Journal of AI and Data Mining
Vol 5, No 2, 2017, 275-284

English-Persian Plagiarism Detection based on a Semantic Approach

F. Safi-Esfahani*, Sh. Rakian and M.-H. Nadimi-Shahraki

Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Isfahan, Iran.

Received 02 February 2016; Revised 15 March 2016; Accepted 17 September 2016
**Corresponding author: fsafi@iaun.ac.ir (F. Safi-Esfahani).*

Abstract
Plagiarism, defined as "the wrongful appropriation of other writers' or authors' works and ideas without citing or informing them", poses a major challenge to knowledge spread publication. Plagiarism has been placed in the four categories of direct, paraphrasing (re-writing), translation, and combinatory. This paper addresses the translational plagiarism, which is sometimes referred to as the cross-lingual plagiarism. In cross-lingual translation, writers meld a translation with their own words and ideas. Based on the monolingual plagiarism detection methods, this paper ultimately intends to find a way to detect the cross-lingual plagiarism. A framework called multi-lingual plagiarism detection (MLPD) has been presented for the cross-lingual plagiarism analysis with the ultimate objective of detection of plagiarism cases. English is the reference language, and Persian materials are back-translated using the translation tools. The data used for MLPD assessment is obtained from English-Persian Mizan parallel corpus. Apache's Solr is also applied to record the creep of the documents and their indexation. The accuracy mean of the proposed method was revealed to be 98.82% when employing highly accurate translation tools, which indicate the high accuracy of the method. Also the Google translation service showed the accuracy mean to be 56.9٪. These tests demonstrate that the improved translation tools enhance the accuracy of the developed method.

Keywords: *Text Retrieval, Cross-lingual, Text Similarity, Translation, Plagiarism, Semantic-based Plagiarism Detection.*

1. Introduction

An easier access to digital information, particularly the internet, has exponentially increased the plagiarism cases. Plagiarism comes in different forms including direct copying of a text without giving credit to the original writer, misappropriation of other's ideas, resources and styles, translation, reproduction of the original works via different visual and audio media, and code plagiarism [1]. Ceska et al. (2008) [2] have produced a new taxonomy of plagiarism, which has been completed later that year by Alzahrani et al. [3], who categorized plagiarism into literal and intelligent based on the plagiarist's behavioral viewpoint, highlighting the differences between these two phenomena.

Figure 1 presents a simple categorization of plagiarism. While the software tools are the most effective ones when they come to detect plagiarism, the final decision should be made based on the manual handling of cases [4].

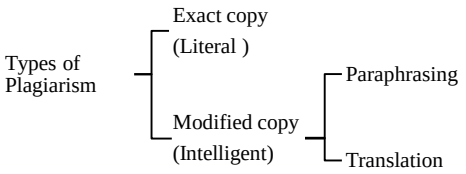

Figure 1: Taxonomy of plagiarism [6].

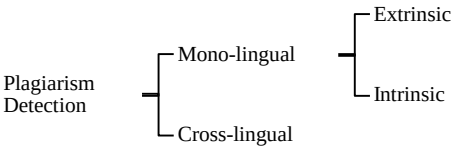

Figure 2. Plagiarism detection techniques.

Plagiarism detection could be categorized into monolingual and cross-lingual based on the varying degrees of homogeneity/heterogeneity of the language of documents (Figure 2).

Cross-lingual plagiarism refers to cases in which the writer melds a translation into his/her work

without giving a proper reference to the original text. Cross-lingual plagiarism refers to the detection and identification of plagiarism in a multilingual environment. It should be noted that detection of translational plagiarism is more challenging than the other categories of plagiarism [3]. In this category, the retrieval of suspicious documents from a large corpus of multilingual documents is intended.

This paper puts forward a machine translation-based detection method in order to identify cases of cross-lingual plagiarism of documents based on the application of a semantic relatedness approach. Semantic similarity or semantic relatedness rates the likeness of words using (WUP) Wu-Palmer for detection of plagiarism. The detection of the cross-lingual plagiarism cases is generally the same as the external ways of detection, yet with some minor changes. Persian fuzzy plagiarism detection (PFPD) [5], which has been primarily designed to detect cases of paraphrasing monolingual texts, does not cover translational plagiarism. This paper attempts to add the capability of detection of translational plagiarism to the store of PFPD. As a matter of fact, we intend to cover translational plagiarism through the development of the PFPD technique to improve its precision. It is hypothesized that through integration of semantic approach with PFPD to measure the similarity of inter-lingual text, the precision of detection of translational plagiarism will improve. It should be noted that this paper does not apply to the detection of inherent plagiarism (changes in stylistics). Suspicious texts (target text) are in Persian, and the source language is English. Detection of cross-lingual plagiarism refers to cases of automatic detection of plagiarism between languages. The purpose of this work is the translational plagiarism detection between Persian and English texts.

Apache's Solr was applied to record the creep of the documents and their indexation. The accuracy mean of the proposed method was revealed to be 98.82%, while employing the highly accurate translation tools, which indicate the high accuracy of the proposed method. The Google translation service was also employed to translate suspicious documents from Persian to English in order to implement the proposed method, and showed 56.9% for the accuracy mean.

The rest of this paper is organized as follows. The second section will offer some explanations and concepts about the cross-lingual plagiarism. The third section reviews a number of relevant research works with accounts of their pros and cons. The fourth section presents the methodology in detail. In the fifth section, we will deal with the implementation and analysis results. Finally, in the last part, we will review the conclusions made.

