based on WordNet structure. J. Syst. Softw. 85 (2), 370–381. https://doi.org/10.1016/j.jss.2011.08.029.

Liuling, D., Bin, L., Yuning, X., ShiKun, W., 2008. Measuring semantic similarity between words using HowNet. In: Computer Science and Information Technology, 2008. ICCSIT '08. International Conference on, Aug. 29 2008–Sept. 2 2008, pp. 601–605.

Lulu, L., Belkhouche, B., Harous, S., 2016. Candidate document retrieval for Arabic-based text reuse detection on the web. In: 2016 12th International Conference on Innovations in Information Technology (IIT), 28–30 Nov. 2016, pp. 1–6.

Magooda, A., Mahgoub, A., Rashwan, M., Fayek, M., Raafat, H., 2015. RDI System for Extrinsic Plagiarism Detection (RDI_RED) Working Notes for PAN-AraPlagDet at FIRE 2015.

Meng, F., Lu, W., Xue, R., 2017. Mapping senses in BabelNet to Chinese based on word embedding. In: 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 14–16 Oct. 2017, pp. 1–6.

Meng, F., Zhang, Y., Lu, W., Zhang, W., Cheng, J., 2017. Chinese word semantic relation classification based on multiple knowledge resources. In: 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 15–17 Dec. 2017, pp. 372–376.

Miller, G.A., 1995. WordNet: a lexical database for English. Commun. ACM 38 (11), 39–41.

Muneer, I., Sharjeel, M., Iqbal, M., Nawab, R.M.A., Rayson, P., 2019. CLEU - a cross-language english-urdu corpus and benchmark for text reuse experiments. J. Assoc. Inform. Sci. Technol. 70 (7), 729–741. https://doi.org/10.1002/asi.24074.

Nagoudi, E.M.B., Cherroun, H., Alshehri, A., 2018. Disguised plagiarism detection in Arabic text documents. In: 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), 25–26 April 2018, pp. 1–6.

Navigli, R., Ponzetto, S.P. BabelRelate! A joint multilingual approach to computing semantic relatedness. In: Paper presented at the Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence Toronto, Ontario, Canada, July 22–26.

Navigli, R., Ponzetto, S.P., 2012. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. 193, 217–250. https://doi.org/10.1016/j.artint.2012.07.0013, 217-250.

Navigli, R., Ponzetto, S.P., 2012. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. 193, 217–250. https://doi.org/10.1016/j.artint.2012.07.001.

Osman, A.H., Salim, N., Binwahlan, M.S., Alteeb, R., Abuobieda, A., 2012. An improved plagiarism detection scheme based on semantic role labeling. Appl. Soft Comput. 12 (5), 1493–1502. https://doi.org/10.1016/j.asoc.2011.12.021.

Paul, M., Jamal, S., 2015. An improved SRL based plagiarism detection technique using sentence ranking. Procedia Comput. Sci. 46, 223–230. https://doi.org/10.1016/j.procs.2015.02.015.

Pawar, A., Mago, V., 2019. Challenging the boundaries of unsupervised learning for semantic similarity. IEEE Access 7, 16291–16308. https://doi.org/10.1109/ACCESS.2019.2891692.

Pecorari, D., 2012. Plagiarism. In: The Encyclopedia of Applied Linguistics.

Pertile, S.d.L., Moreira, V.P., Rosso, P., 2016. Comparing and combining content- and citation-based approaches for plagiarism detection. J. Assoc. Inform. Sci. Technol. 67 (10), 2511–2526. https://doi.org/10.1002/asi.23593.

Pierce, J., Zilles, C., 2017. Investigating student plagiarism patterns and correlations to grades. In: Paper presented at the Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, Seattle, Washington, USA.

Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P., 2011. Cross-language plagiarism detection. Language Resour. Eval. 45 (1), 45–62.

Pu, H., Fei, G., Zhao, H., Hu, G., Jiao, C., Xu, Z., 2017. Short text similarity calculation using semantic information. In: 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM), 10–11 Aug. 2017, pp. 144–150.

