Mahak Samim: A Corpus of Persian Academic Texts for Evaluating Plagiarism Detection Systems

Morteza Rezaei Sharifabadi
Computer Research Center of Islamic Sciences
Tehran, I. R. Iran.
m.rezaei@noornet.net

Seyed Ahmad Eftekhari
Computer Research Center of Islamic Sciences
Tehran, I. R. Iran.
s.ahmad.ef@gmail.com

ABSTRACT
In this paper we introduce Mahak Samim, a plagiarism detection corpus that consists of Persian academic texts in which plagiarism cases are embedded. This corpus, which can be used for evaluating plagiarism detection systems, consists of more than five thousand artificial plagiarism cases with various lengths and diverse degrees of obfuscation. The development process and the features of the corpus are described here.

CCS Concepts
• Information systems → Information retrieval → Retrieval tasks and goals → Near-duplicate and plagiarism detection.

Keywords
plagiarism detection; evaluation corpus; Persian; academic texts.

1. INTRODUCTION
Plagiarism is defined as "copying or closely imitating the work of another writer, composer, etc., without permission and with the intention of passing the results off as original work" [12]. Plagiarism detectors are software programs developed to detect cases of such misconduct in documents. The PAN evaluation lab series has provided a framework for evaluating plagiarism detection systems. This framework relies on plagiarism corpora which are basically collections of text that include cases of plagiarism. Plagiarism detection systems receive the corpus texts as input and their ability to detect the plagiarism cases embedded in the texts are examined.

Since plagiarism detectors are not entirely language independent, there is a need for plagiarism corpora in various languages. In recent years a couple of Persian plagiarism detection systems have been developed. Proper evaluation of these systems is dependent on reliable Persian plagiarism corpora.

In this paper we introduce Mahak Samim[1], a corpus suitable for evaluating Persian plagiarism detectors. We first briefly review previous works in this field and then we introduce our own approach. The paper concludes with a summary and an outlook for further work.

2. RELATED WORK
Prior to PAN evaluation lab series, plagiarism corpora were rare. The corpora used in PAN labs held in 2009 [4], 2010 [5] and 2011 [6] had basically the same structure. The documents used in these corpora were books from Project Gutenberg. The corpora were used to evaluate both external and intrinsic plagiarism detection. In external plagiarism detection suspicious documents are checked against a collection of source documents, but in intrinsic plagiarism suspicious documents are analyzed in isolation for changes in writing style etc. Fifty percent of the documents were used as source documents and fifty percent as suspicious documents. The corpora contain plagiarism cases with different lengths and various degrees of artificial and simulated obfuscation. Artificial obfuscation includes techniques such as automatically shuffling and replacing words and simulated obfuscation was achieved through crowdsourcing the obfuscation task. The major shortcoming of corpora presented in these years was their relatively small size. The plagiarism detectors were expected to include a stage of heuristic retrieval in which they selected a group of candidate documents among the total collection of source documents. However, since the size of the corpora were not large enough, the systems skipped this stage. In PAN 2012 [7] this issue is addressed and a new approach is adopted for developing the plagiarism corpus. For this purpose a number of professional writers were asked to write articles – containing plagiarism - on a set of topics. A one billion document corpus resembling the web was used as the collection of source documents. The writers compiled their articles by searching through this huge collection. In PAN 2013 [9] and PAN 2014 [8] expanded versions of the 2012 corpus were used.

In PAN 2015 [10] a task of corpus construction was introduced. In this task, participants were asked to provide their own plagiarism corpora. Eight plagiarism corpora were provided for this task among which two included Persian documents. Khoshnavataher et.al. [3] present a monolingual Persian corpus based on about 2100 Wikipedia articles with plagiarism cases obfuscated artificially and intended for evaluation of extrinsic plagiarism detection. Asghari et.al. [1] use Wikipedia documents and a Persian-English sentence-aligned corpus to develop a bilingual plagiarism detection corpus.

