Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus

Notebook for PAN at CLEF 2015

Habibollah Asghari[1], Khadijeh Khoshnava[1], Omid Fatemi[2], Heshaam Faili[2]

[1] ICT Research Institute,Academic Center for Education, Culture and Reseach (ACECR), Iran
[2] Department of Electrical and Computer Engineering, University of Tehran, Iran

{habib.asghari, khadijeh.khoshnava}@ictrc.ir,
omid@fatemi.net, hfaili@ut.ac.ir

Abstract. Plagiarism detection is the process of locating text reuse within a suspicious document. The plagiarism detection corpora are used for evaluating plagiarism detection systems. In this paper, we present a bilingual Persian-English plagiarism detection corpus. We provide our corpus for the task of text alignment corpus construction in the PAN 2015 competition. Our approach is based on parallel corpus sentences. We have used a Persian-English sentence aligned parallel corpus in a combination with Wikipedia articles to create our corpus. Paired sentences in parallel corpus have a similarity score between 0 and 1. We have used similarity scores to establish the degree of obfuscation for constructing the plagiarism cases.

Keywords: Plagiarism Detection, Evaluation Corpus, Bilingual Corpus, Persian-English Corpus

1 Introduction

Plagiarism detection is the automatic identification of plagiarism and the retrieval of the original sources [1, 2]. The suspicious and source documents can be written either in the same language or in different languages. Particularly cross lingual plagiarism detection (CLPD) refers to cases where an author translates text from another language and then integrates the translated text into his/her own article [3].

The cross lingual plagiarism detection corpora are used to evaluate the cross lingual plagiarism detection systems. Since the creation of plagiarism corpora is very time demanding, so an alternative approach is to construct a corpus consisting of artificial plagiarized passages [4].

In this paper, we have proposed an approach to construct a bilingual Persian-English plagiarism detection corpus by using a Persian-English parallel corpus. The parallel corpus consists of aligned parallel sentences with similarity scores. Sentence

similarity scores have been used for establishing obfuscation degree in plagiarism cases. The paper is organized as follow: In section 2 we introduce the preparation of data sources needed to construct our corpus. In section 3 we will describe our approach in detail. Then, we will discuss the results of corpus building in section 4. Finally, we will conclude and explain about some future works in section 5.

2 Data Source preparation

We have used Wikipedia documents for constructing the main body of source and suspicious documents. Moreover, we exploited a parallel Persian- English sentence aligned corpus to construct the plagiarized passages. By inserting plagiarized passages with specific degrees of obfuscation into the document with related topics, a bilingual Persian–English plagiarism detection corpus was established. In the following subsections we provide a brief overview of these two resources.

2.1 Wikipedia

Wikipedia is a rich multilingual web-based encyclopedia. Each document in Wikipedia is represented as a page. The text of pages is partially structured [5]. We have crawled Persian Wikipedia documents in accordance with corresponding pages in English language. In the process of crawling, we have considered and extracted the following fields:

- Title of the page
- Url of the page
- Text of the page
- Categories field of the page

It should be noted that pages less than 300 words were removed from the collected data due to low information content.

2.2 Persian – English Parallel Corpus

We have exploited a parallel English-Persian sentence aligned corpus to construct paired plagiarism passages to be inserted into source (English) and suspicious (Persian) documents. A collection of 12 features were used into a Maximum Entropy (MaxEnt) log linear model in order to compute the similarity scores between paired sentences. The features are in four categories including: Features based on sentence length, Features related to dictionary (IBM model 1), Features based on alignment and, Miscellaneous features. The total score resulted from the mentioned features has been used to determine the various degrees of obfuscation in plagiarized passages; the more similar sentences can be used to build less obfuscated passages.

3 Our Approach

In this section we describe our approach to generate a bilingual Persian-English plagiarism detection corpus. We use a sentence aligned parallel corpus to create plagiarism cases. In the following, we explain our approach in five steps: preprocessing, clustering, building plagiarism cases, fragment obfuscation and inserting plagiarized cases into source and suspicious documents.