Qin, P., Lu, Z., Yan, Y., Wu, F., 2009. A New Measure of Word Semantic Similarity Based on WordNet Hierarchy. IEEE Computer Society.

Qu, R., Fang, Y., Bai, W., Jiang, Y., 2018. Computing semantic similarity based on novel models of semantic representation using Wikipedia. Inf. Process. Manage. 54 (6), 1002–1021. https://doi.org/10.1016/j.ipm.2018.07.002.

Quan, Z., Wang, Z., Le, Y., Yao, B., Li, K., Yin, J., 2019. An efficient framework for sentence similarity modeling. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (4), 853–865. https://doi.org/10.1109/TASLP.2019.2899494.

Rashidghalam, H., Mahmoudi, F., 2015. Web query classification using improved visiting probability algorithm and babelnet semantic graph. In: 2015 AI & Robotics (IRANOPEN), 12–12 April 2015, pp. 1–5.

Rashidghalam, H., Taherkhani, M., Mahmoudi, F., 2016. Text summarization using concept graph and BabelNet knowledge base. In: 2016 Artificial Intelligence and Robotics (IRANOPEN), 9–9 April 2016, pp. 115–119.

Roberts, K., Harabagiu, S.M. UTD-SpRL: a joint approach to spatial role labeling. In: Paper presented at the First Joint Conference on Lexical and Computational Semantics (*SEM), Montréal, Canada, June 7–8.

Rubenstein, H., Goodenough, J.B., 1965. Contextual correlates of synonymy. Commun. ACM 8 (10), 627–633.

Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Networks 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Schmidt, A., Bühler, S., Senger, R., Scholz, S., Dickerhof, M., 2016. Detection and visual inspection of highly obfuscated plagiarisms. In: 2016 49th Hawaii International Conference on System Sciences (HICSS), 5–8 Jan. 2016, pp. 4113–4122.

Shajalal, M., Aono, M., 2018. Sentence-level semantic textual similarity using word-level semantics. In: 2018 10th International Conference on Electrical and Computer Engineering (ICECE), 20–22 Dec. 2018, pp. 113–116.

Shumin, W., Choi, J.D., propbanks., M.P.D.c.-l.s.s.u.p. Detecting Cross-lingual Semantic Similarity Using Parallel PropBanks. In: Paper presented at the 9th Conference of the Association for Machine Translation in the Americas Denver, Colorado.

Sorg, P., Cimiano, P., 2010. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In: Horacek, H., Métais, E., Muñoz, R., Wolska, M. (Eds.), Natural Language Processing and Information Systems, vol. 5723. Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 36–48.

Steinberger, R., 2012. Cross-lingual similarity calculation for plagiarism detection and more - tools and resources. In: Paper presented at the CLEF 2012 Evaluation Labs and Workshop, Rome, Italy, September 17–20.

Stoyanova, I., Koeva, S., Leseva, S. Wordnet-based cross-language identification of semantic relations. In: Paper presented at the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, 8–9 August.

Suleiman, D., Awajan, A., Al-Madi, N., 2017. Deep learning based technique for plagiarism detection in arabic texts. In: 2017 International Conference on New Trends in Computing Sciences (ICTCS), 11–13 Oct. 2017, pp. 216–222.

Tiedemann, J., 2016. Parallel corpora for everyone. Baltic J. Modern Comput. (BJMC) 4 (2).

Tiedemann, J., 2012. Parallel data, tools and interfaces in OPUS. In: Paper presented at the International Conference on Language Resources and Evaluation (LREC'2012), Istanbul, Turkey, May.

Tomassetti, F., Rizzo, G., Torchiano, M., 2014. Spotting automatically cross-language relations. In: Software Maintenance, Reengineering and Reverse Engineering (CSMR-WCRE), 2014 Software Evolution Week - IEEE Conference on, 3–6 Feb. 2014, 2014, pp. 338–342.

