Noname manuscript No.
(will be inserted by the editor)

On the Use of Word Embedding for Cross Language Plagiarism Detection

Habibollah Asghari · Omid Fatemi · Salar Mohtaj · Heshaam Faili · Paolo Rosso

Received: date / Accepted: date

Abstract Cross language plagiarism is the unacknowledged reuse of text across language pairs. It occurs if a passage of text is translated from source language to target language and no proper citation is provided. Although various methods have been developed for detection of cross language plagiarism, less attention has been paid to measure and compare their performance, especially when tackling with different types of paraphrasing through translation. In this paper, we present a novel approach to cross language plagiarism detection using word embedding methods and explore its performance against other state-of-the-art plagiarism detection algorithms. In order to evaluate the methods, we have constructed an English-Persian bilingual plagiarism detection corpus (referred to as HAMTA-CL) including seven types of obfuscation. This corpus can measure the effectiveness of cross language plagiarism detection methods against a low resource language like Persian. The results show that the word embedding approach outperforms the other approaches with respect to recall when encountering heavily paraphrased passages. On the other hand, translation based approaches perform well when the precision is the main consideration of the cross language plagiarism detection system.

Keywords Cross-Language Plagiarism Detection · Low resource Languages · Distant language pairs · Text re-use

H. Asghari
School of Electrical and Computer Engineering, University of Tehran, Iran
E-mail: habib.asghari@ictrc.ac.ir

O. Fatemi
School of Electrical and Computer Engineering, University of Tehran, Iran

S. Mohtaj
ICT Research Institute of ACECR, Tehran, Iran

H. Faili
School of Electrical and Computer Engineering, University of Tehran, Iran

P. Rosso
Universitat Politecnica de Valencia, Spain

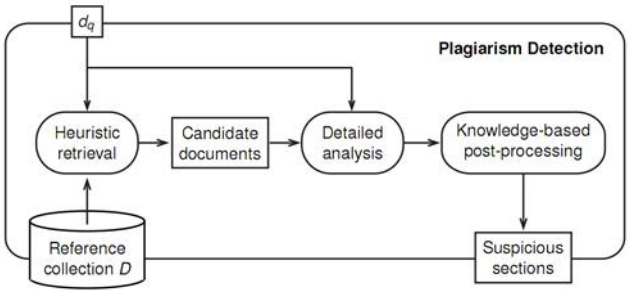

Fig. 1 Generic retrieval process for external plagiarism detection

1 Introduction

Plagiarism is the unacknowledged reuse of others' ideas or text without giving a proper credit (Stein, Stamatatos, & Koppel, 2008). In recent years, researchers enjoy easy access to a wide range of information via the Internet, especially across languages. Unfortunately, this also causes plagiarism occurs more simply. There are many attempts to detect plagiarism, especial across languages. In a research accomplished by (Stein, zu Eissen, & Potthast, 2007), a generic three-step retrieval process for a plagiarism detection (PD) system was proposed. They have presented the retrieval process for external plagiarism detection as depicted in Figure 1. In the candidate retrieval step, a heuristic task of retrieval potential source documents is done. In the text alignment step, an exhaustive comparison of suspicious document against selected source documents is applied. In the final stage, named as knowledge-based post-processing step, those detected fragments with proper citation are discarded as they are not plagiarized. The result is offered to the human expert to take the final decision. It should be noted that textual similarity detection methods are not exactly the methods to detect plagiarism. Plagiarism occurs when someone deliberately copy a passage of text without attribution, while these methods only detect textual similarities. Therefore, it is not enough to just recognize text similarities and to consider these similarities to plagiarism (Ferrero, Agnès, Besacier, & Schwab, 2016).When plagiarism is generated by a translation process, it is known as cross-language plagiarism. In other words, the problem does not end at language boundaries. Nowadays, a vast amount of knowledge is created in rich resource languages like English, and students and researchers in low resource languages have a motivation to bring the knowledge to their language through translation. Cross Language Plagiarism Detection (CLPD) systems try to find plagiarism cases between language pairs. Cross-language plagiarism detection is to identify the text reuse given suspicious documents in one language L_1 and the possible source document in L_2. Text reuse detection across languages is even harder if the detection of text reuse is between distant language pairs (Barrón-Cedeño, Rosso, Agirre, & Labaka, 2010). In cross-language plagiarism detection, the source and suspicious documents are in different languages. In this paper we have focused on English-Persian language pairs. Therefore, the task of a CLPD system is to find the source document(s) in English for the given suspicious document in Persian for the probable text re-use.

There are some drawbacks of applying current algorithms to CLPD in Persian. Our research objective is to challenge with the following problems:

- Persian is a less-resourced language which has a low degree of representation on the Web. They are often referred to as low profile languages. There is a pressing need to develop NLP algorithms and techniques for less-resourced languages. Because of low resources in Persian, machine translation tools do not work well, especially when we deal with passages that are heavily paraphrased.
- Persian and English are distant languages, so some approaches such as ordinary cross-lingual character n-gram (CL-CNG) cannot be applied in an English-Persian text reuse detection systems.
- Persian is an Arabic-Script based language. There are many problems to basic preprocessing tasks in this language such as normalization, stemming and recognizing word and multi-token word boundaries (Farghaly & Shaalan, 2009).
- Paraphrasing and translation can be considered as connected natural language tasks. Various types of paraphrasing can be done through translating a text passage from a language into other languages. CLPD systems would have different performance facing different types of paraphrasing. We compiled a plagiarism detection corpus with different types of paraphrasing such as summarizing, splitting the sentence to two or more sentences, merging two or more sentence to one sentence and heavy paraphrasing of the sentence in the target language. To the best of our knowledge, no work has been done considering these differing types of paraphrasing.

Word embedding methods showed their effectiveness in text similarity in recent years (Gouws, Bengio, & Corrado, 2015). Moreover, the approaches based on semantic networks are also of great importance (Franco-Salvador, Gupta, & Rosso, 2013). In this paper, we compare the effectiveness of five different categories of algorithms for the task of CLPD and compare them with the simple Translation plus Mono-lingual Analysis (T+MA) approach. T+MA is a simple way of translating the suspicious document and doing a monolingual plagiarism detection task. However, it is constrained by the availability and quality of translators; in other words, it is limited to and upper bounded by quality of Machine Translation tools (Gupta & Singhal, 2011).

The rest of the paper is organized as follows. Section 2 describes some of the recent works in the field of cross-language plagiarism analysis. Section 3 presents our methodology for approaching various CLPD algorithms with broad types of parameters. The data preparation and corpus construction for evaluating the algorithms are also described in this section. Section 4 gives a detailed description of the experiments carried out in our work. Finally, Section 5 includes discussion and future work.

2 Related Work

In this section, some of the previous methods on cross-language plagiarism detection are presented. Figure 2 depicts the taxonomy of various approaches on CLPD. As shown in the figure, there are four main categories for CLPD. Moreover, recent works pay attention to combine different approaches to benefit from advantages of two or more methods. In the following subsections, we describe the recent approaches based on the above mentioned taxonomy.

