A Pairwise Document Analysis Approach for Monolingual Plagiarism Detection

Nava Ehsan
School of Electrical and Computer Engineering
College of Engineering
University of Tehran
n.ehsan@ece.ut.ac.ir

Azadeh Shakery
School of Electrical and Computer Engineering
College of Engineering
University of Tehran
shakery@ut.ac.ir

ABSTRACT

The task of plagiarism detection entails two main steps, suspicious candidate retrieval and pairwise document similarity analysis also called detailed analysis. In this paper we focus on the second subtask. We will report our monolingual plagiarism detection system which is used to process the Persian plagiarism corpus for the task of pairwise document similarity. To retrieve plagiarised passages a plagiarism detection method based on vector space model, insensitive to context reordering, is presented. We evaluate the performance in terms of precision, recall, granularity and plagdet metrics.

CCS Concepts

•Information systems → Near-duplicate and plagiarism detection; •Applied computing → Document analysis;

1. INTRODUCTION

The task of plagiarism detection comprises two main steps: candidate document retrieval and pairwise document detection or detailed analysis. Some researchers also included a third step, called post-processing, where the extracted passage pairs are cleaned, filtered, and possibly visualized for later presentation [18]. Detailed analysis of plagiarism detection task is retrieving passages of text which have originated from another document. This process is comparing source and suspicious pairs and retrieving plagiarized fragments. In this paper we introduce a vector space model for this task. The proximity of words are considered by dividing the text into passages. After creating the passages we didn't take into account the order of the words in each passage. Thus, this approach is insensitive to punctuations, extra white spaces and permutation of the document context in the passages. We also didn't use any language specific feature. Thus, our approach is applicable in any language. The result we obtained on the training data was about 0.82 with respect to Plagdet score.

The rest of the paper is organized as follows: Section 2 outlines related works in monolingual plagiarism detection. Section 3 describes the pairwise document similarity approach. Finally experimental results are discussed in Section 4. Conclusion and future work are reported, in Section 5.

2. RELATED WORK

Plagiarism detection could be classified into two classes, intrinsic and external also called without reference or with reference, respectively. Intrinsic evaluation is referred to those methods, which use style analysis to detect parts of the text that are inconsistent in terms of writing style [15, 14]. The aim of external plagiarism detection is not only finding the suspicious text but also finding the reference of the suspicious text. In monolingual plagiarism detection, the suspicious text could be an exact copy or a modified copy [17] and it should be large enough to be more than just a coincidence.

One method for monolingual plagiarism detection is comparing fragments of suspicious and source documents using fingerprint indexing. Winnowing approach [22], which is used in the widely used plagiarism detection tool, MOSS, is based on fingerprint indexing.

There are different approaches for detection monolingual plagiarism detection. Some of these approaches could be classified into fingerprinting [22, 13], string matching [8], using stopwords [24], vector space models [12, 5], probability models [3], classification [16], semantic models [6, 7] and structural models [23]. According to monolingual experiments in [4] paraphrasing could make plagiarism detection more difficult. There have been some works for detecting paraphrased sentences in monolingual texts [1].

In recent years PAN competition offers evaluation environment to evaluate plagiarism detection algorithms. This competition also offers evaluation corpora. PAN@FIRE Persian Plagdet 2016 competition [2] prepares a set of suspicious and source documents written in Persian and the task is to find all plagiarized sections in the suspicious documents and, if available, the corresponding source sections. This external plagiarism detection task provides a situation to evaluate Persian plagiarism detection systems.

3. PAIRWISE DOCUMENT SIMILARITY ANALYSIS

In this section we deal with pairwise document similarity analysis and we will introduce a method based on a vector space model to detect plagiarized fragments of specified set of suspicious and source document pairs. We assume that the source and suspicious pairs are pre-defined for the system. According to PAN competition this process is called, detailed analysis stage. The problem of pairwise document analysis is defined as follows: let S' be a suspicious document and S be a source document that is likely to contain similar passages to some passages in S'. S and S' are compared section-wise using a language retrieval model. A plagiarism is considered, if for a pair of sections (S_f and $S'_{f'}$) a similarity above a threshold is detected.

According to [18] the detection approaches of this sub-task includes three building blocks named (1) seeding, (2) match merging, and (3) extraction filtering. In the following subsections, we describe them in detail.

3.1 Seeding

Given a suspicious document and a source document, matches between the two documents are identified using some seed [18]. In our approach, first preprocessing phase is performed. Given a document, a preprocessing phase is performed. We substitute Arabic ي and ك with Persian ی and ک. The reason is that the mentioned two letters have different character encodings. Then, stopwords, punctuations and extra white spaces are removed and tokens are extracted.

Since, plagiarism usually happens in parts of the text, a plagiarism detection method should be able to detect local similarities where only a short passage may be in common in both documents. Thus, there is a requirement to segment the texts into fragments. For each document pair, we split the texts into sentences by using ".", "?" and "!" marks. We choose an amount of consecutive sentences as the smallest unit of plagiarism. Documents are divided into some fragments each containing n sentences with one sentence overlap. The sensitivity of the algorithm with respect to parameter n is shown in Section 4.