Toutanova, K., Klein, D., Manning, C.D., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, Edmonton, Canada.

Ustalov, D., Panchenko, A., 2017. A tool for effective extraction of synsets and semantic relations from BabelNet. In: 2017 Siberian Symposium on Data Science and Engineering (SSDSE), 12–13 April 2017, pp. 10–13.

Vani, K., Gupta, D., 2017. Detection of idea plagiarism using syntax–semantic concept extractions with genetic algorithm. Expert Syst. Appl. 73, 11–26. https://doi.org/10.1016/j.eswa.2016.12.022.

Vani, K., Gupta, D., 2017. Text plagiarism classification using syntax based linguistic features. Expert Syst. Appl. 88, 448–464. https://doi.org/10.1016/j.eswa.2017.07.006.

Vani, K., Gupta, D., 2018. Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: comparisons, analysis and challenges. Inf. Process. Manage. 54 (3), 408–432. https://doi.org/10.1016/j.ipm.2018.01.008.

Volk, M., Ripplinger, B., Vintar, Š., Buitelaar, P., Raileanu, D., Sacaleanu, B., 2002. Semantic annotation for concept-based cross-language medical information retrieval. Int. J. Med. Inf. 67 (1–3), 97–112. https://doi.org/10.1016/S1386-5056(02)00058-8.

Vulic, I., Moens, M., 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In: Paper presented at the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, June 9–14.

Vulic, I., Moens, M.-F., 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In: Paper presented at the The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014) Doha, Qatar October 25–29.

Wanjawa, B.W., Muchemi, L., 2018. Automatic semantic network generation from unstructured documents – the options. In: 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI), 21–22 Nov. 2018, pp. 72–78.

Wu, Z., Palmer, M., 1994. Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico State University, New Mexico, pp. 133–139.

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009

Xu, W., Callison-Burch, C., Dolan, B., 2015. SemEval-2015 task 1: paraphrase and semantic similarity in Twitter (PIT). In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics, Denver, Colorado, pp. 1–11.

Xu, L., Sun, S., Wang, Q., 2016. Text similarity algorithm based on semantic vector space model. In: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 26–29 June 2016, pp. 1–4.

Ye, P., Baldwin, T., 2006. Semantic role labeling of prepositional phrases. ACM Trans. Asian Lang. Inf. Process. 5 (3), 228–244. https://doi.org/10.1145/1194936.1194940.

Ye, Z., Huang, J.X., He, B., Lin, H., 2012. Mining a multilingual association dictionary from Wikipedia for cross-language information retrieval. J. Am. Soc. Inform. Sci. Technol. 63 (12), 2474–2487. https://doi.org/10.1002/asi.22696.

You, B., Liu, X.-r., Li, N., Yan, Y.-s., 2012. Using information content to evaluate semantic similarity on HowNet. In: Computational Intelligence and Security (CIS), 2012 Eighth International Conference on, 17–18 Nov. 2012, pp. 142–145.

Zhang, P., Zhang, Z., Zhang, W., Wu, C., 2014. Semantic similarity computation based on multi-feature combination using HowNet. J. Softw. 9 (9), 2461–2466.

Zhou, D., Truran, M., Brailsford, T., Wade, V., Ashman, H., 2012. Translation techniques in cross-language information retrieval. ACM Comput. Surv. 45 (1), 1–44. https://doi.org/10.1145/2379776.2379777.

Zhu, G., Iglesias, C.A., 2017. Sematch: semantic similarity framework for knowledge graphs. Knowl.-Based Syst. 130, 30–32. https://doi.org/10.1016/j.knosys.2017.05.021.

Ziemski, M., Junczys-Dowmunt, M., Pouliquen, B., 2016. The United Nations Parallel Corpus. In: Paper presented at the Language Resources and Evaluation (LREC'16), Portorož, Slovenia, May.

Please cite this article as: S. Alzahrani and H. Aljuaid, Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.04.009