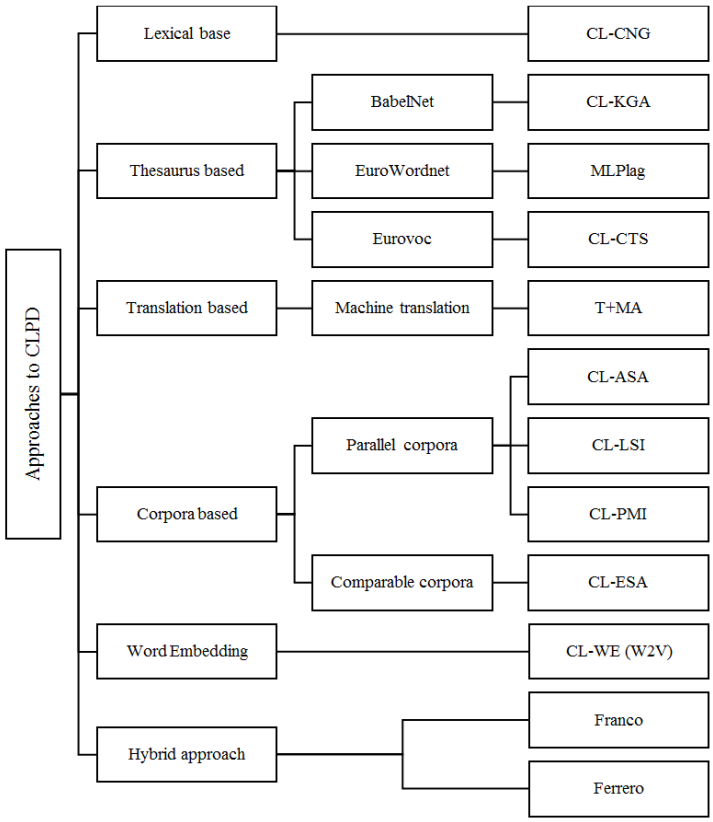

Fig. 2 Taxonomy of approaches to CLPD

2.1 Lexical based approaches

Lexical based approaches try to compare multilingual documents without using translation systems or any multi-lingual resources of data. They analyze cross-lingual similarity considering the structural and lexical similarity between languages. Cross-Language Character N-Gram model (CL-CNG), which uses overlapping character N-gram tokenization, has been proposed in (McNamee & Mayfield, 2004). The method is based on the fact of lexical similarity between languages sharing similar syntactic structure (e.g., related European language pairs). The obtained results show a competitive accuracy with respect to language-specific approaches for European languages. This approach can compare multilingual documents without using translation systems. However, due to lexical differences and different writing alphabets between distant languages with different lexicon, this method cannot be applied for detecting cases of similarity while encountering different lexicon.

2.2 Thesaurus based approaches

Thesaurus based approaches use multi-lingual resources to transform passages in different languages into a unique language independent form. The BabelNet and EuroWordNet are the most popular resources for different cross language tasks including CLPD. BabelNet is a very large, wide-coverage multilingual semantic network which is constructed automatically by integrating lexicographic and encyclopedic knowledge from WordNet and *Wikipedia* (Navigli & Ponzetto, 2010). The BabelNet version 3.7 covers more than 270 languages and made up of about 14 million entries, called Babel synsets [1]. The EuroWordNet is a multilingual database of words and their relations for most European languages (i.e. English, Danish, Italian, Spanish, German, French and Czech) and contains sets of synonyms and relations between them (Ceska, Toman, & Jezek, 2008).

MLPlag system is proposed in (Ceska et al., 2008) based on analysis of word positions for plagiarism detection across languages. The proposed approach utilizes the EuroWordNet thesaurus which transforms words into language independent form. In the case of ambiguous words, two words from different languages have been considered as plagiarized if one of the senses matches with the one in the other language. They compared the influence of multilingual pre-processing on the accuracy and also two different similarity measures, named as symmetric and asymmetric measures (Ceska et al., 2008).

An approach to identify very similar documents among a collection of candidate documents has been proposed in (Pouliquen, Steinberger, & Ignat, 2006). The proposed method is based on representing the document contents by a vector of thesaurus terms from a multilingual thesaurus, and measuring similarity between the vectors. In their proposed method, they used a "Length Factor based on the observation of differences between the lengths of original and translated texts in Spanish, French and English. They found that the variation of the length difference approximately follows a normal distribution and considered it as a factor for computing similarity between documents. The proposed length factor has also been used as a separated score for measuring cross-lingual text similarity for CLPD in (Gupta, Barrón-Cedeño, & Rosso, 2012).

CL-CTS method is proposed by (Gupta et al., 2012) to measures the cross-lingual similarity based on a conceptual thesaurus by representing documents in the conceptual space using a domain specific Eurovoc conceptual thesaurus. The proposed model represents documents as vectors after filtering stop words, stemming and using term frequency weighting schema to build the vectors. At the final step, they compare the similarity between vectors using the cosine similarity measure in addition to named entities matching and "Length Model similarity.

A knowledge graph-based approach is proposed in (Franco-Salvador et al., 2013) by using BabelNet to obtain and compare context models of document fragments in different languages. To build their knowledge graph, at the first step a set of concepts in each fragment of text is extracted in different languages. In the next step, they obtain a set of paths (P) by searching the BabelNet for paths between each pairs of concepts. The knowledge graph is constructed by joining the paths from P. Then the concepts and relations have been weighted based on the degree of relatedness. Finally, to compare pairs of fragments in different languages,

[1] http://www.babelnet.org/stats

the resulted graphs are compared based on conceptual graph similarity algorithm. Their results show better performance of the proposed model with respect to other lexical and distributional based models (Franco-Salvador et al., 2013).

2.3 Translation based approaches

Translation based models use dictionaries and machine translation systems to translate suspicious document into the source language and then do a mono-lingual analysis. Automatic machine translation tools have been used by many researches for detecting cases of cross-lingual plagiarism (Pereira, Moreira, & Galante, 2010; Nawab, Stevenson, & Clough, 2010; Oberreuter, L'Huillier, Rios, & Velásquez, 2011; Kent & Salim, 2009). In this approach, the suspicious documents translated to the same language as the source documents, and mono-lingual PD methods applied to find plagiarized cases. Different monolingual PD methods (e.g. word N-Gram similarity and vector space model similarity detection) can be applied to compare resulted mono-lingual documents. The accuracy of CLPD systems based on these approaches is restricted by the accuracy of available machine translation systems.

2.4 Corpora based approaches

Corpora based approaches use different multi-lingual resources to train similarity detection models. Although most methods use sentence aligned parallel corpora, some approaches have been proposed based on using comparable resources.

Cross-language explicit semantic analysis (CL-ESA) retrieval model is proposed in (Potthast, Stein, & Anderka, 2008) for cross-language similarity analysis. The proposed model is an extension to previously proposed explicit semantic analysis (ESA) model. ESA uses a document collection (D) with n documents, and measure the cosine similarity of target document (d) with the collection. In the proposed model, each document can be represented with a vector of n dimensions, where the i_{th} index in V shows the cosine similarity between d and the i_{th} document of D. The similarity between two documents under the ESA model is defined as the similarity between resulted vectors (e.g. cosine similarity). CL-ESA uses same principle to compare documents in different languages. For this purpose, a collection of comparable *Wikipedia* documents in different languages (D) is used to measure the cosine similarity between document d in the language L1 with the collection D in the same language. Like ESA, the similarity between documents can be calculated by measuring the cosine similarity between the resulted vectors (Potthast et al., 2008).

Cross-lingual latent semantic indexing (CL-LSI) has been proposed in (Rehder, Littman, Dumais, & Landauer, 1997) to construct a multilingual semantic space. LSI creates a reduced-dimension feature space by applying singular value decomposition (SVD) on word-document matrix, in which words that occur in similar contexts are near to each other. The proposed CL-LSI method uses manually or automatically translated documents to create a set of bilingual training documents. Based on the structure of training documents that contain terms from both languages, the resulting LSI model is a bilingual vector space.

An approach for cross lingual plagiarism detection by using statistical bilingual dictionaries based on the IBM-1 alignment model has been proposed in (Barrón-Cedeño, Rosso, Pinto, & Juan, 2008; Pinto, Civera, Barrón-Cedeño, Juan, & Rosso, 2009). Given the suspicious and reference texts x and y (written in different languages), their goal is to answer the question "Is x plagiarized (and translated) from y? To achieve this goal, they divided documents into fragments and the objective was to know if a suspicious fragment x was a plagiarism case from one of the reference fragments y. In order to determine if x is plagiarized from any y fragment, the following probability had been calculated for each pairs of fragments.

This model has been tested on a mini-corpus of original plagiarized pair of texts. Moreover, in (Pinto et al., 2009) the proposed statistical approach based on IBM1 has been evaluated on different cross- lingual tasks of NLP such as bilingual text classification, cross-language information retrieval and cross-language plagiarism detection. In contrast to current approaches that ignore or do not take full advantage of multi-linguality, the aim of the presented approach was to capture word correlation across languages. The obtained results in different tasks show the benefits of the IBM1 model and the advantageous of learning cross-lingual information directly from cross-lingual resources.