3. CORPUS DEVELOPMENT
3.1 Document Collection
Academic papers are one of the major types of texts subject to plagiarism. In order to cover such texts in our corpus, we collected Persian papers from peer reviewed journals. We crawled the websites of journals introduced in the System for Evaluation of Scientific Journals[2] (affiliated to Iran's Ministry of Science, Research and Technology) and we downloaded papers from journal websites that provide free full-text access to their articles in plain-text format. Table1 shows the statistics of the number of

[1] Samim-Noor is a commercial plagiarism detection system developed by the Computer Research Center of Islamic Sciences. Mahak in Persian means "Touchstone".

[2] http://journals.msrt.ir/

documents in each subject, as grouped by the System for Evaluation of Scientific Journals, and Table 2 provides information about the document lengths.

Table 1. Statistics of number of documents per subject

Subject	Number of documents
Humanities	2697
Science	1204
Veterinary Science	469
Agriculture and Natural Resources	281
Engineering	38
Art and Architecture	18
Total	4707

Table 2. Statistics of document lengths

Document Length	Percent of Documents
short (1-3000 words)	20 %
medium (3000-6000 words)	50 %
long (6000-30000 words)	30 %

3.2 Source / suspicious documents

In plagiarism corpora, the documents collection is usually split into two main subgroups i.e. source documents and suspicious documents. Source documents are documents from which parts of text are selected as plagiarism cases. These parts are then inserted inside the text of so-called suspicious documents. In other words, suspicious documents are documents which include text used in source documents. We follow PANs tradition of using half of the documents as source documents and half as suspicious documents. It is noteworthy that the subjects of the papers were taken into consideration while dividing the collection into halves. i.e. 50 percent of the papers in humanities were used as source documents and 50 percent as suspicious documents etc.

3.3 Plagiarism per document

50 percent of the suspicious documents have no plagiarism cases. As mentioned in [11], the documents without plagiarism allow to determine whether or not a detector can distinguish plagiarism cases from overlaps that occur naturally between random documents. Statistics of plagiarism per document in the rest of the suspicious documents, i.e. 25 percent of the whole corpus, is available in Table 3.

Table 3. Statistics of plagiarism per document in documents with plagiarism

Plagiarism Per Document	Percent of Documents
hardly (5%-20%)	30 %
medium (20%-50%)	25 %
much (50%-80%)	30 %
entirely (>80%)	15 %

3.4 Plagiarism case length

Our corpus consists of a total of 5862 plagiarism cases with lengths between 50 and 5000 words. Table 4 shows the statistics. Long plagiarism cases may include more than one sentence.

Table 4. Statistics of lengths of plagiarism cases

Plagiarism Case Length	Percent of Cases
Short (50-150 words)	34 %
Medium (300-500 words)	33 %
Long (3000-5000 words)	33 %

3.5 Topic match

The six general topic categories of the papers used in our corpus were introduced in table 1. Fifty percent of the plagiarism cases were made between papers with same topics (intra-topic cases) and fifty percent between papers with different topics (inter-topic cases).

3.6 Obfuscation types

In many cases, plagiarized texts are manipulated by those committing plagiarism in order to avoid being detected by plagiarism detection systems or human readers. Plagiarism corpora developers use different techniques to include such obfuscations in their plagiarism cases. An overview of different types of obfuscation in our plagiarism cases is available in Table 5.

Table 5. Statistics of Obfuscation types

Obfuscation	Percent of Cases
None	40 %
Random Text Operations	
> low obfuscation	20 %
> high obfuscation	20 %
Semantic Word Variation	
> low obfuscation	10 %
> high obfuscation	10 %

As shown in table 4, 40 percent of the plagiarism cases have no obfuscation. As explained in [11], since the writing style of the original author is preserved in plagiarism cases without obfuscation, these cases are especially appropriate for evaluating intrinsic plagiarism detection. Random text operations are operations such as adding, deleting and substituting words, which are all done randomly. Semantic word variation, on the other hand, is the random substitution of words with their synonyms. We use the Comprehensive Dictionary of Persian Synonyms and Antonyms[3] as a resource for extracting synonyms. The terms "low obfuscation" and "high obfuscation" mentioned in table 4 show the degree of obfuscation i.e. how many words have been added, deleted or substituted etc.