After sentence splitting, a vector space model is created. The terms of the source document are considered as the vocabulary and the binary weighting schema is used by setting the i^{th} index 1 if the i^{th} term occurs in the fragment and 0 otherwise. We regard suspicious passage S_f and reference passage $S'_{f'}$ as pairs of plagiarism candidate sentences whose cosine similarity is greater than a threshold t_1, and at least three terms of the source text are found in the suspicious fragment. The last criterion is added to avoid retrieving fragments with coincidental similarity. S_f will be considered as a plagiarism source of $S'_{f'}$ if it has maximum cosine similarity among all source fragments with similarity above the threshold.

The vector creation approach is similar to Eurovoc-based model proposed in [20] except that we use it for monolingual texts rather than cross-lingual texts. Thus, this approach could be easily adapted to cross language plagiarism detection.

3.2 Match Merging

Finding seed matches between a suspicious and a source document, they are merged into aligned passages of maximal length which are then reported as plagiarism detections [18]. To improve performance with respect to the granularity metric, we merge adjacent suspicious and source fragments to report a single plagiarism case. If the number of characters between two detected fragments of source and suspicious documents are below a threshold t_2 those fragments are considered as adjacent fragments and they are merged to report a single plagiarism case.

3.3 Extraction Filtering

Given a set of aligned passages, a passage filter removes all aligned passages that do not meet certain criteria. For example dealing with overlapping passages or extremely short passages [18].

After retrieving potential plagiarized fragments in previous steps, the sentences within a fragment are partitioned into non-overlapping n-grams. For extraction filtering step we applied a method similar to result filtering approach proposed in [10] for excluding false positive detections and improving precision. We used the dynamic algorithm to find the alignments between the n-grams of the source and suspicious texts and then the null alignments are excluded from the start and end of the reported fragments.

4. RESULTS

The results of Persian detailed analysis subtask on PAN@FIRE Persian Plagdet 2016 [2] using the training data is summarized in Tables 1, 2 and 3. The experimentation platform TIRA is used for our evaluations [11, 19].

Parameters t_1 and n are respectively the similarity threshold and number of consecutive sentences described in Section 3.1. The adjacency threshold t_2, described in Section 3.2, is set to 1500 characters. The n value for n-gram creation described in Section 3.3 for each sentence is chosen to be 9, except for the final partition of the sentence that may include fewer or more than 9 terms.

The evaluation metrics are described in [21]. The evaluation is based on macro-average precision and recall. Also, the granularity measure characterizes the power of a detection algorithm. It shows whether a plagiarism case is detected as a whole or in several pieces. The Plagdet score is the combination of the three metrics, precision, recall and granularity defined as follows for plagiarism cases S and plagiarism detections R [21]:

$$plagdet(S, R) = \frac{F_1}{\log(1 + granularity(S, R))} \quad (1)$$

The results of the tables show that the plagdet score improves by decreasing the amount of the sentences in a fragment. The reason could be that there are more short plagiarized texts than long ones in the dataset. The increase in precision comparing different number of sentences in a fragment using 0.3 for the value t_1, could show that decreasing the amount of sentences may have higher impact on decreasing rather than increasing the false positive detections.

We also analysed the effect of extraction filtering part of the approach and we realized that without applying this stage for (t_1 = 0.3, n = 3) the precision was 0.6026 and the plagdet score was 0.7087. This shows that this step improved the plagdet score about 14 percent.

We used 0.3 for t_1 and 5 sentences for parameter n (number of consecutive sentences) for the test set. The results on the test set are shown in Table 4.

5. CONCLUSION AND FUTURE WORK

The task of plagiarism detection entails two sub-tasks, suspicious candidate retrieval and pairwise document similarity. We introduce a pairwise document analysis approach for Persian language. An approach based on a vector space model is described for computing pairwise document similarity. The principle of this approach is that sentences containing more common words are likely to be a source of plagiarism. The method contains three building blocks named seeding, match merging and extraction filtering. Our work is tested on a Persian corpora which offers evaluation environment to evaluate plagiarism detection algorithms. The proposed approach is insensitive to context reordering and could be applied in any language.

Detailed analysis subtask will be improved by expanding the representative words of the document to find appropriate substitutes for a word in the context in order to capture intelligent plagiarisms. For this reason, there is requirement to minimize the risk of noises that word expansion may cause. Other weighting schemas such as $tf-idf$ weighting could be applied in comparing the vectors of the texts. A complete plagiarism detection system could be developed by adding a candidate selection [9] step before pairwise document analysis.