2.5 Word Embedding approaches

Word Embedding (WE) methods, which map words or phrases to vectors of real numbers, have shown tremendous success in numerous NLP tasks in recent years. According to good performance of word embedding methods, some of the more traditional distributional representation models have been fully replaced with these novel approaches. Cross-lingual word embedding models try to learn features (embedding) for each word in such a way that similar words in each language are assigned similar embedding (that meets monolingual objective function), and also similar words across languages to have similar representations (that meets cross-lingual objective function) (Gouws et al., 2015). To achieve this goal, different bilingual resources (i.e. parallel corpora, word aligned corpora or comparable corpora) have been used by different approaches. BILBOWA (Bilingual Bag-of-Words without Alignments) is proposed as a model to learn bilingual distributed representations of words which can scale to large monolingual datasets and don't require word-aligned parallel training data (Gouws et al., 2015). The BILBOWA combines advances of monolingual word embedding with a particularly efficient novel sampled cross-lingual objective function.

2.6 Hybrid Approaches

In addition to the mentioned approaches for measuring cross-lingual similarity and cross-language plagiarism detection which use different resource to train their models, some of the recent hybrid methods try to combine the benefits of different approaches to improve accuracy.

A new model based on knowledge graph and continuous space representation of words has been proposed in (Franco-Salvador, Rosso, & Montes-y Gómez, 2016; Franco-Salvador, Gupta, Rosso, & Banchs, 2016). The presented method

basically follows the previously proposed CL-KGA model (Franco-Salvador et al., 2013). For weighting the obtained BabelNet semantic relations, instead of using the BabelNet's relation weights, the continuous Skip-gram and SenVec models have been used in this model (Franco-Salvador, Rosso, & Montes-y Gómez, 2016). Moreover, in this research the impact of relevant aspects of the model has been studied for the task of CLPD which includes: word sense disambiguation (WSD), vocabulary expansion, language independence and representation by similarities with a collection of concepts. The obtained results show the importance of WSD for improving the model's performance for the task of cross-language plagiarism detection (Franco-Salvador, Rosso, & Montes-y Gómez, 2016).

Ferrero et al. presented different syntax-based, dictionary-based, context-based and MT-based methods and a hybrid method by combining some of these approaches for cross-lingual textual similarity for SemEval-2017 shared task, named as CompiLIG system (Ferrero, Besacier, Schwab, & Agnès, 2017). Among all of their runs, the Cross-Language Conceptual Thesaurus-based Similarity (CL-CTS) achieved the best result, which consists of representing texts as bag of words (or concepts) to compare them (Ferrero et al., 2017). As a hybrid method, the most similar words from the embedding space have been added to the main concepts of the sentences from a multi-lingual semantic network. In other words, they use word embedding methods to enrich the basically extracted concept from the thesaurus (Ferrero et al., 2017).

In this paper, we have investigated various approaches that are proved to be efficient in CLPD and compared them with each other. It should be noted that a comprehensive investigation and comparison of monolingual plagiarism detection algorithms in Persian has been done in a PAN-FIRE shared task on plagiarism detection (Asghari et al., 2016). In this paper, we focused on English-Persian CLPD. Shortly our contributions are as follow:

- Benchmarking of recent approaches to CLPD and cross-lingual text similarity detection
- Investigating the performance of CLPD approaches applied to low resource languages (e.g. Persian)
- Applying the above mentioned approaches on the HAMTA-CL corpus with various types of obfuscation (paraphrasing)

Moreover, in this research we have investigated cross lingual word embedding methods on the task of plagiarism detection and compared it to the other previously proposed approaches. In order to compare the performance of the proposed approach, we have applied the above mentioned approaches on the HAMTA-CL corpus with various types of obfuscation.

3 Our Approach

In this section, the selected methods and evaluation framework for measuring the performance of algorithms are described in detail.

3.1 Selecting the Algorithms

In order to compare the performance of the word embedding method (BILBOWA) against the other algorithms, a collection of state-of-the-art approaches including CL-ESA, CL-KGA, CL-LSI and T+MA were selected and were applied on the proposed CLPD corpus. Due to lexical and syntactical differences between Persian and English as distant languages, the CL-CNG method cannot be applied for detecting cases of similarity, so we ignored this method.

CL-ESA: Cross-language explicit semantic analysis (CL-ESA) proposed in (Potthast et al., 2008) as a cross-lingual retrieval model. In the proposed model, a document d in the language l can be represented as an ESA vector d, using the cosine similarity with the index collection D in the corresponding language l. Also, a document d' in the language l' can be presented as a vector d' by computing the cosine similarity of d' with the index collection D' in language l'. The similarity between two documents under the ESA model is defined as the similarity between the resulted vectors (e.g. cosine similarity).

As mentioned in (Potthast et al., 2008), the collection D should contain documents from a broad range of domains, and each index document should be of "reasonable length. While a subset of the documents in *Wikipedia* can fulfills both properties, for the training phase, we used a collection of 200 comparable articles from *Wikipedia*. The selected articles covers a board ranges of topics, contains both Persian and English *Wikipedia* pages and also contains more than pre-defined 500 words length in both languages.

In our experiments, the suspicious and corresponding source documents have been split into sentences. We embed each sentence in the source and suspicious documents into vectors using CL-ESA. For this purpose, each sentence has been compared with a collection of 20000 documents under cosine similarity measure. The Persian sentences have been compared with Persian *Wikipedia* pages and English ones have been compared against equivalent English pages. To detect cases of plagiarism between documents, the cosine similarity between derived vectors in two documents has been computed.

BILBOWA: This algorithm is a fast and simple method for learning distributed representation of bilingual words (Gouws et al., 2015). BILBOWA does not rely on word aligned parallel data, and this makes the algorithm appropriate for less resourced languages (e.g. Persian). The model tries to learn both mono and cross lingual word embedding using joint optimization. The well-known monolingual word embedding methods (e.g. CBOW and Skip-Gram (Mikolov, Chen, Corrado, & Dean, 2013)) have been used to train mono-lingual vectors.

In our experiments, the source and suspicious documents are split into sentences. The BILBOWA has been used to convert constitutive words into 200 dimensional vectors. We used a simple averaging approach to combine word-vectors to create vectors of sentences. It has shown that the averaging approach has the best performance for the task of sentence embedding for semantic similarity detection (Wieting, Bansal, Gimpel, & Livescu, 2015). The resulted vectors of sentences have been compared using cosine similarity measure to detect cases of similarity between source and suspicious documents.

CL-LSI: The goal of cross-lingual latent semantic indexing (CL-LSI) is to construct a multilingual semantic space. The proposed CL-LSI method uses manually or automatically translated documents to create a set of bilingual training docu-

ments. Based on the structure of training documents that contain terms from both languages, the resulting LSI model is a bilingual vector space. In our experiment, we train a model using CL-LSI on sentence aligned parallel corpora. For detecting cases of text similarity between source and suspicious documents, both documents have been split to sentences. The created LSI model has been used to convert each sentence into low-dimensional LSI space. The resulted vectors of sentences have been compared using cosine similarity measure to detect cases of similarity between source and suspicious documents.

T+MA: In the Translation plus Mono-lingual Analysis approach, the suspicions documents have been translated from Persian into English, using Google translate API. The Vector Space Model (VSM) method is used to convert sentences of resulted English documents and source documents into vectors. Like previous models, the resulted vectors of sentences have been compared using the cosine similarity measure to detect cases of similarity between source and suspicious documents.

3.2 Evaluation Framework

For investigating the performance of the CLPD algorithms, an evaluation framework is required. The framework is comprised of an evaluation corpus along with evaluation measures. In the following subsections we will thoroughly describe the HAMTA-CL English-Persian corpus that we built and the measure that we employed for evaluation.

