Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation

Notebook for PAN at CLEF 2015

Khadijeh Khoshnavataher, Vahid Zarrabi, Salar Mohtaj, Habibollah Asghari

ICT Research Institute,
Academic Center for Education, Culture and Reseach (ACECR), Iran

{khadijeh.khoshnava, vahid.zarrabi, salar.mohtaj,habibollah.asghari}
@ ictrc.ir

Abstract. The task of text alignment corpus construction at PAN 2015 competition consists of preparing a plagiarism corpus so that it can provide various obfuscation types and versatile obfuscation degrees. Meanwhile, its format and metadata structure should follow previous PAN plagiarism corpora. In this paper, we describe our approach for construction of a monolingual Persian plagiarism corpus that can be used to evaluate the performance of Persian plagiarism detection systems.

Keywords: Plagiarism Detection, Corpus Construction, Text Alignment Corpus

1 Introduction

Plagiarism is the re-use of another person's ideas, processes, results, or words without explicitly acknowledging the original source [1]. Plagiarism detection is the algorithms for retrieval and extraction of text reuse within a suspicious document and corresponding source documents. The suspicious and source documents can be written either in the same language named as monolingual (MLPD) or in different languages named as cross lingual (CLPD) [2].

In the process of developing a system for plagiarism detection in natural language texts, the system should be trained and tested on a text corpus containing known plagiarized passage. The PAN is a major international competition for the task of plagiarism detection, and provides corpora for the evaluation of plagiarism detection systems. The corpus used in the first PAN competition known as PAN-PC-09 which, contain only artificial plagiarism cases [3]. In later revisions of the corpus, a variety of obfuscation strategies have been applied in corpora, including artificial obfuscation, simulated (or manual) obfuscation, translation obfuscation and summary obfuscation [4, 5, 6, 7, 8].

In this paper, in order to construct a monolingual plagiarism detection corpus for Persian language, we have deployed our approach based on PAN corpora strategies. We have used an artificial obfuscation strategy to create plagiarism cases.

The paper is organized as follow: In section 2, we describe our approach for corpus construction. Then in section 3 we will discuss the results of the corpus. Finally, we will conclude and discuss about some future works in section 4.

2 Our Approach

In this section the overall procedure of our approach for building monolingual Persian corpus is described. It is organized in five steps: preprocessing, documents clustering, fragment extraction, fragment obfuscation and inserting plagiarism cases in suspicious documents. In the following subsections, we describe the process of each step.

2.1 Preprocessing

Persian language belongs to the category of Arabic-Scripted based languages. There are some problems dealing with preprocessing in this language [9]. There are some efforts to develop Persian preprocessing algorithms [10, 11]. In this paper, in the preprocessing stage of the system, we have applied some algorithms such as normalization, tokenization, stemming and part of speech (POS) tagging.

2.2 Documents Clustering

Establishing topically similarity between suspicious passage and its corresponding source documents is an important issue for corpus construction. By inserting plagiarized passages into topically related surrounding text, the corpus may become more realistic. Therefore, in this step, collection of Wikipedia documents clustered into different topically related groups. A bipartite graph of documents-categories was created to cluster the documents. In the next step, the info- map community detection algorithm was applied to the graph and all communities were detected. Finally, Documents within a community are considered as one cluster. Each suspicious document and its corresponding source documents are selected from one cluster.

2.3 Fragments Extraction

The documents used in the corpus are divided into two categories: 50% of the documents selected as source documents and 50% are designated as suspicious documents. Note that only 50% of suspicious documents contain plagiarism cases.

The task of the fragment extraction is to extract fragments from source documents. The length of fragments is evenly distributed between 30 and 500 words. The length of fragments is shown in Table 1.