	Precision	Recall	Granularity	Plagdet
($t_1 = 0.4$, $n = 5$)	0.8532	0.5357	1	0.6582
($t_1 = 0.3$, $n = 5$)	0.7630	0.7486	1	0.7557
($t_1 = 0.2$, $n = 5$)	0.4004	0.8151	1	0.5370
($t_1 = 0.3$, $n = 3$)	0.7867	0.8304	1	0.8080
($t_1 = 0.3$, $n = 2$)	0.8482	0.7876	1	0.8168

Table 1: Results of detailed analysis sub-task using macro-averaged precision, recall, granularity, and the plagdet score

	Precision	Recall	F_1
($t_1 = 0.4$, $n = 5$)	0.9434	0.5478	0.6931
($t_1 = 0.3$, $n = 5$)	0.8484	0.7631	0.8035
($t_1 = 0.2$, $n = 5$)	0.4379	0.8292	0.5731
($t_1 = 0.3$, $n = 3$)	0.8931	0.8518	0.8720
($t_1 = 0.3$, $n = 2$)	0.9069	0.8162	0.8592

Table 2: Results of detailed analysis sub-task in case-level

	Precision	Recall	F_1
($t_1 = 0.4$, $n = 5$)	0.9457	0.5534	0.6982
($t_1 = 0.3$, $n = 5$)	0.8798	0.7701	0.8213
($t_1 = 0.2$, $n = 5$)	0.6266	0.8342	0.7156
($t_1 = 0.3$, $n = 3$)	0.9312	0.8555	0.8918
($t_1 = 0.3$, $n = 2$)	0.9452	0.8220	0.8793

Table 3: Results of detailed analysis sub-task in document-level

	Precision	Recall	Granularity	Plagdet
($t_1 = 0.3$, $n = 5$)	0.7496	0.7050	1	0.7266

Table 4: Results of detailed analysis sub-task on the test set

Acknowledgement

We would like to acknowledge the assistance and information provided by Hossein Nasr Esfahani and Mahsa Shahshahani.

6. REFERENCES

[1] I. Androutsopoulos and P. Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135 – 187, 2009.

[2] H. Asghari, S. Mohtaj, O. Fatemi, H. Faili, P. Rosso, and M. Potthast. Algorithms and corpora for persian plagiarism detection: Overview of pan at fire 2016. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation*, CEUR Workshop Proceedings. CEUR-WS.org, 2016.

[3] A. Barrón-Cedeño, P. Rosso, and J.-M. Benedí. Reducing the plagiarism detection search space on the basis of the kullback-leibler distance. In *Computational Linguistics and Intelligent Text Processing*, pages 523–534. Springer, 2009.

[4] A. Barrón-Cedeno, M. Vila, M. A. Martí, and P. Rosso. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947, 2013.

[5] S. Brin, J. Davis, and H. Garcia-Molina. Copy detection mechanisms for digital documents. In *ACM SIGMOD Record*, volume 24, pages 398–409. ACM, 1995.

[6] C.-Y. Chen, J.-Y. Yeh, and H.-R. Ke. Plagiarism detection using rouge and wordnet. *Journal of Computing*, pages 34 – 44, 2010.

[7] M. Chong and L. Specia. Lexical generalisation for word-level matching in plagiarism detection. In *RANLP*, pages 704–709, 2011.

[8] P. Clough and M. Stevenson. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45(1):5–24, 2011.

[9] N. Ehsan and A. Shakery. Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information. *Information Processing & Management*, 52(6):1004–1017, 2016.

[10] N. Ehsan, F. W. Tompa, and A. Shakery. Using a dictionary and n-gram alignment to improve fine-grained cross-language plagiarism detection. In *Proceedings of the 2016 ACM Symposium on Document Engineering*, pages 59–68. ACM, 2016.

[11] T. Gollub, B. Stein, and S. Burrows. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In B. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*, pages 1125–1126. ACM, Aug. 2012.

[12] C. Grozea, C. Gehl, and M. Popescu. Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection. In *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*, pages 10 – 18, 2009.

[13] G. S. Manku, A. Jain, and A. Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150. ACM, 2007.

[14] S. Meyer zu Eißen and B. Stein. Intrinsic Plagiarism Detection. In *Advances in Information Retrieval. 28th European Conference on IR Research (ECIR 06)*, volume 3936 LNCS of *Lecture Notes in Computer Science*, pages 565–569, Berlin Heidelberg New York, 2006. Springer.

[15] G. Oberreuter and J. D. Velásquez. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications*, 40(9):3756–3763, 2013.

[16] R. C. Pereira, V. P. Moreira, and R. Galante. A new approach for cross-language plagiarism analysis. In *Multilingual and Multimodal Information Access Evaluation*, volume 6360, pages 15–26. 2010.

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

[18] M. Potthast, T. Gollub, M. Hagen, J. Kiesel, M. Michel, A. Oberländer, M. Tippmann, A. Barrón-Cedeño, P. Gupta, P. Rosso, et al. Overview of the 4th international competition on plagiarism detection. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[19] M. Potthast, T. Gollub, F. Rangel, P. Rosso, E. Stamatatos, and B. Stein. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, and E. Toms, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, Sept. 2014. Springer.

[20] M. Potthast, B. Stein, and M. Anderka. A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530. 2008.

[21] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 997–1005. Association for Computational Linguistics, 2010.

[22] S. Schleimer, D. S. Wilkerson, and A. Aiken. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 76–85. ACM, 2003.

[23] A. Si, H. V. Leong, and R. W. Lau. Check: a document plagiarism detection system. In *Proceedings of the 1997 ACM symposium on Applied computing*, pages 70–77. ACM, 1997.

[24] E. Stamatatos. Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, 62(12):2512–2527, 2011.