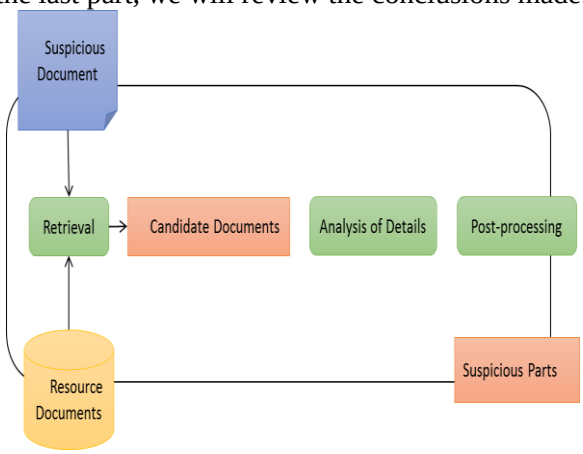

Figure 3. External plagiarism detection framework [6].

2. Concepts of cross-lingual plagiarism

Research on the detection of cross-lingual plagiarism has attracted many researches in the recent years, with a focus on the cross-lingual relatedness of the texts [2, 7-11]. The detection of the cross-lingual plagiarism cases is generally the same as the external ways of detection, yet with some minor changes. (The framework for the external detection methods is presented in Figure 3.) The differences are as follows:

(1) In the retrieval phase, it is required to write the suspicious texts in their source language.

(2) In the analysis of the details, the relatedness of the original and suspicious texts should be investigated. (It is also possible to back-translate the suspicious text, and then apply the monolingual plagiarism detection methods).

Query document and sets of documents serve as inputs in the operational framework of the cross-lingual plagiarism detection. Writing language, query document, and set of documents are not the same. Basically, there are three primary phases. In the first phase, a list of candidate documents is retrieved based on the CLIR models. If the suspicious document has been translated via a machine translation, it is possible to retrieve the candidate documents using IR models. In the second phase, a two-by-two study is conducted to find all suspicious parts of the query document that are similar to the candidate documents. In this phase, the language of the query and candidate texts is different. In the 3rd phase, post-processing is performed by a human agent to render the results obtained in a readable format [10].

The set of candidate texts is the distinguishing factor between the methods of cross-lingual

plagiarism detection and external detection. While in cross-lingual plagiarism detection, the language of the suspicious and the original texts is heterogeneous and thus the comparison is made between other languages and the suspicious text, in the external plagiarism detection, the language of the suspicious and original texts is the same. The syntactical and lexical features are not sufficient to create a cross-lingual environment. To establish the relation between a cross-lingual text and detection of plagiarism, the syntactical features are usually combined with the semantic or statistical features [3].

Known methods for cross-lingual information retrieval (CLIR) may be applied to retrieve candidate documents; 1. Extracting the key words of the suspicious text to obtain a set of words to represent it, translating these words and searching them in the original text, 2. Back-translating the suspicious text, extracting the key words, and searching them in the original text [12].

The output of this stage is a set of documents that might have been plagiarized. It should be noted that it is possible to use various cross-lingual information retrieval techniques such as comparable corpus, parallel corpus, multi-lingual dictionary, and machine translation [13].

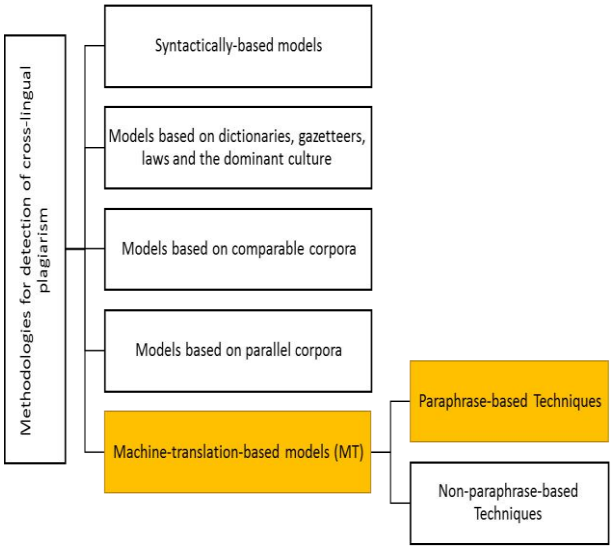

Figure 4. Taxonomy of different methodologies for detection of cross-lingual plagiarism.

3. Literature review

In the cross-lingual detail analysis phase, the similarity and relatedness of the suspicious documents and candidate documents are measured. For this, 5 models are available: 1. Syntactically-based models, 2. Models based on dictionaries, gazetteers, laws, and dominant culture, 3. Models based on comparable corpora, 4. Models based on parallel corpora, and 5. Machine-translation-based models [12]. Figure 4 shows the taxonomy of different methodologies for the detection of cross-lingual plagiarism.

Machine-translation-based models were applied in the proposed method. Other methods utilize the machine translation principles as well, yet they do not cover translating suspicious documents. Many methods use machine translation for analyzing documents to detect cross-lingual re-use of documents. This turns the issue into detecting monolingual plagiarism, which has gained popularity in the recent years [12]. The proposed method falls in this category.

So far, few quantitative research works have been conducted about cross-lingual plagiarism. However, the interest is growing fast in this regard. According to [12], there was no technology available before 2008 to detect the cross-lingual plagiarism cases. However, detection of cross-lingual plagiarism may benefit from the research works carried out in the other fields [2,8,9,11,14,28].

In [11], the proposed method is based upon statistical bilingual dictionary, which is comprised of parallel corpora and a bilingual algorithm text. The authors have conducted a test on a 5-piece set of plagiarized documents. The results obtained have revealed that the similarity between the original source and the plagiarized texts is remarkably higher than the unaffected documents.