3.1 Preprocessing

Persian is one of the Indo-European languages which have borrowed its script from Arabic, a member of the Semitic language family [6]. In the process of developing a Persian corpus, we faced a lot of problems due to some special features of Persian language [7]. The control characters for Persian are very similar to Arabic, but with some differences. One discrepancy is that the written texts sometimes employ Arabic or ASCII characters beside the range of Unicode characters designed for Persian. While the Arabic and Persian codes coming together, processing through text is difficult. Another importance issues for Persian texts is the internal word boundary that should be presented with a zero-width non-joiner space named pseudo-space. Typically, typists completely ignore the internal word boundary or enter a white space instead of it. Moreover, optionality of the internal word boundary raises problems in processing of Persian texts [6].

Therefore, to overcome these problems and challenging issues, we have applied some algorithms such as normalization in the preprocessing stage of the system. Unification of letters to Unicode characters designed for Persian and using zero-width non-joiner space are applied in normalization algorithm.

3.2 Clustering

Our purpose is to establish topically similarity between suspicious documents, source documents and their plagiarism cases, so as to make plagiarism corpus to be more realistic and make plagiarism cases hard to find.

We have proposed our approach for clustered parallel sentences and Wikipedia documents into different topically related groups. Therefore, this step is organized in two subsections: parallel sentence clustering and documents clustering. In the following, we describe the process of each subsection.

Parallel Sentence Clustering. Given a collection of parallel sentences, the clustering procedure of parallel sentences is accomplished to detect the presence of distinct groups and assign parallel sentences to groups, such that the parallel sentences within a group are very similar and also parallel sentences in apart clusters are different from one another.

Since the parallel corpus we have used, has been extracted from Wikipedia, so we used the structure of the wiki pages for clustering of sentences. The algorithm for clustering of parallel sentences is as follow:

1. Persian Wikipedia documents were indexed by the Apache Lucene library.
2. A query was built from each Persian sentence.
3. The query was searched in the indexed documents and returns the top document.
4. A bipartite graph of return documents-categories was created. Then, the info- map community detection algorithm was applied to the graph and all communities were detected. Documents within a community are considered as one cluster.
5. Finally, parallel sentences were assigned to the documents in the same cluster.

Documents Clustering. For clustering of documents which includes source and suspicious documents, we used the results of parallel sentences clustering stage. For each cluster of return documents in the previous stage, the categories of documents have been extracted and considered as label of that cluster. Then, we collected basic documents into different topically related clusters based on their categories. The documents are assigned to the cluster with maximum common categories.

3.3 Building Plagiarism cases

In this step, we have used paired sentences from parallel corpus to create plagiarism cases. For constructing a plagiarism case, we put together some of the sentences of parallel corpus. Note that source fragments were generated from sentences in the English language and plagiarized fragments were constructed by Persian sentences paired with English sentences.

The length of fragments is evenly distributed between 3 and 15 sentences. The length of fragments is shown in table 1.

Table 1. Fragment lengths in words

Fragment Length	
Short	3 – 5 sentences
Medium	5 – 10 sentences
Long	10 – 15 sentences

3.4 Fragment Obfuscation

Plagiarism cases in bilingual corpus are constructed from parallel sentences. Plagiarized fragments have been constructed from Persian sentences and corresponding source fragments have been constructed from English sentences parallel with source sentences. To consider the degree of obfuscation in plagiarized fragments, a combination of sentences with different similarity score were chosen. The number of sentences and their similarity score in a fragment specifies the degree of obfuscation in that

fragment. Different degrees of obfuscation are “Low”, “Medium”, and “High” obfuscation which is shown in Table 2.

Table 2. Degree of obfuscation in plagiarism cases

Degree	Similarity scores of sentences in fragments		
	1- 0.85	**0.85 – 0.65**	**0.65 – 0.85**
Low	100%	-	-
Medium	55% - 75%	25% - 45%	-
High	35% - 55%	-	45% - 65%

3.5 Inserting Plagiarism Cases into Source and Suspicious Documents

In this step, according to the length of suspicious document, one or more plagiarism cases which are in the same cluster of suspicious document are selected. Then, each of them is inserted at random positions in suspicious document. Persian documents considering as suspicious documents and source documents are English documents. Source fragments also, inserted at random positions in source documents. In other words, Persian translation of English fragments has been inserted into suspicious documents.