Table 1. Fragment lengths in words

Fragment Length	
Short	30 – 50 words
Medium	150 – 250 words
Long	300 – 500 words

2.4 Fragments Obfuscation

We have used an artificial obfuscation strategy for the plagiarism corpus. To create artificial plagiarism, we have used five obfuscation strategies as follows:

- None (No Obfuscation)

Source fragment without any change consider as obfuscated fragment. In other words, obfuscation fragment is an exact copy of source fragment.

- Random Change of Order

Given source fragment, obfuscation fragment is created by shuffling words at random. Since the tokenization of words in Persian is a challenging issue, so this task should be done under supervision.

- POS-preserving Change of Order

In order to accomplish this obfuscation, the sequence of parts of speech (POS) in source fragment is determined. Then, words are shuffling randomly, while retaining the original POS sequence.

- Synonym Substitution

The plagiarized fragment is created in such a way to replace some words by one of their synonyms.

- Addition / Deletion

Obfuscation fragment is created by inserting or removing words at random.

The number of operations made on source fragment specifies the degree of obfuscation. Different degrees of obfuscation are “None”, “Low”, “Medium”, and “High” obfuscation.

2.5 Insert Plagiarism Cases in Suspicious Documents

In this step, according to suspicious document’s length, one or more plagiarism cases which are in the same cluster of suspicious documents are selected. Then, each of them inserted at random positions in suspicious document.

The fraction of plagiarism in each document is not fixed. The percentage of plagiarism in each suspicious document is distributed between 5% and 100% of its length. The ratio of plagiarism per suspicious documents is shown in Table 2.

Table 2. Ratio of Plagiarism fragments in Documents

Plagiarism per Document	
Little	5% - 20%
Medium	20% - 50%
Much	50% - 80%
Very Much	80% - 100%

Finally, for each pair of source and suspicious documents, an XML file is created which contains Meta information about the plagiarism cases. These include:

- this_length: Length of plagiarism case in 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 source fragment in the source document.

3 Results

In this section, we have presented the result and statistics of our corpus. Because of the lack of Persian plagiarism detection corpora, which contain manual cases of plagiarism, we cannot compare plagiarism cases in our corpus to actual cases. An overview of important corpus statistics is shown in Table 3. The corpus is based on 2114 documents from Wikipedia articles.

For developing this corpus and other corpora in our laboratory, we have developed a web based application that can process the input documents and construct various plagiarism corpora based on corpus builder settings. The reason for implementing the corpus builder in a web application is for crowd sourcing the simulated plagiarism cases and inserting them in resulted corpus.

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

[1] http://www.ictrc.ir/plaglab/corpora/Monolingual_Persian_Corpus(khoshnava15).zip

Table 3. Monolingual Persian Corpus statistics

Documents	
The number of source documents:	1057
The number of suspicious documents:	
With plagiarism:	529
No plagiarism:	528
Plagiarism cases	
The number of plagiarism cases:	
No obfuscation cases:	259
With obfuscation cases:	564
Plagiarism per Document	
The number of Little plagiarized documents:	301
The number of Medium plagiarized documents:	80
The number of Much plagiarized documents:	96
The number of Very much plagiarized documents:	52