3.2.1 Corpus Construction

In order to compare the performance of different algorithms on English-Persian plagiarism detection, a CLPD evaluation corpus should be constructed. In this section, we first review some of the recently developed corpora and then describe our methodology for building an English-Persian plagiarism detection corpus.

The PAN plagiarism detection corpus PAN-PC-09 includes a set of cross-language plagiarism cases across two language pairs (Potthast, Stein, Barrón-Cedeño, & Rosso, 2010). Out of different types of obfuscation, more than 10% have been covered by cross-language cases of plagiarism, which includes automatically translated plagiarized fragments from German and Spanish to English. Subsequent PAN-PC-10 (Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010) and PAN-PC-11 (Potthast, Eiselt, Barrón-Cedeño, Stein, & Rosso, 2011) corpora contains 14% and 11% cross-language cases of plagiarism, respectively. Moreover, for improving the quality of cross-language corpus, 1% of automatically translated fragments of PAN-PC-11 have been manually corrected. A cross-language plagiarism detection corpus is constructed in (Ceska et al., 2008) for evaluating CLPD methods using JRC-EU and Fairy-tale multilingual corpora. The proposed corpus consists of 200 English reports from JRC-EU and 27 English document of Fairy-tale as source documents and the same number of documents in Czech as the suspicious ones. A cross-language PD corpus has been compiled in (Potthast, Barrón-Cedeño, Stein, & Rosso, 2011) using 23,000 JRC-Acquis parallel corpus documents and 45,000 Wikipedia documents, in which 10,000 aligned documents have been used to test the algorithms. In PAN 2015 text alignment shared task, The first English-Persian

corpus has been proposed in (Asghari, Khoshnava, Fatemi, & Faili, 2015). An English-Persian sentence aligned parallel corpus is used to compile cases of plagiarism across the two languages. Plagiarized fragments in suspicious document have been constructed from Persian sentences and corresponding source fragments have been constructed from English sentences. To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with different similarity scores were chosen. The number of sentences and their similarity score in a fragment specifies the four degree of obfuscation in the fragments. Unlike the prosed corpus in (Asghari et al., 2015), in this paper we proposed a CLPD corpus with different type of obfuscations. For the CL!TR task on cross-language text re-use detection across Hindi and English languages a corpus which includes 5032 English documents from Wikipedia and 388 Hindi documents has been used (Barrón-Cedeño, Rosso, Devi, Clough, & Stevenson, 2011). To generate cases of plagiarism, the participants are asked to write a short answer to a set of questions either by re-using the source documents or by using learning materials. To simulate real cases of plagiarism, they asked participants to answer questions with 4 different levels of obfuscation including: near copy, light revision, heavy revision and no-plagiarism (Barrón-Cedeño et al., 2011). A multi-style multi-granularity corpus for cross-language textual similarity detection has been proposed by (Ferrero et al., 2016). The proposed corpus is in French, English and Spanish and is based on a parallel corpus along with a comparable corpus. Both human translated texts from multiple types of authors and also machine translated texts have been used for constructing the corpus. They have prepared different granularities in document-level, sentence level and chunk-level (noun chunks).In constructing a plagiarism detection corpus, some text fragments from source document should be inserted into the suspicious document in order to simulate plagiarism. In order to have more realistic cases of plagiarism, the text fragments should be paraphrased (obfuscated). Paraphrasing are alternative ways of conveying the same information (Bannard & Callison-Burch, 2005), so plagiarists use paraphrasing to change the word forms while keeping the same meaning. In our approach to construct the HAMTA-CL corpus, we have focused on creating various types of paraphrasing. None of the above mentioned methods have considered such versatile types of paraphrasing in creating bilingual PD corpora.It should be mentioned that translation is inherently a paraphrasing mechanism. In (Bannard & Callison-Burch, 2005) is showed that the task of generating paraphrases can be accomplished using bilingual parallel corpora. They have also defined a paraphrase probability derived from a phrase-based statistical machine translation (SMT) approach that allows paraphrases to be ranked by translation probabilities. Callison-Burch has investigated how paraphrasing can be accomplished via translation (Bosma & Callison-Burch, 2006). As a result, in order to incorporate different kinds of paraphrasing techniques into a bilingual CLPD corpus, we have considered the following obfuscation approaches:

- **Simple Translation (STR):** Creating plagiarized passages by combining topically related sentences from a parallel corpus.
- **Artificial (ART):** Creating plagiarized passages by combining topically related sentences from a parallel corpus, along with artificial obfuscation in the target language.
- **Paraphrasing (PAR):** Creating plagiarized passages by combining topically related sentences from a parallel corpus, along with human aided paraphrasing

Table 1 Corpus statistics

Document Purpose	Number of Documents	1904
	% of Source Documents (English)	59%
	% of Suspicious Documents (Persian)	41%
Document Length	Short (1-400 words)	67%
	Medium (400-2000 words)	28%
	Long (2000-17000 words)	5%
	Average number of words per document	482
	Average number of sentences per document	23
	Smallest document (by words)	55
	Largest document (by words)	16685

Table 2 Plagiarism case statistics

Case Length	Short (20 - 50 words)	36%
	Medium (50-100 words)	42%
	Long (100-300 words)	22%

in the target language. In this type of obfuscation, a monolingual paraphrasing is done in the target language regardless of the source language.
- **Summarization (SUM):** Translation plus human aided summarization of the passage in the target language.
- **Circular Translation (CTR):** Translation from source language L1 to a different language L3 and then translate it back into the target language L2.
- **Split (SPL):** Translation plus dividing the sentence in the target language into two or more sentences.
- **Merge (MRG):** Translation plus combining two or more sentences in the target language into one sentence.

Figure 3 demonstrates the flow diagram for construction the cross-language PD corpus with the above mentioned paraphrasing techniques.

Wikipedia articles are used as primary resource to create the HAMTA-CL corpus. Due to its scale, context and open accessibility, *Wikipedia* is the best available resource to compile such a corpus. Due to the importance of documents' length in compiling a realistic PD corpus, among the whole *Wikipedia* documents, 1904 documents with variety of lengths have been used to compile the proposed corpus. The statistics of documents is represented in Table 1.

Since in real situations the plagiarism can be done in different lengths, a broad range of lengths is considered to create potential plagiarized fragments. The lengths of fragments are distributed between 20 and 300 words to simulate all types of plagiarism as shown in Table 2, and the distribution of fragments' length is depicted in Figure 4. Moreover, the statistics of the proposed corpus and ratio of different types of obfuscations is represented in Table 3 .

As shown in the figure above, the Merge obfuscated fragments have the shortest length on average among all the passages. Also, the average length of Summarization obfuscated fragments are the longest among different types of passages. This is because we have selected long passages for the source documents in such a way that the summarization process could be easier for crowd-workers. Moreover, the mean length of all plagiarized passages except summarization is almost the same.