The research work [2] has suggested MLPlag as a cross-lingual analysis tool for the detection of plagiarism based on the positions of words. This tool uses European Word net TM to convert words into an independent format from the language in question. The authors have created two multilingual corpora: 1. Fairy-tale, made up of 400 legal texts of EU, which were randomly selected and included 200 reports in English and a corresponding number in Czech, 2. JRC-EU, a set of textual documents in simple English and their 27 Czech English. This method has revealed good results. However, the authors have stated that an incomplete Word net may result in difficulties in the detection of plagiarism, especially while dealing with the less common languages.

A number of research works have been carried out in the field of retrieval of multilingual documents, which may contribute to the detection of cases of plagiarism [15]. A system was proposed to identify the original source of a translated text among a large number of candidate texts. The content was represented with a vector of words using a comprehensive dictionary. The textual similarity was measured independent from the language of the documents. The writers conducted

the test on a number of French and Spanish translated tests (from English) via a number of parallel corpora, which included 795 to 1130 pairs of texts and 1640 documents. The results obtained show that the system is capable of identifying the translation with an accuracy of 96%.

In [16], Kullback-Leibler have used divergence to reduce the number of documents to be compared with suspicious documents. In this study, a feature vector was created for each document of the reference set to be compared with the vector of the suspicious document. Ten documents with the lowest divergence with the vector of the suspicious documents were selected for the plagiarism analyses purposes.

Table 1. Comparison between Previous Research Works on Translational Plagiarism.

Reference/ Year/Authors	Method	Language	Main Features	Defects
[11] 2008 Barrón-Cedeno, A., et al.	Based on parallel corpora	Spanish-English	Based on a statistical dictionary created from parallel corpora and bilingual text document	Requires data for education
[2] 2008 Ceska, Z., M. Toman, and K. Jezek	Based on dictionaries, gazetteers, comprehensive laws	Czech-English	Analysis of word positions, European WordNet to convert words into formats independent of the target language	Incomplete WordNet may cause difficulties, especially while dealing with less common languages
[18] 2010 Kent, C.K. and N. Salim	MT- based	Malay-English	Analysis was based on three least-frequent 4-grams fingerprint matching	Fingerprint matching fails to identify cases where a word has been translated with its synonyms
[19] 2010 Muhr, M., et al.	Parallel corpora	European-English	Translation model output is used. Bekeley Aligner and EU corpora were employed to create a word-based model	Low accuracy
[17] 2011 Gupta, P., et al.	Vector Space Model (VSM)	German-Spanish & English	Analyze the monolingual paraphrases of English and cross-lingual paraphrases for German and Spanish languages	Low accuracy. This approach can be more effective by considering synonyms using thesauri, dictionary, and WordNet.
PFPD [5] 2015 Rakian, S., et al.	Fuzzy analysis in plagiarism detection and candidate documents retrieval	All Languages/ Tested for English/Persian Languages	Increasing precision and recall in candidate documents retrieval and in measuring the similarity. Avoid unnecessary comparisons	Paraphrasing detection monolingual texts
MLDP 2016 (Proposed Model)	MT-based	Persian-English	Promoting the fuzzy method to detect cases of rewriting of monolingual texts to help detect translational plagiarism cases	Extensive operations required

The research work [17] has focused on paraphrasing detection for both the monolingual and cross-lingual aspects applying Vector Space Model (VSM). The authors have considered English language for monolingual and German-Spanish languages for cross-lingual paraphrasing. This approach can be more effective by considering synonyms using thesauri, dictionary, and WordNet.

PFPD (Persian fuzzy plagiarism detection), presented in [5], is an approach to the external plagiarism detection in Persian texts. The aim of this framework is to make a compatible fuzzy method in Persian language. PFPD, which has been primarily designed to detect cases of paraphrasing monolingual texts, does not cover translational plagiarism either.

Normalizing language in the pre-processing phase is a common measure taken for the cross-lingual information retrieval techniques, in particular, the cross-lingual plagiarism detection. In [7], the authors have proposed using English as the source language for 2 reasons: 1. Most of the internet content is in English, 2. Non-English to English translation tools are easier to access. In the first phase, a language detector was employed to determine the language of the documents. If the language is shown to be other than English, it will be rendered into English. The second phase is dedicated to detection of monolingual cases of plagiarism. This method covers five phases: language normalization, retrieval of candidate documents, education of categorization, comprehensive analysis of plagiarism, and post-processing. This study has been used as an automatic translation tool to translate the texts into a single language. To detect the plagiarized texts form the unaffected ones, the researchers have turned to the categorization algorithm. In the retrieval phase, the documents that are suspected to have been the subject of plagiarism will be extracted. This is a crucial phase as it is not possible to search such a large set of documents.

The research work [18] has suggested applications of API translation and Google search options. In the first phase, a suspicious text in Malay has been translated into English as the source language. The text has been entered in Google as a query after removal of stop and stemming words. Exact analysis was performed during the retrieval phase of the set of candidate documents. This analysis was based upon three least-frequent 4-grams fingerprint matching. The failure of fingerprint matching to detect words that were translated with synonyms is the main drawback of the method.

The research work [19] has attempted to utilize a partial machine translation process to detect the cross-lingual plagiarism cases. End translation has been replaced with the output of the translation model. The word-based model in this research work has been produced by Berkeley Aligner using the European parallel corpora. Each token has been substituted with five candidate translations, and if it was not possible to translate, the token would be directly used. Table 1 compares different plagiarism methods.