4. SUMMARY AND FUTURE WORK

As explained above, Mahak Samim is a plagiarism corpus which can be used for evaluating both intrinsic and external plagiarism detection systems. In order to preserve overall balance, many factors – plagiarism per document, plagiarism case length, topic match, obfuscation type, and obfuscation degree – were taken into consideration while preparing each plagiarism case. The corpus files are prepared according to the format of previous PAN

[3] The plain-text version of this dictionary can be downloaded from this link: http://dadegan.ir/catalog/D3911124a

corpora which include xml files that have information about the starting point of the plagiarism in relevant source and suspicious documents and the length of the plagiarism case.

Plagiarism cases in our corpus are cases of "artificial plagiarism". Using "real plagiarism" cases in plagiarism corpora is problematic due to ethical, legal, and financial issues [11]. However, we may enrich our corpus by adding cases of simulated plagiarism. Other types of artificial obfuscation, such as POS-preserving word shuffling could also be employed. The corpus may be easily expanded with both academic papers and other types of documents such as books, web articles, etc.

This paper has been submitted to The PAN@FIRE2016 Shared Task on Persian Plagiarism Detection and Text Alignment Corpus Construction [2] and the corpus is available through Peykaregan[4] website.

5. ACKNOWLEDGMENTS

Special thanks to Dr. Martin Potthast for his valuable help and to Dr. Mahdi Behnia and Mr. Amirhossein Rajabzadeh Assarha, our colleagues in the Computer Research Center of Islamic Sciences, for their support and their comments.

6. REFERENCES

[1] Asghari, H., Khoshnava, K., Fatemi, O. and Faili, H., 2015. Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus. In *Working Notes Papers of the CLEF 2015*.

[2] Asghari, H., Mohtaj, S., Fatemi, O., Faili, H., Rosso, P., and Potthast, M., 2016. Algorithms and Corpora for Persian Plagiarism Detection: Overview of PAN at FIRE 2016. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[3] Khoshnavataher, K., Zarrabi, V., Mohtaj, S. and Asghari, H., 2015. Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation. In *Working Notes Papers of the CLEF 2015*.

[4] Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B. and Rosso, P., 2009. Overview of the 1st international competition on plagiarism detection. In *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*.

[5] Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P., 2010. Overview of the 2nd international competition on plagiarism detection. In *Notebook Papers of CLEF 2010 LABs and Workshops.*

[6] Potthast, M., Eiselt, A., Barrón Cedeño, L.A., Stein, B. and Rosso, P., 2011. Overview of the 3rd international competition on plagiarism detection. In*CEUR Workshop Proceedings.*

[7] Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B., 2012. Overview of the 4th international competition on plagiarism detection. In *Working Notes Papers of CLEF 2012 Evaluation Labs and Workshop.*

[8] Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, Rosso, P. and Stein, B., 2014. Overview of the 6th international competition on plagiarism detection. *Working Notes Papers of the CLEF 2014 Evaluation Labs*, *CEUR Workshop Proceedings.*

[9] Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E. and Stein, B., 2013. Overview of the 5th international competition on plagiarism detection. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation* (pp. 301-331). CELCT.

[10] Potthast, M., Hagen, M., Göring, S., Rosso, P. and Stein, B., 2015. Towards data submissions for shared tasks: first experiences for the task of text alignment. *Working Notes Papers of the CLEF 2015*, pp.1613-0073.

[11] Potthast, M., Stein, B., Barrón-Cedeño, A. and Rosso, P., 2010. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 997-1005). Association for Computational Linguistics.

[12] Reitz, J.M., 1996. *ODLIS: Online dictionary for library and information science*. Libraries Unlimited.

[4] www.peykaregan.com