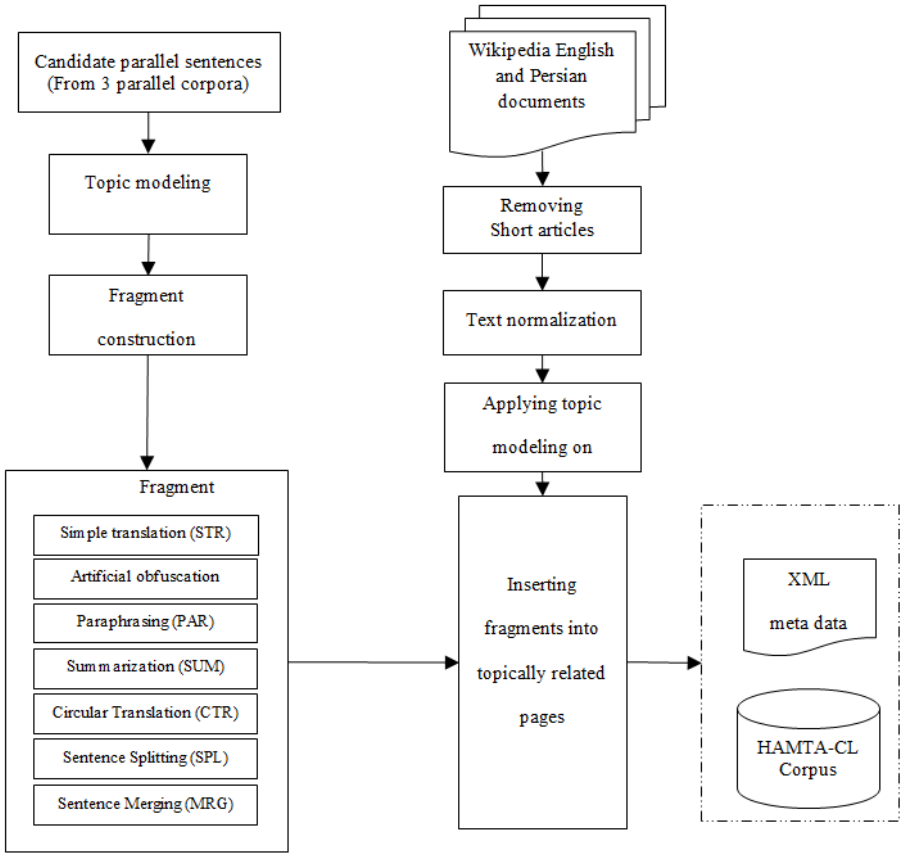

Fig. 3 Flow Diagram of Bilingual PD Corpus Construction

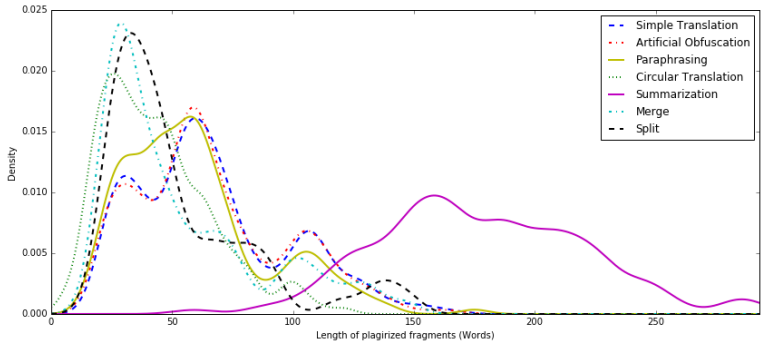

Fig. 4 Length distribution of the fragments

Table 3 Statistics of different types of obfuscation

Obfuscation	Number of fragments	% of fragments
Simple Translation (STR)	498	29%
Artificial (ART)	495	29%
Paraphrasing (PAR)	185	10%
Summarization (SUM)	134	8%
Circular Translation (CIR)	187	10%
Split (SPL)	144	9%
Merge (MRG)	58	5%

Table 4 Ratio of plagiarism fragments in documents

Plagiarism Per Document	Ratio
Hardly (5% - 10%)	50%
Medium (11% - 25%)	17%
Much (26% - 60%)	33%

We should also cover different situations concerning ratio of plagiarized fragments per suspicious document. To this aim, a wide range of plagiarism ratio is considered from hardly (i.e., low ratio of plagiarized fragments per suspicious document) to much (i.e., most parts of the document is plagiarism) as shown in Table 4.

Some efforts have been made in previous works to evaluate plagiarism detection corpora. Potthast et al. proposed some automatic and manual methods to evaluate and validate submitted corpora on the first shared task on plagiarism detection data submission (Potthast, Goering, Rosso, & Stein, 2015). Also, manual and automatic evaluation measures to evaluate PD corpora have been proposed in (Zarrabi, Rafiei, Khoshnava, Asghari, & Mohtaj, 2015). In the proposed automatic method, the character n-grams similarity has been used as a language independent measure to compare the ratio of obfuscation between the source and suspicious passages (Zarrabi et al., 2015).

The corpus was automatically validated considering the ratio of the length of plagiarized passages to the length of the documents, and the distribution of plagiarized passages across the documents. Moreover, a manual checking was done for evaluating the quality of plagiarized fragments. It should be noted that the constructed corpus is freely available to use for the research community [2].

3.2.2 Evaluation Measure

The ordinary measures for evaluating the performance of NLP algorithms are precision, recall and F-measure. In plagiarism detection tasks, we use character-level precision and recall. Besides this performance measures, another measure that characterizes the goodness of a detection algorithm have been defined in (Potthast, Stein, et al., 2010; Barrón-Cedeño, Potthast, Rosso, & Stein, 2010); whether a plagiarism case is detected as a whole or it has been detected in several pieces. Granularity quantifies whether the contiguity between plagiarized text passages is properly recognized. A low granularity simplifies both the human inspection of

[2] http://www.ictrc.ac.ir/corpus/HAMTA-CL.rar

algorithmically detected passages as well as an algorithmic style analysis within a potential post-process (?, ?). To capture this characteristic, they have introduced the granularity of R under S as follows:

$$gran(S, R) = \frac{1}{|S_{\mathrm{R}}|} \sum_{s \in S_{\mathrm{R}}} |R_{\mathrm{S}}| \tag{1}$$

The range of gran(S, R) is between [1, R], with 1 indicating the desired one-to-one correspondence and R indicating the worst case. Precision, recall (both at character-level) and granularity have been combined to an overall score based on following equations:

$$Precision(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{\bigcup_{s \in S}(s \sqcap r)}{|r|} \tag{2}$$

$$Recall(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{\bigcup_{r \in R}(s \sqcap r)}{|s|} \tag{3}$$

$$F_1 = 2.\frac{Precision.Recall}{Precision + Recall} \tag{4}$$

The three measures can be applied in isolation but also be combined into a single, overall performance score as follow:

$$Plagdet(S, R) = \frac{F_1}{(1 + gran(S, R))} \tag{5}$$

Where S denote the set of plagiarism cases in the suspicious documents of the corpus, and R denote the set of plagiarism that detected by detector for these documents, and F_1 denotes the $F - Measure$.

4 Experiments

We investigate the performance of CL-LSI, CL-ESA, BILBOWA, CL-KGA and T+MA on the task of English-Persian plagiarism detection in the following experiments.

4.1 Experiment 1

In this experiment, we have measured the performance of CL-ESA, CL-LSI, BILBOWA, CL-KGA and T+MA algorithms in detecting cases of plagiarism on the whole HAMTA-CL corpus. Our goal is to measure the performance of CLPD methods on the proposed corpus that contains various types of obfuscation.

The graphs of precision, recall and F1 measure versus different similarity thresholds are depicted in Figure 5. As shown in the figure, the T+MA algorithm obtains the best F1 and precision in the whole corpus. Moreover, BILBOWA outperforms other approaches with respect to recall for different ranges of cosine similarity threshold. In comparison to other approaches, CL-ESA obtains the worst results among all algorithms, especially in the case of precision. In the graph that

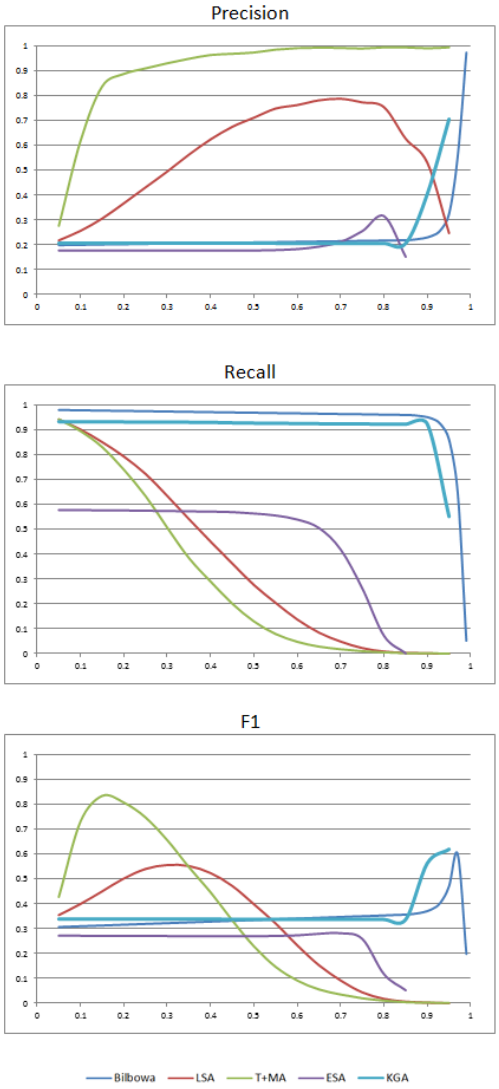

Fig. 5 Performance of algorithms on the HAMTA-CL corpus in terms of Cosine similarity between sentence pairs

represents F1, T+MA performs well when the threshold is less than 0.15. After this threshold, the performance of T+MA decreases monotonically. There is a similar trend in CL-LSI except that the best performance achieved with threshold of about 0.35. On the contrary, the behavior of BILBOWA and CL-ESA remains constant in most of the ranges of similarity thresholds. The best performance for BILBOWA is achieved with threshold of about 0.98 which obtaining on F1 measure of 0.61.