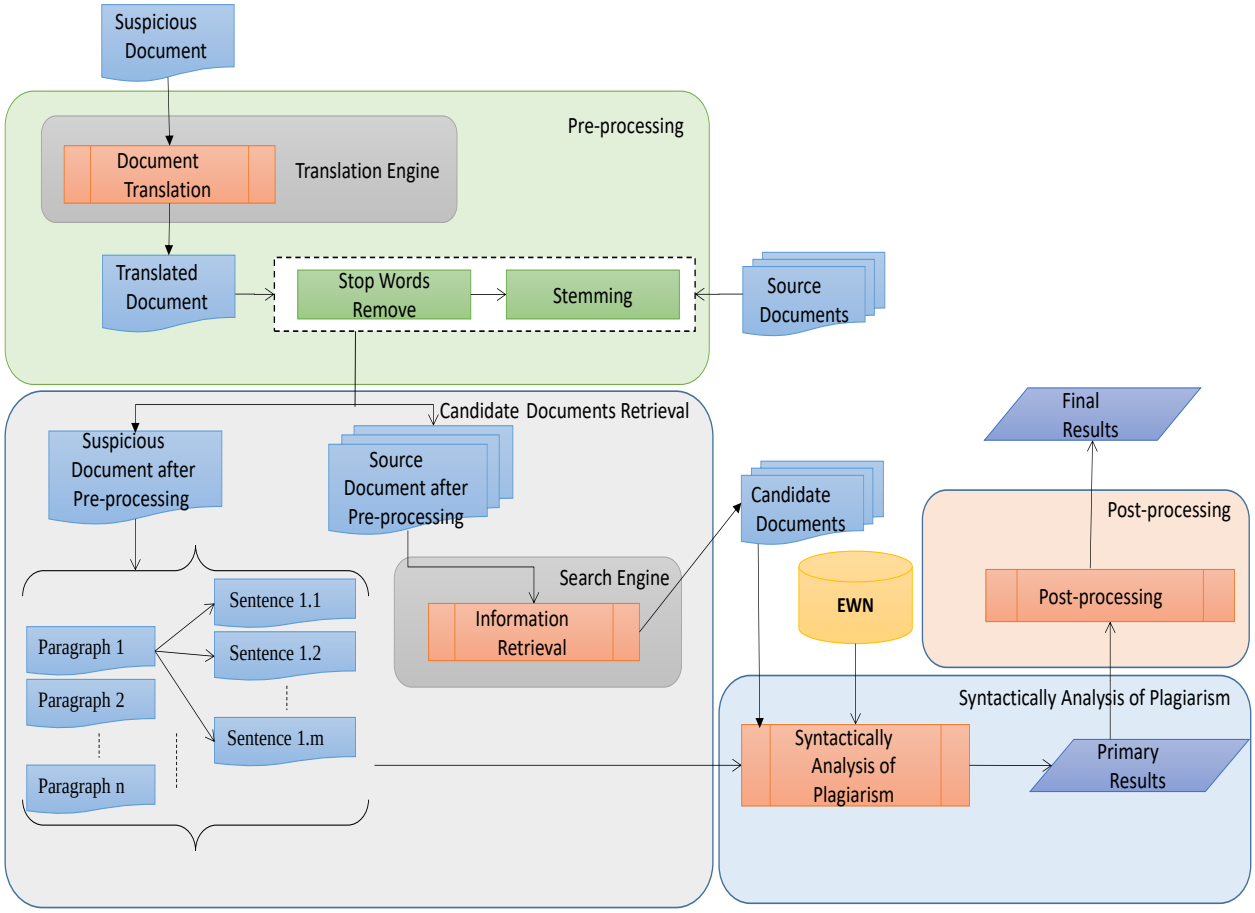

Figure 5. Recommended Framework- MLPD.

4. MLPD approach

This paper proposes MLPD (multi-lingual plagiarism detection) for a cross-lingual plagiarism analysis with the ultimate objective of detection of plagiarism cases. As the proposed model is designed to detect cross-lingual plagiarism, an automatic translation tool was employed to translate suspicious documents in English as the source language to make the analysis consistent. The proposed model intends to improve PFPD [5], a method used in detection of re-writing of monolingual texts that attempts to offer highly accurate detection results. It should be noted that methods of detection of monolingual plagiarism cannot be directly used for cases of cross-lingual plagiarism due as the words of suspicious texts and source texts will fail to match. Even if the plagiarized text is the exact translation of the source text, there will be at least changes in order of the words. As a result, MLPD tries to overcome this problem via employing an automatic translation tool to translate suspicious documents in English as the source language in order to make the analysis consistent. The primary difference between MLPD and PFPD, which is a fuzzy method for monolingual plagiarism detection, lies in the fact that in the former, occurrences of cross-lingual plagiarism are revealed in the analysis phase. English is set as the default language as the primary objective is to

detect cases of plagiarized cases of Persian texts translated from English, and the fact that most documents have been developed in English. The analysis will be performed after completion of the translations. The point is that even with employment of a superb translation tool, a partial content loss is inevitable. Figure 5 shoes how the proposed model works.

4.1. Text and word normalization

This phase includes translation of suspicious texts and removal of stop and stemming words. It is necessary to back-translate the suspicious texts before application of the plagiarism detection algorithms to detect potential cases of violation.

By stop words, we mean words that frequently occur in the texts without imparting any significant meaning [20]. Therefore, these words are removed for a faster analysis and shrinking the index store. The list of English stop words are available for processing.

Stemming removes suffixes and prefixes to produce word roots. Roots improve the information retrieval process. There are many stemmers for English language from which Porter and Kstem are very popular [21]. Here, we used Kstem for its higher precision.

```
    Input: candidateDocs[], susdoc
1   for each sentence i in susdocs
2   for each sentence j in docs
3   double similarity = compareSentences(sentence[i],
    sentence[j]);
4   if (similarity >=SIMTHRESHOLD)
5     result.add(sentence[j]);
6   Endfor
7   return result;
8   Endfor
    Output: result
```

Figure 6. Pseudo-code for measuring similarity of sentences.