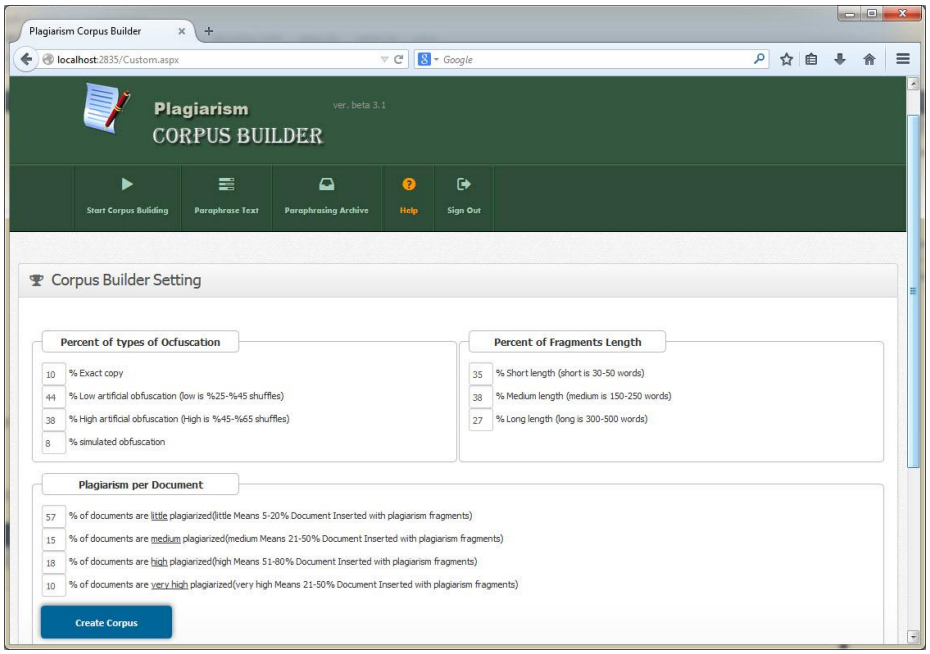

Fig. 1. Snapshot of our Plagiarism Corpus Builder

4 Conclusion and Future Work

We have discussed our approach to the task of text alignment in the context of PAN 2015 competition. We describe a system that generates a monolingual Persian plagiarism detection corpus. This corpus is the first plagiarism detection corpus in Persian language and is intended to be used for testing the performance of plagiarism detection systems.

In our future work, we plan to improve our corpus by implementing obfuscation techniques such that simulated obfuscation and other obfuscation strategies using plagiarism corpus builder.

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 express their gratitude to M.R. Ghahari and Samira Rezaei.

References

1. Barrón-Cedeño, Alberto, Marta Vila, M. Antònia Martí, and Paolo Rosso. "Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection." *Computational Linguistics* 39, no. 4 (2013): 917-947.

2. Barrón-Cedeno, Alberto, and Paolo Rosso. "Monolingual and Crosslingual Plagiarism Detection. Towards the Competition@ SEPLN09." (2009): 29-32.

3. Eiselt, Andreas, Martin Potthast, Benno Stein, and Alberto Barrón-Cedeno Paolo Rosso. "Overview of the 1st international competition on plagiarism detection." In *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*, p. 1. 2009.

4. Potthast, Martin, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. "Overview of the 2nd International Competition on Plagiarism Detection." In *CLEF (Notebook Papers/LABs/Workshops)*. 2010.

5. Potthast, Martin, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, and Paolo Rosso. "Overview of the 3rd International Competition on Plagiarism Detection." In *CLEF (Notebook Papers/LABs/Workshops)*. 2011.

6. Potthast, Martin, Tim Gollub, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. Overview of the 4th International Competition on Plagiarism Detection. In Working Notes Papers of the CLEF 2012 Evaluation Labs, September 2012. ISBN 978-88-904810-3-1. ISSN 2038-4963.

7. 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.

8. Potthast, Martin, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein "Overview of the 6th International Competition on Plagiarism Detection." *CLEF (Online Working Notes/Labs/Workshop)*. 2014.

9. Shamsfard, Mehrnoush. "Challenges and open problems in Persian text processing." *Proceedings of LTC* 11 (2011).

10. Sarabi, Zahra, Hamidreza Mahyar, and Mojgan Farhoodi. "ParsiPardaz: Persian Language Processing Toolkit." In *Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on*, pp. 73-79. IEEE, 2013.

11. Seraji, Mojgan, Beáta Megyesi, and Joakim Nivre. "A basic language resource kit for Persian." In *Eight International Conference on Language Resources and Evaluation (LREC 2012), 23-25 May 2012, Istanbul, Turkey*, pp. 2245-2252. European Language Resources Association, 2012.