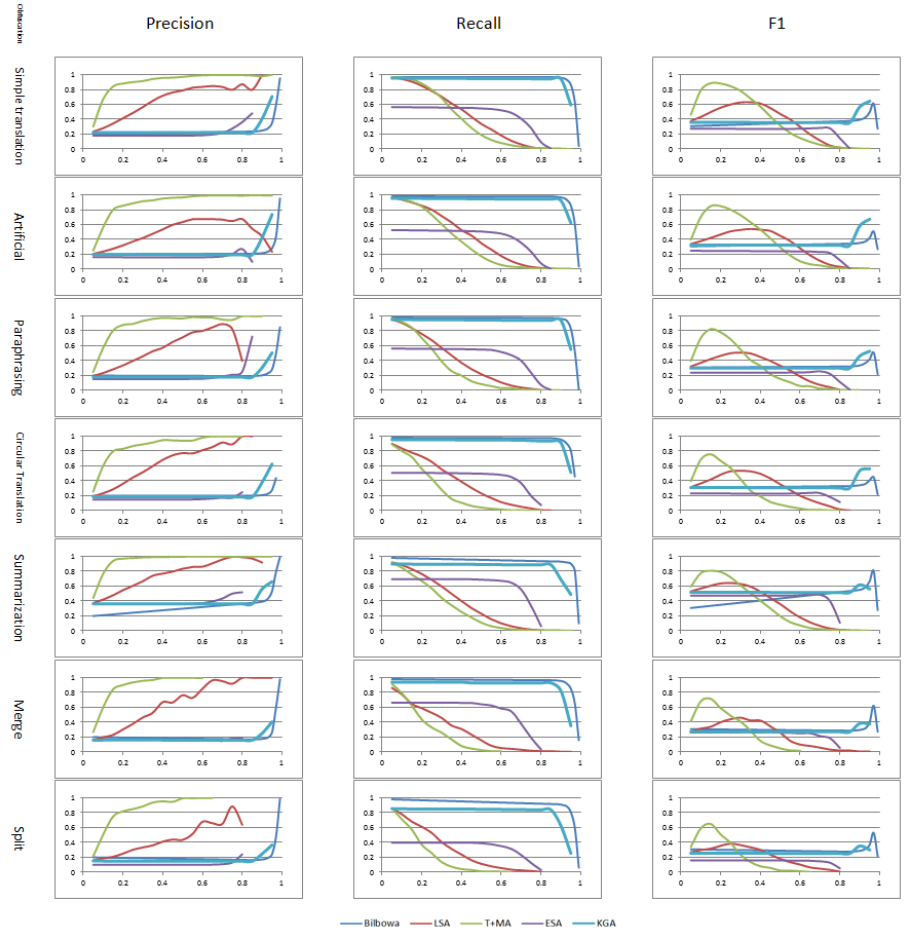

Fig. 6 Performance measure for different types of obfuscation

4.2 Experiment 2

In the second experiment, the performance of the CLPD methods is computed against separate sub-corpora containing different types of paraphrasing. Our purpose is to evaluate the capability of the methods on detecting different types of obfuscation.

The performance of the methods on detecting cases of plagiarism in different types of paraphrasing is presented in Figure 6. As shown in the figure, the T+MA method outperforms other approaches in F1 and precision. Also, BILBOWA achieved the best results in the case of recall measure for different types of obfuscation. Although T+MA performs better than other methods, however BILBOWA shows a close performance to T+MA for more complicated types of paraphrasing (e.g Merge and Split) and outperforms T+MA in detecting summarization cases of obfuscation. Table 5 shows Plagdet and granularity obtained by each method. We can discuss the performance of methods on the sub-corpora from

three perspectives as follow.

First, we focus on the performance of algorithms over sub-corpora. Since T+MA is a precision-oriented method, its best performance is when the obfuscation complexity is low. For instance, the performance of T+MA on STR sub-corpus is highest with respect to the other sub-corpora. Since BILBOWA is a semantic approach, so it is a recall oriented method. It can be seen when the obfuscation complexity is high (such as summarization), it outperforms the other algorithms. In LSI algorithm, the precision value increases monotonically, but when the threshold value reaches around 0.7, the precision of LSI decreases. Since the recall is very low at the values above this threshold, the change in F1 is not perceptible. ESA is a recall oriented semantic approach and it can be seen that its precision is very low. The precision doesn't leverage with the increase in threshold value. This method has also the lowest recall among the other algorithms which decreases rapidly in threshold value around 0.7. Since ESA is inherently a semantic approach, it works better in fragments with complex obfuscation.

Table 5 Comparison of performance measures in each method vs. different types of obfuscation

Obfuscation	**CL-ESA**				**CL-LSI**				**BILBOWA**				**T+MA**				**CL-KGA**			
	Recall	Precision	Granularity	Plagdet	Recall	Precision	Granularity	Plagdet	Recall	Precision	Granularity	Plagdet	Recall	Precision	Granularity	Plagdet	Recall	Precision	Granularity	Plagdet
Whole Corpus	.41	.21	**1**	.28	.63	.49	**1**	.55	.64	.55	**1**	.59	**.83**	**.83**	**1**	**.83**	.55	.70	**1**	.61
STR	.40	.21	**1**	.28	.61	.64	**1**	.62	.62	.59	**1**	.60	**.93**	**.84**	**1**	**.88**	.59	.70	**1**	.64
ART	.52	.16	**1**	.24	.62	.47	**1**	.54	.60	.43	**1**	.50	**.90**	**.80**	**1**	**.84**	.62	.73	**1**	.67
PAR	.39	.18	**1**	.25	.57	.45	**1**	.50	.52	.49	1.001	.50	**.84**	**.80**	**1**	**.82**	.55	.50	**1**	.52
SUM	.64	.39	**1.004**	.48	.68	.60	1.07	.61	**.80**	.79	1.008	**.79**	.71	**.90**	1.03	.78	.68	.56	1.03	.60
CTR	.45	.16	**1**	.23	.55	.51	**1**	.53	.45	.43	**1**	.44	**.71**	**.79**	**1**	**.75**	.51	.62	**1**	.56
SPL	.39	.09	**1.001**	.15	.52	.30	1.01	.38	**.59**	.47	1.02	.52	.57	**.75**	1.05	**.62**	.62	.24	1.04	.34
MRG	.66	.16	1.01	.26	.45	.46	**1.002**	.46	**.65**	.59	1.066	.59	.62	**.84**	1.004	**.71**	.82	.25	1.09	.36

Table 6 Changes in the performance of algorithms with respect to the whole corpus

Obfuscation	CL-ESA	CL-LSI	BILBOWA	T+MA	CL-KGA
Simple translation (STR)	0	+7.4	+1	+5.1	+2.5
Artificial (ART)	-3.6	-1.4	-9.2	+1.5	+5.5
Paraphrasing (PAR)	-2.5	-4.7	-9.1	-0.9	-9.1
Summarization (SUM)	+20.7	+5.6	+20.0	-5.1	-1.3
Circular Translation (CIR)	-4.4	-2.1	-15.1	-7.8	-5.6
Split (SPL)	-12.3	-17.5	-7.6	-20.8	-27.6
Merge (MRG)	-1.9	-9.5	-0.3	-12.2	-25.8