The fraction of plagiarism in each document is not a fixed value. The percentage of plagiarism in each suspicious document is distributed between 5% and 60% of its length. The ratio of plagiarism per suspicious documents is shown in Table 3.

Finally, for each pair of source and suspicious documents, an XML file was generated which contains meta information about the plagiarism cases. The metadata XML file includes:

- this_length: Length of plagiarism case in the suspicious document.
- this_offset: Start offset of the plagiarism case in the suspicious document.
- source_reference: Name of source file.
- source_length: Length of source fragment in source document.
- source_offset: Start offset of the source fragment in the source document.

Table 3. Ratio of Plagiarism fragments in Documents

Plagiarism per Document	
Low	5% - 20%
Medium	20% - 40%
High	40% - 60%

4 Results

In this section, the statistics of our bilingual corpus are represented. An overview of important corpus statistics is shown in Table 4.

Table 4. Bilingual Persian-English Corpus statistics

Documents	
The number of source documents (English):	19973
The number of suspicious documents (Persian):	
With plagiarism:	3571
No plagiarism:	3571
Plagiarism cases	
The number of plagiarism cases:	11200
Plagiarism per Document	
The number of Little plagiarized documents:	2035
The number of Medium plagiarized documents:	536
The number of Much plagiarized documents:	642
The number of Very much plagiarized documents:	358

The established bilingual Persian-English plagiarism detection corpus is available at the website[1] of "Research Institute for Information and Communication Technology" for research purposes.

5 Conclusion and Future Works

In this paper we have described our approach to the task of text alignment corpus construction in the context of PAN 2015 competition. This corpus is intended to be used to evaluate the performance of bilingual plagiarism detection systems. We have exploited a sentence aligned parallel corpus to construct a bilingual Persian–English plagiarism detection corpus. Our main contribution is to use a novel obfuscation strategy by using the similarity scores between parallel sentences in such a way that the obfuscation degree can be adjusted in plagiarized passages. This corpus is the first bilingual plagiarism corpus for Persian language.

In the future works, we plan to improve our corpus by incorporating other obfuscation strategies such as manual obfuscation and artificial obfuscation in the corpus. We also plan to extend our corpus in other languages.

[1] http://www.ictrc.ir/plaglab/corpora/Bilingual_Persian_English_Corpus(asghari15).zip

Acknowledgement

This work has been accomplished in ICT research Institute, ACECR, under the support of Vice Presidency for Science and Technology of Iran - grant No. 1164331. The authors gratefully acknowledge the support of aforementioned organizations. Special thanks go to the members of ITBM research group for their valuable collaboration. The authors also would like to express their gratitude to Leila Tavakoli and Hamed Zamani.

References

1. Potthast, Martin, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. "Overview of the 5th international competition on plagiarism detection." In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pp. 301-331. CELCT, 2013.

2. Potthast, Martin, Matthias Hagen, Steve Göring, Paolo Rosso, and Benno Stein. Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. In Working Notes Papers of the CLEF 2015 Evaluation Labs, CEUR Workshop Proceedings, September 2015. CLEF and CEUR-WS.org. ISSN 1613-0073.

3. Potthast, Martin, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. "Cross-language plagiarism detection." *Language Resources and Evaluation* 45, no. 1 (2011): 45-62.

4. Juričić, Vedran, Vanja Štefanec, and Siniša Bosanac. "Multilingual plagiarism detection corpus." In *MIPRO, 2012 Proceedings of the 35th International Convention*, pp. 1310-1314. IEEE, 2012.

5. Kittur, Aniket, Ed H. Chi, and Bongwon Suh. "What's in Wikipedia?: mapping topics and conflict using socially annotated category structure." In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1509-1512. ACM, 2009.

6. Ghayoomi, Masood, Saeedeh Momtazi, and Mahmood Bijankhan. "A study of corpus development for Persian." In *International Journal on ALP*. 2010.

7. Bijankhan, Mahmood, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. "Lessons from building a Persian written corpus: Peykare." *Language resources and evaluation* 45, no. 2 (2011): 143-164.