4.2. Candidate documents retrieval

In this phase, a maximum of five documents with the highest frequency of suspicious sentences are selected from the source documents. This is done via the *Solr* search engine later.

4.3. Semantic analysis of plagiarism

In an attempt to detect monolingual plagiarism, Li et al. [22] have used the depth and length of the shortest route to the word in WorldNet synset. Similar to [22], our MLPD used the depth and length of the shortest route to the word in WordNet synset to detect monolingual plagiarism. WordNet is a lexical database for the English language that includes the lexical categories nouns, verbs, adjectives, and adverbs but ignores prepositions, determiners, and other function words. Sets of synonyms were linked to each other through semantic-conceptual and lexical ties. WordNet structure has become very useful in the process of natural language, thanks to its structure.

This paper differs from [22], as it uses a different procedure for measurement of relatedness level.

The pseudo-code for this algorithm is presented in Figure 6. In this code, suspicious and candidate documents are used as input.

For each sentence of the suspicious paragraph, first a matrix is formed. This matrix is measured for different words of each pair of sentences. If the bigger sentence is called s_1 and the shorter one s_2, then the matrix (1) is measured for them. In this matrix, the columns represent words from s_1 that do not occur in s_2, and the rows represent words from s_2 that do not occur in s_1. The internal volumes of this matrix are calculated with Wu-Palmer (WUP) semantic relation metric. This metric has been introduced by Wu and Palmer in 1994 [23]. If the output is bigger than one, number one is designated for them.

Subsequently, both α and β will be calculated based on the formula (2) and (3) (n is the number of the words of s_1, and *m* the number of the words of s_2). Then δ_{12} and δ_{21} are calculated as (4) and (5). In the final stage, the similarity of the two sentences is measured using the formula (6).

If the similarity rank was higher than T threshold, that sentence would be marked as plagiarized. Otherwise, it would be labeled as unaffected. Alzahrani and Salim [24] have set the suitable limit for T at 0.65. This paper follows the suit as well.

$$\begin{array}{c} \\ w_{21} \\ w_{22} \end{array} \begin{array}{c} \begin{array}{ccc} w_{11} & w_{12} & w_{13} \end{array} \\ \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \end{bmatrix} \end{array} \quad (1)$$

$$\alpha = \sum_{i=1}^{n} \max_{j=1}^{m} \left(w_{ij} \right) \quad (2)$$

$$\beta = \sum_{j=1}^{m} \max_{i=1}^{n} \left(w_{ij} \right) \quad (3)$$

$$\delta_{12} = (\alpha + Number\ of\ the\ same\ words) / \#s_1 \quad (4)$$

$$\delta_{21} = (\beta + Number\ of\ the\ same\ words) / \#s_2 \quad (5)$$

$$\gamma = \left| \delta_{12} - \delta_{21} \right| > 0.28 \text{ and } \max\left(\delta_{12}, \delta_{21} \right) < 0.83 \quad (6)$$

$$sim\left(s_1, s_2 \right) = \begin{cases} min\left(\delta_{12}, \delta_{21} \right), \gamma\ is\ TRUE \\ max\left(\delta_{12}, \delta_{21} \right), \gamma\ is\ FALSE \end{cases}$$

Multi-Lingual Plagiarism Detection (MLPD)

	W_{11} = various	W_{12} = carry	W_{13} = man	$\beta = \sum_{j=1}^{m} \max_{i=1}^{n}\left(w_{ij}\right) = 1.6$
W_{21} = person	0	0.31	0.8	0.8
W_{22} = case	0	0.57	0.8	0.8
$\alpha = \sum_{i=1}^{n} \max_{j=1}^{m}\left(w_{ij}\right) = 1.37$	0	0.57	0.8	

$$\delta_{12} = (\alpha + Number\ of\ the\ same\ words)/\#s_1 = (1.37+1)/4 = 0.59$$

$$\delta_{21} = (\beta + Number\ of\ the\ same\ words)/\#s_2 = (1.6+1)/3 = 0.86$$

$$(\gamma = |\delta_{12} - \delta_{21}| > 0.28)\ and\ (\max(\delta_{12}, \delta_{21}) < 0.83) = false$$

$$sim(s_1, s_2) = \begin{cases} min(\delta_{12}, \delta_{21}), \gamma\ is\ TRUE \\ max(\delta_{12}, \delta_{21}), \gamma\ is\ FALSE \end{cases}$$

$$sim_{MLPD}(s_1, s_2) = 0.86$$

Persian Fuzzy Plagiarism Detection (PFPD)

$$Same\ words\ (synonym):\ N / A$$

$$\delta = (0.5 \times \text{Number of the same words}) + (1 \times \text{ Number of the same words}) = 0.5 \times 0 + 1 \times 1 = 1$$

$$\alpha = \frac{\delta}{|s_1|} = \frac{1}{4} = 0.25$$

$$\beta = \frac{\delta}{|s_2|} = \frac{1}{3} = 0.33$$

$$\gamma = \frac{\delta}{(|S_1| + |S_2| - \delta)} = \frac{1}{(4+3)-1} = 0.17$$

$$Sim_{PFPD}(S_1, S_2) = \frac{A.min(\alpha,\beta) + B.max(\alpha,\beta) + C.\gamma}{A+B+C} = \frac{20 \times 0.25 + 8 \times 0.33 + 3 \times 0.17}{20+8+3} = 0.26$$

Figure 7. Case study calculations.