Second, we focus on the effect of obfuscation complexity on the performance of the above mentioned methods. From a general point of view, it is expected that the performance of the methods decreases when the obfuscation complexity increases. In simple translation (STR), since there is no obfuscation in the fragments, all of the methods have their best results. In the artificial obfuscation (ART), the fragments are constructed automatically, while the fragments in paraphrase obfuscation (PAR) have been manually changed by human. As a result, the PAR fragments have more complexity with respect to the ART fragments. As shown in Figure 6, the algorithms have better performances in artificial obfuscation with respect to the paraphrase obfuscation. Merge (MRG) and Split (SPL) obfuscations cause the structure of sentences to be messed up, whereas all of the methods work on a sequence of individual sentences to detect cases of plagiarism. Therefore, the worst performance of the methods occurs with MRG and SPL types of obfuscations. Moreover, the performance of all of the methods in SPL obfuscation is lower than MRG. As a last point, it seems that the most complex obfuscation is summarization (SUM), but since the summarized passages are relatively long (with respect to merge and split passages), the performance of the methods on SUM is better than MRG and SPL.

Third, we consider the sensitivity of methods on each sub-corpus with respect to the whole HAMTA-CL corpus as shown in Table 6. It can be seen in the table that the performance of BILBOWA on Merge (MRG) and Split (SPL) obfuscation has the lowest change with respect to the whole corpus among the other methods. In other words, MRG and SPL obfuscation decreases the performance of all algorithms except BILBOWA. On the other hand, BILBOWA has the most change in the performance among all of the approaches in the case of Artificial (ART) and Paraphrase (PAR) obfuscation. Another issue is that all of the methods have the lowest change in performance in simple translation (STR), while all of them have the highest growth of performance when facing summarization (SUM) obfuscation.

5 Conclusion and future works

In this paper we presented a novel approach to cross language plagiarism detection using word embedding methods and explore its performance against other state-of-the-art plagiarism detection algorithms. Moreover, we investigated various algorithms on the task of cross language plagiarism detection. We categorized the methods, described their pros and cons and compared them in the task of CLPD, focusing on English and Persian as two distant languages. We also investi-

gated the performance of CLPD approaches applied to Persian as a low resource language.

For investigating the performance of the algorithms, a corpus comprised of seven different types of obfuscation was constructed. The simulated cases of plagiarism were compiled by graduated crowd workers, while the artificial ones were compiled automatically. For validation of the corpus, it was automatically checked considering the ratio of length of plagiarized passages to length of the documents and the distribution of plagiarized passages across the documents. Moreover, for evaluation of the corpus, a manual checking was done for investigating the quality of plagiarized fragments. We compared the performance of the algorithms on the whole corpus and also on separate sub-corpora containing different types of paraphrasing as well.

For comparing the methods on CLPD, we implemented five algorithms and evaluated them using the constructed corpus. The performance of the algorithms on detecting cases of plagiarism in different types of paraphrasing showed that T+MA method outperforms other approaches in F1 and precision. Also, BILBOWA achieved the best results in the case of recall for different types of obfuscation. The results can also show that BILBOWA can detect more complicated types of plagiarism (Merge, Split and Summarization).

As a future work, we plan to focus our research on improving the performance of the above mentioned algorithms concerning Persian specific features. Another research that can be investigated in the future is to work on bilingual plagiarism detection when the source and target languages are both less resourced.

References

Asghari, H., Khoshnava, K., Fatemi, O., & Faili, H. (2015). Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus: Notebook for {PAN} at {CLEF} 2015. In L. Cappellato, N. Ferro, G. J. F. Jones, & E. SanJuan (Eds.), *Working notes of {clef} 2015 - conference and labs of the evaluation forum, toulouse, france, september 8-11, 2015.* (Vol. 1391). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1391/148-CR.pdf

Asghari, H., Mohtaj, S., Fatemi, O., Faili, H., Rosso, P., & Potthast, M. (2016). Algorithms and Corpora for Persian Plagiarism Detection: Overview of {PAN} at {FIRE} 2016. In P. Majumder, M. Mitra, P. Mehta, J. Sankhavara, & K. Ghosh (Eds.), *Working notes of {fire} 2016 - forum for information retrieval evaluation, kolkata, india, december 7-10, 2016.* (Vol. 1737, pp. 135–144). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1737/T4-1.pdf

Bannard, C. J., & Callison-Burch, C. (2005). Paraphrasing with Bilingual Parallel Corpora. In K. Knight, H. T. Ng, & K. Oflazer (Eds.), {*ACL*} *2005, 43rd annual meeting of the association for computational linguistics, proceedings of the conference, 25-30 june 2005, university of michigan,* {*usa*} (pp. 597–604). The Association for Computer Linguistics. Retrieved from http://aclweb.org/anthology/P/P05/P05-1074.pdf

Barrón-Cedeño, A., Potthast, M., Rosso, P., & Stein, B. (2010). Corpus and Evaluation Measures for Automatic Plagiarism Detection. In

N. Calzolari et al. (Eds.), *Proceedings of the international conference on language resources and evaluation, {lrec} 2010, 17-23 may 2010, valletta, malta.* European Language Resources Association. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/summaries/35.html

Barrón-Cedeño, A., Rosso, P., Agirre, E., & Labaka, G. (2010). Plagiarism Detection across Distant Language Pairs. In C.-R. Huang & D. Jurafsky (Eds.), *{COLING} 2010, 23rd international conference on computational linguistics, proceedings of the conference, 23-27 august 2010, beijing, china* (pp. 37–45). Tsinghua University Press. Retrieved from http://aclweb.org/anthology/C10-1005

Barrón-Cedeño, A., Rosso, P., Devi, S. L., Clough, P. D., & Stevenson, M. (2011). PAN@FIRE: Overview of the Cross-Language !ndian Text Re-Use Detection Competition. In P. Majumder, M. Mitra, P. Bhattacharyya, L. V. Subramaniam, D. Contractor, & P. Rosso (Eds.), *Multilingual information access in south asian languages - second international workshop, {fire} 2010, gandhinagar, india, february 19-21, 2010 and third international workshop, {fire} 2011, bombay, india, december 2-4, 2011, revised selected papers* (Vol. 7536, pp. 59–70). Springer.

Barrón-Cedeño, A., Rosso, P., Pinto, D., & Juan, A. (2008). On Cross-lingual Plagiarism Analysis using a Statistical Model. In B. Stein, E. Stamatatos, & M. Koppel (Eds.), *Proceedings of the ecai'08 workshop on uncovering plagiarism, authorship and social software misuse, patras, greece, july 22, 2008* (Vol. 377). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-377/paper1.pdf

Bosma, W., & Callison-Burch, C. (2006). Paraphrase Substitution for Recognizing Textual Entailment. In C. Peters et al. (Eds.), *Evaluation of multilingual and multi-modal information retrieval, 7th workshop of the cross-language evaluation forum, {clef} 2006, alicante, spain, september 20-22, 2006, revised selected papers* (Vol. 4730, pp. 502–509). Springer.

Ceska, Z., Toman, M., & Jezek, K. (2008). Multilingual Plagiarism Detection. In D. Dochev, M. Pistore, & P. Traverso (Eds.), *Artificial intelligence: Methodology, systems, and applications, 13th international conference, {aimsa} 2008, varna, bulgaria, september 4-6, 2008. proceedings* (Vol. 5253, pp. 83–92). Springer.

Farghaly, A., & Shaalan, K. F. (2009). Arabic Natural Language Processing: Challenges and Solutions. *{ACM} Trans. Asian Lang. Inf. Process.*, *8*(4), 14:1–14:22.

Ferrero, J., Agnès, F., Besacier, L., & Schwab, D. (2016). A Multilingual, Multi-style and Multi-granularity Dataset for Cross-language Textual Similarity Detection. In N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation {lrec} 2016, portorož, slovenia, may 23-28, 2016.* European Language Resources Association {(ELRA)}. Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/summaries/304.html

Ferrero, J., Besacier, L., Schwab, D., & Agnès, F. (2017). CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity. In S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. M. Cer, & D. Jurgens (Eds.), *Proceedings of the 11th international*

workshop on semantic evaluation, semeval@acl 2017, vancouver, canada, august 3-4, 2017 (pp. 109–114). Association for Computational Linguistics.

Franco-Salvador, M., Gupta, P., & Rosso, P. (2013). Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing. In N. Ferro (Ed.), *Bridging between information retrieval and databases - {promise} winter school 2013, bressanone, italy, february 4-8, 2013. revised tutorial lectures* (Vol. 8173, pp. 227–236). Springer.

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.

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.

Gouws, S., Bengio, Y., & Corrado, G. (2015). BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In F. R. Bach & D. M. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning, {icml} 2015, lille, france, 6-11 july 2015* (Vol. 37, pp. 748–756). JMLR.org. Retrieved from `http://jmlr.org/proceedings/papers/v37/gouws15.html`

Gupta, P., Barrón-Cedeño, A., & Rosso, P. (2012). Cross-Language High Similarity Search Using a Conceptual Thesaurus. In T. Catarci, P. Forner, D. Hiemstra, A. Peñas, & G. Santucci (Eds.), *Information access evaluation. multilinguality, multimodality, and visual analytics - third international conference of the {clef} initiative, {clef} 2012, rome, italy, september 17-20, 2012. proceedings* (Vol. 7488, pp. 67–75). Springer.

Gupta, P., & Singhal, K. (2011). Mapping Hindi-English Text Re-use Document Pairs. In P. Majumder, M. Mitra, P. Bhattacharyya, L. V. Subramaniam, D. Contractor, & P. Rosso (Eds.), *Multilingual information access in south asian languages - second international workshop, {fire} 2010, gandhinagar, india, february 19-21, 2010 and third international workshop, {fire} 2011, bombay, india, december 2-4, 2011, revised selected papers* (Vol. 7536, pp. 79–85). Springer.

Kent, C. K., & Salim, N. (2009). Web Based Cross Language Plagiarism Detection. *CoRR*, *abs/0912.3*. Retrieved from `http://arxiv.org/abs/0912.3959`

McNamee, P., & Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. *Inf. Retr.*, *7*(1-2), 73–97.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR*, *abs/1301.3*. Retrieved from `http://arxiv.org/abs/1301.3781`

Navigli, R., & Ponzetto, S. P. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In J. Hajic, S. Carberry, & S. Clark (Eds.), {*ACL*} *2010, proceedings of the 48th annual meeting of the association for computational linguistics, july 11-16, 2010, uppsala, sweden* (pp. 216–225). The Association for Computer Linguistics. Retrieved from `http://www.aclweb.org/anthology/P10-1023`

Nawab, R. M. A., Stevenson, M., & Clough, P. D. (2010). University of Sheffield - Lab Report for {PAN} at {CLEF} 2010. In M. Braschler, D. Harman, & E. Pianta (Eds.), {*CLEF*} *2010 labs and workshops, notebook papers, 22-23 september 2010, padua, italy* (Vol. 1176). CEUR-WS.org. Retrieved from

http://ceur-ws.org/Vol-1176/CLEF2010wn-PAN-NawabEt2010.pdf

Oberreuter, G., L'Huillier, G., Rios, S. A., & Velásquez, J. D. (2011). Approaches for Intrinsic and External Plagiarism Detection - Notebook for {PAN} at {CLEF} 2011. In V. Petras, P. Forner, & P. D. Clough (Eds.), {*CLEF*} *2011 labs and workshop, notebook papers, 19-22 september 2011, amsterdam, the netherlands* (Vol. 1177). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-OberreuterEt2011.pdf

Pereira, R. C., Moreira, V. P., & Galante, R. (2010). A New Approach for Cross-Language Plagiarism Analysis. In M. Agosti, N. Ferro, C. Peters, M. de Rijke, & A. F. Smeaton (Eds.), *Multilingual and multimodal information access evaluation, international conference of the cross-language evaluation forum, {clef} 2010, padua, italy, september 20-23, 2010. proceedings* (Vol. 6360, pp. 15–26). Springer.

Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., & Rosso, P. (2009). A statistical approach to crosslingual natural language tasks. *J. Algorithms*, *64*(1), 51–60.

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., & Rosso, P. (2010). Overview of the 2nd International Competition on Plagiarism Detection. In M. Braschler, D. Harman, & E. Pianta (Eds.), {*CLEF*} *2010 labs and workshops, notebook papers, 22-23 september 2010, padua, italy* (Vol. 1176). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1176/CLEF2010wn-PAN-PotthastEt2010a.pdf

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, *45*(1), 45–62.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. In V. Petras, P. Forner, & P. D. Clough (Eds.), {*CLEF*} *2011 labs and workshop, notebook papers, 19-22 september 2011, amsterdam, the netherlands* (Vol. 1177). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-PotthastEt2011a.pdf

Potthast, M., Goering, S., Rosso, P., & Stein, B. (2015). Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. In L. Cappellato, N. Ferro, G. J. F. Jones, & E. SanJuan (Eds.), *Working notes of {clef} 2015 - conference and labs of the evaluation forum, toulouse, france, september 8-11, 2015.* (Vol. 1391). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1391/inv-pap11-CR.pdf

Potthast, M., Stein, B., & Anderka, M. (2008). A Wikipedia-Based Multilingual Retrieval Model. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. W. White (Eds.), *Advances in information retrieval , 30th european conference on {ir} research, {ecir} 2008, glasgow, uk, march 30-april 3, 2008. proceedings* (Vol. 4956, pp. 522–530). Springer.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An Evaluation Framework for Plagiarism Detection. In C.-R. Huang & D. Jurafsky (Eds.), {*COLING*} *2010, 23rd international conference on computational linguistics, posters volume, 23-27 august 2010, beijing, china* (pp. 997–1005). Chinese Information Processing Society of China. Retrieved from http://aclweb.org/anthology/C/C10/C10-2115.pdf

Pouliquen, B., Steinberger, R., & Ignat, C. (2006). Automatic Identification of Document Translations in Large Multilingual Document Collections. *CoRR*,

abs/cs/060. Retrieved from http://arxiv.org/abs/cs/0609060

Rehder, B., Littman, M. L., Dumais, S. T., & Landauer, T. K. (1997). Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing. In E. M. Voorhees & D. K. Harman (Eds.), *Proceedings of the sixth text retrieval conference, {trec} 1997, gaithersburg, maryland, usa, november 19-21, 1997* (Vol. Special Pu, pp. 233–239). National Institute of Standards and Technology {(NIST)}. Retrieved from http://trec.nist.gov/pubs/trec6/papers/lsi.ps

Stein, B., Stamatatos, E., & Koppel, M. (2008). Proceedings of the ECAI'08 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, Patras, Greece, July 22, 2008. In (Vol. 377). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-377

Stein, B., zu Eissen, S. M., & Potthast, M. (2007). Strategies for retrieving plagiarized documents. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, & N. Kando (Eds.), {*SIGIR*} *2007: Proceedings of the 30th annual international {acm} {sigir} conference on research and development in information retrieval, amsterdam, the netherlands, july 23-27, 2007* (pp. 825–826). ACM.

Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2015). Towards Universal Paraphrastic Sentence Embeddings. *CoRR*, *abs/1511.0*. Retrieved from http://arxiv.org/abs/1511.08198

Zarrabi, V., Rafiei, J., Khoshnava, K., Asghari, H., & Mohtaj, S. (2015). Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection. In L. Cappellato, N. Ferro, G. J. F. Jones, & E. SanJuan (Eds.), *Working notes of {clef} 2015 - conference and labs of the evaluation forum, toulouse, france, september 8-11, 2015.* (Vol. 1391). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1391/147-CR.pdf