4.4. Post-processing

Once the results are produced, a summary of the results of plagiarism detection including plagiarized parts, source of plagiarism, and similarity percentage are presented. Post-processing is used to report the results and integrate the plagiarized parts.

5. Case study

To shed light on how the similarity measurement method of the MLPD model works and highlight its differences with PFPD [5], an example is given.

Example: Imagine that two sentences to be compared have been represented with S_1 and S_2. S_3 is the translation of S_2, obtained from the Google translation engine.

S_1: Sometimes one man carried various names.

The words in S_1 after equalization of the text, removal of stop words, and stemming are:

[name, various, carry, man]; then $|S_1| = 4$

S2: در برخی موارد شخصی دارای چندین نام است

S3: In some cases, a person has several names.

The words in S3 after equalization of the text, removal of stop words, and stemming are:

[case, person, name]; then $|S_3|=3$.

It should be noted that the word order is not important in measurement of similarity rate in [25], and the data structure used is a set. The same words in two sentences (shared points of S_1 and S_3): $\boldsymbol{S_1 \cap S_3 = [name]}$.

The calculation of the formulas is illustrated in Figure 7. As the results indicate, while MLPD can detect the similarity of two sentences, PFPD fails to do so.

6. Assessment and trials

The dataset for assessment of MLPD were obtained from the standard English-Persian Mizan parallel corpus. This corpus is free for all, and has been used in the research works [26] and [27]. It contains one million parallel sentences of Gutenberg's novels along with their Persian

translations that have been keyed in, spellchecked, and parallelized semi-automatically.
These trials attempt to prove the hypothesis of this research work and investigate the retrieval precision of the proposed model. A total of two sets of trial were conducted to test the hypothesis. First, each sentence is compared with its translation.
This reveals the accuracy of the proposed model when highly accurate translation tools are employed. In the second stage, the Google translation program translates the sentence, and then the results obtained are compared with the first phase.
Both trials were performed with MLPD and PFPD [5]. PFPD is a fuzzy method used in identification of re-writing. These trials were conducted to prove that the re-writing identification methods are not successful in detection of translational plagiarism.

6.1. Trial environment

In MLPD, first the Persian input is back-translated via the Google API free tool. In the retrieval phase, Apache's Solr was also applied to record the creep of the documents and their indexation. An HP Pavilion dv4-1515tx was used to complete the trials. Figure 8 gives a general scheme for the trial environment.

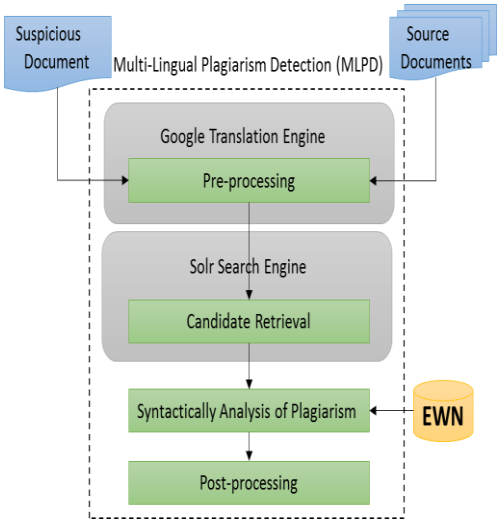

Figure 8. Experimental environment.

6.2. First set of trials

In the first set, each sentence is compared with its translation. This reveals the accuracy of the proposed model when highly accurate translation tools are employed. The trials were conducted on1021596 sentences. The results obtained are presented in Figure 9 and Table 2, respectively.

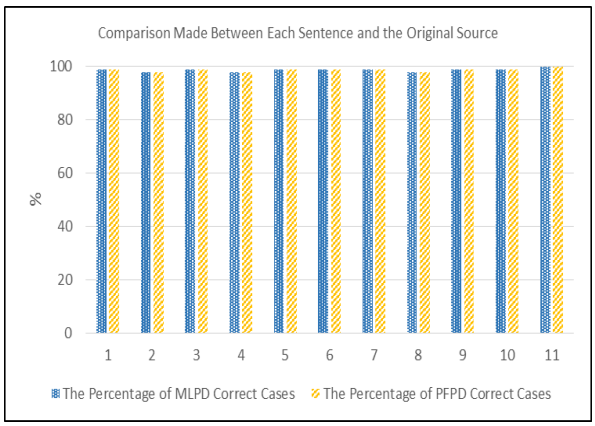

Figure 9. Percentage of results of comparison made between each sentence and original source.

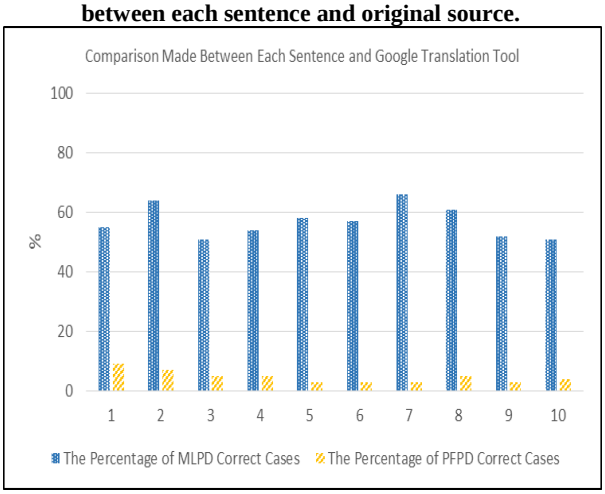

Figure 10. Percentage of results of comparison made between each sentence and google translation tool.

The results obtained reveal the accuracy mean of the proposed method to be 98.82% when employing highly accurate manually translated by an English expert, which indicate the high accuracy of the proposed method. This supports the idea that if suitable translation is used, the re-writing identification methods may contribute to the translational plagiarism detection.

6.3. Second set of trials

In the Second set, each sentence is translated by the Google translation program. The trials were conducted on 10000 sentences. The obtained results are presented in Figure 10 and Table 3. The results of the second trial indicate that when the Google translation program is employed, the accuracy mean is 56.9%, and the accuracy mean of PFPD is 4.7%. These tests demonstrate that the improved translation tools enhance the accuracy of the proposed method. Also after comparing the proposed method with PFPD, which is a method of monolingual plagiarism detection, it becomes obvious that the monolingual methods cannot be practiced for the cross-lingual plagiarism detection. This is because even if the plagiarized

text is an exact translation of the original text, the word order will not be the same. As a result, the words of suspicious and original texts will not match.

Table 2. Results of comparison made between each sentence and original source.

No.	From Sentence:	To:	Correct Cases of MLPD	Mistakes of MLPD	Percentage of MLPD Correct Cases	Correct Cases of PFPD	Mistakes of PFPD	Percentage of PFPD Correct Cases
1	1	100001	98095	1905	99	98095	1905	99
2	100001	200000	97484	2516	98	97484	2516	98
3	200001	300000	98295	1705	99	98295	1705	99
4	300001	400000	97946	2054	98	97946	2054	98
5	400001	500000	98059	1941	99	98059	1941	99
6	500001	600000	98484	1516	99	98484	1516	99
7	600001	700000	98650	1350	99	98650	1350	99
8	700001	800000	97342	2658	98	97342	2658	98
9	800001	900000	98290	1710	99	98290	1710	99
10	900001	1000000	98627	1373	99	98627	1373	99
11	1000001	1021596	21447	149	100	21447	149	100

Table 3. Results of comparison made between each sentence and google translation tool.

No.	From Sentence	To Sentence	Correct Cases of MLPD	Mistakes of MLPD	Percentage of MLPD Correct Cases	Correct Cases of PFPD	Mistakes of PFPD	Percentage of PFPD Correct Cases
1	1	1000	550	450	55	84	916	9
2	1001	2000	632	368	64	70	930	7
3	2001	3000	501	499	51	48	952	5
4	3001	4000	537	463	54	43	957	5
5	4001	5000	574	426	58	30	970	3
6	5001	6000	564	436	57	25	975	3
7	6001	7000	651	349	66	24	976	3
8	7001	8000	605	395	61	41	959	5
9	8001	9000	511	489	52	27	973	3
10	9001	10000	504	496	51	33	967	4

7. Conclusion and future works

In this paper, we attempted to propose a method for a cross-lingual plagiarism detection based on a semantic approach. The accuracy mean of the proposed method when employing highly accurate translation tools was compared with the Google translation program. These tests demonstrate that the improved translation tools enhance the accuracy of the proposed method. In the first set, each sentence is compared with its translation. This reveals the accuracy of the proposed model when highly accurate translation tools are employed. The results obtained reveal the accuracy mean of the proposed method to be 98.82% when employing highly accurate translation tools, which indicate the high accuracy of the proposed method. The results of the second trial indicate that when the Google translation program is employed, the accuracy mean is 56.9%. These trials revealed that improving the existing translation tools would enhance the accuracy of the proposed method. Also they showed that monolingual methods could not be practiced for the cross-lingual plagiarism detection.

In the proposed MLPD, the Google translation machine was employed to translate suspicious texts. Other translation machines could be used to draw a comparison with the results of MLPD. Also the results obtained could be compared with the other translational plagiarism methods.

References

[1] Lukashenko, R., Graudina, V. & Grundspenkis, J. (2007). Computer-based plagiarism detection methods and tools: an overview, In: Proceedings of the international conference on Computer systems and technologies, Bulgaria, 2007.

[2] Ceska, Z., Toman, M. & Jezek, K. (2008). Multilingual plagiarism detection, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence Lecture Notes in Bioinformatics), pp. 83-92.

[3] Alzahrani, S. M., Salim, N. & Abraham, A. (2012). Understanding Plagiarism linguistic patterns, textual features, and detection methods, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 42, no. 2, pp. 133-149.

[4] Gruner, S. & Naven, S. (2005). Tool support for plagiarism detection in text documents, In: Proceedings of the ACM symposium on Applied computing, Santa Fe, New Mexico, 2005.

[5] Rakian, S., Safi-Esfahani, F. & Rastegari, H. (2015). A Persian Fuzzy Plagiarism Detection Approach, Journal of Information Systems and Telecommunication (JIST), vol. 3, no. 11.

[6] Stein, B., Meyer zu Eissen, S. & Potthast, M. (2007). Strategies for retrieving plagiarized documents,

In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, NY, USA, 2007.

[7] Corezola Pereira, R., Moreira, V. & Galante, R. (2010). A New Approach for Cross-Language Plagiarism Analysis, Multilingual and Multimodal Information Access Evaluation, pp. 15-26.

[8] Lee, C. H., Wu, C. H. & Yang, H. C. (2008). A Platform Framework for Cross-lingual Text Relatedness Evaluation and Plagiarism Detection, In: Innovative Computing Information and Control, ICICIC'08, IEEE, 2008.

[9] Pinto, D., Civera, J., Barrón-Cedeòo, A., Juan, A., & Rosso, P. (2009). A statistical approach to crosslingual natural language tasks, Journal of Algorithms, vol. 64, no. 1, pp. 51-60.

[10] Potthast, M., Barrón-Cedeño, A., Stein, B. & Rosso, P. (2010). Cross-language plagiarism detection. Language Resources and Evaluation, vol. 45, no. 1, pp. 45-62.

[11] Barrón-Cedeno, A., Rosso, P., Pinto, D. & Juan, A. (2008). In: Proceeding of PAN'2008, Greece, 2008.

[12] Cedeno, L. A. B. (2012). On the mono-and cross-language detection of text re-use and plagiarism, Universidad de La Rioja, 2012.

[13] He, D. & J. Wang. (2009). Cross-language information retrieval, Information Retrieval: Searching in the 21st Century, pp. 233-254.

[14] Potthast, M., Stein, B. & Anderka, M. (2008). A Wikipedia-based multilingual retrieval model, In: Advances in Information Retrieval. Springer, pp. 522-530.

[15] Pouliquen, B., Steinberger, R. & Ignat, C. (2006). Automatic identification of document translations in large multilingual document collections, arXiv preprint cs/0609060.

[16] Barrón-Cedeño, A., P. Rosso, & Benedí, J.M. (2009). Reducing the plagiarism detection search space on the basis of the kullback-leibler distance, Computational Linguistics and Intelligent Text Processing, pp. 523-534.

[17] Gupta, P., Singhal, K., Majumder, P. & Rosso, P. (2011). Detection of Paraphrastic Cases of Mono-lingual and Cross-lingual Plagiarism, In: Proceedings of ICON, Chennai, India, 2011.

[18] Kent, C. K. & Salim, N. (2010). Web based cross language plagiarism detection, In: Computational Intelligence, Modelling and Simulation (CIMSiM), Second International Conference on, IEEE, 2008.

[19] Muhr, M., Kern, R., Zechner, M. & Granitzer , M. (2010). External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System, Braschler and Harman.

[20] Van Rijsbergen, C.J. (1986). A new theoretical framework for information retrieval, In: Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1986.

[21] Scherbinin, V. & S. Butakov. (2008). Plagiarism Detection: The Tool and The Case Study. In e-Learning, 2008.

[22] Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D. & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics, IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 8, pp. 1138-1150.

[23] Palmer, Z. W. A. M. (1994). Verbs semantics and lexical selection, In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133-138.

[24] Alzahrani, S. M. & Salim, N. (2010). Fuzzy semantic-based string similarity for extrinsic plagiarism detection: Lab report for PAN at CLEF'10, presented at the 4th International Workshop PAN'10, Padua, Italy, 2010.

[25] Meyer, D., Hornik, K. & Feinerer, I. (2008). Text mining infrastructure, Journal of Statistical Software, vol. 25, no. 5, pp. 1-54.

[26] Sarmad, M., IBM word-alignment model I for statistical machine translation.

[27] Azarbonyad, H. Shakery, A. & Faili, H. (2014). Learning to Exploit Different Translation Resources for Cross Language Information Retrieval, arXiv preprint arXiv:1405.5447.

[28] Rafieian, S. & Baraani-Dastjerdi, A. (2015). Plagiarism Checker for Persian (PCP) Texts Using Hash-based Tree Representative Fingerprinting, Journal of AI and Data Mining, vol. 4, no. 2, pp. 125-133.

نشریه هوش مصنوعی و داده کاوی

تشخیص سرقت زبانی فارسی-انگلیسی بر اساس یک روش معنایی

فرامرز صافی‌اصفهانی*، شیما راکیان و محمدحسین ندیمی‌شهرکی

دانشکده کامپیوتر، واحد نجف‌آباد، دانشگاه آزاد اسلامی، نجف‌آباد، اصفهان، ایران.

ارسال ۲۰۱۶/۰۲/۰۲؛ بازنگری ۲۰۱۶/۰۳/۱۵؛ پذیرش ۲۰۱۶/۰۹/۱۷

چکیده:

سرقت متون به استفاده از کارها و ایده‌های نویسندگان یا مولفین بدون ارجاع و بدون اطلاع آنها اطلاق می‌شود و یک مشکل جدی برای نشر دانش می‌باشد. سرقت متون به چهار گروه مستقیم، بازگویی (بازنویسی)، ترجمه و ترکیبی دسته‌بندی می‌گردد. این مقاله به گروه ترجمه می‌پردازد که به سرقت متون بین زبانی نیز معروف می‌باشد. در ترجمه بین زبانی نویسندگان متنی را با عبارات و ایده‌های خود ترجمه می‌کنند. این مقاله در تلاش برای یافتن راهی جهت تشخیص این نوع سرقت بر اساس تشخیص سرقت تک زبانی است و چارچوبی بنام MLPD برای آنالیز و تشخیص سرقت علمی بین زبانی ارائه می‌دهد. زبان انگلیسی، زبان مرجع می‌باشد و متون فارسی با استفاده از ابزارهای ترجمه به انگلیسی برگردانده می‌شوند. برای ارزیابی چارچوب MLPD از دادگان متون موازی المیزان که به زبان انگلیسی-فارسی می‌باشند، استفاده شده است. همچنین ابزار Apapche Solr برای خزش و ایندکس گزاری اسناد بکارگرفته شده‌است. هنگامی‌که از ابزارهای ترجمه دقیق استفاده شد، دقت متوسط به دست آمده ۹۸,۹۲٪ بدست آمد که نشان دهنده دقت بالای این روش می‌باشد. همچنین با استفاده از سرویس ترجمه گوگل متوسط دقت ۵۶,۹٪ حاصل شد. این آزمایش‌ها نشان می‌دهد که با بهبود ابزارهای ترجمه، دقت این روش پیشنهادی نیز بهبود می‌یابد.

کلمات کلیدی: بازیابی متن، بین-زبانی، شباهت متون، ترجمه، سرقت متن، تشخیص سرقت متن معنایی.