External Plagiarism Detection using Information Retrieval and Sequence Alignment

Notebook for PAN at CLEF 2011

Rao Muhammad Adeel Nawab, Mark Stevenson and Paul Clough

University of Sheffield, UK.
{r.nawab|m.stevenson}@dcs.shef.ac.uk, p.d.clough@sheffield.ac.uk

Abstract This paper describes the University of Sheffield entry for the 3rd International Competition on Plagiarism Detection which attempted the monolingual external plagiarism detection task. A three stage framework was used: preprocessing and indexing, candidate document selection (using an Information Retrieval based approach) and detailed analysis (using the Running Karp-Rabin Greedy String Tiling algorithm). The submitted system obtained an overall performance of 0.0804, precision of 0.2780, recall of 0.0885 and granularity of 2.18 in the formal evaluation.

Keywords: external plagiarism detection, information retrieval, greedy string tiling

1 Introduction

In recent years, plagiarism and its detection has received significant attention within both academia and industry [1,3]. The task of plagiarism detection itself can be divided into two main categories: (1) external plagiarism detection and (2) intrinsic plagiarism detection. The goal of *external plagiarism detection* is to identify the source (or original) document(s) that have been used to plagiarise a suspicious document. On the other hand, in *intrinsic plagiarism detection* the source documents are not available and plagiarised text is identified by looking for stylistic inconsistencies or text which is different from the rest.

In the 2011 PAN competition the University of Sheffield entry attempted the monolingual external plagiarism detection task. Our system did not attempt translated or multilingual plagiarism detection.

1.1 Related Work

The field of plagiarism detection has been a well-studied area over the years. However, direct comparison of performance using different existing approaches was hampered by the lack of a standard evaluation resource. Since 2009, PAN has been organising an international competition on plagiarism detection to evaluate the performance of different approaches using a common data set.

The systems that participated in 1st and 2nd PAN competitions [9,8] normally used a multi-stage process for plagiarism detection: pre-processing, candidate retrieval, detailed analysis and post-processing. The pre-processing step normally involved stemming, stop word removal, sorting word n-grams etc. The aim of applying different pre-processing techniques was to normalize the effect of obfuscation. The majority of the systems used an IR based approach (with and without hashing) for candidate document selection. Using this approach the entire source collection is converted to fixed length word n-grams or fingerprints and indexed. Each word n-gram or fingerprint in a suspicious document is queried in the index and source documents with word n-grams or fingerprints above some pre-defined threshold are marked as potential candidates. For the detailed analysis stage, heuristic sequence alignment algorithms are often used to extract suspicious-source section pairs. Portions of text that match exactly are used as seeds to identify longer passages using match merging heuristics. In the post-processing step, passages shorter than a pre-defined length or whose similarity score was less than given threshold under a retrieval model were discarded. In addition, passages that are ambiguous (could have been derived from multiple source documents) were discarded.

2 External Plagiarism Detection

Our proposed system consists of three stages: 1) pre-processing and indexing, 2) candidate document selection and 3) detailed analysis using Running Karp-Rabin Greedy String Tiling (RKR-GST).

2.1 Preprocessing and Indexing

Each document in the source and suspicious collections was split into sentences using the NLTK sentence detector [2]. The text was converted into lower case and all the non-alphanumeric characters removed. Documents in the source collection were then indexed with the Terrier IR system [6].

2.2 Candidate Document Selection

The aim of candidate document selection stage is to identify the source documents for each suspicious document. This is an important part of the process in a multi-stage approach to external plagiarism detection since source documents missed at this stage cannot be retrieved in a later stage. As many of the source documents as possible should be obtained at this stage, however, the total number of documents that are identified is limited by the processing required for the detailed analysis stage. Our system uses Information Retrieval for Candidate Document Selection.

The process of candidate document retrieval works as follows. A suspicious document is split into sentences which are used as queries. The index is queried against each sentence to retrieve a set of source (or original) document(s) and the top K documents selected for each query. Results of multiple queries are merged using a score-based fusion approach to generate final list of ranked source documents. A linear combination of the scores based on the CombSUM approach was used [4]. In the CombSUM

method, the final score, $S_{finalscore}$, is obtained by adding the scores obtained against each query q:

$$S_{finalscore} = \sum_{q=1}^{N_q} S_q\,(d) \tag{1}$$

where N_q is the total number of queries to be combined and $S_q\,(d)$ is the similarity score of a document d for a query q .

2.3 Detailed Analysis

For the detailed analysis stage, we used the same sequence alignment algorithm, Running Karp-Rabin Greedy String Tiling (GST), as our entry for last year's competition (see [5] for a detailed description).

The suspicious and candidate source document pairs identified in the candidate selection stage are each represented as a sequence of tokens. The sequences are aligned using the GST algorithm after which aligned tokens are merged using match merging heuristics to generate longer aligned sections. In post-processing, sections whose length is less than a certain threshold are discarded and a final set of source-suspicious section pairs are reported.

The behavior of the detailed analysis stage can be changed by adjusting various parameters: 1) *length of longest match* (α_{length}) filters candidate documents for further analysis. If α_{length} between a pair of aligned documents was greater than a certain threshold then it is analysed to identify suspicious-source section pairs, 2) *minimum match length* (mml) defines the minimum length of a match in aligning two sequences of tokens, 3) *length of gap* (α_{merge}) defines the distance between pairs of aligned passages which are merged into a single passage and 4) *discard length* $(\alpha_{discard})$ defines the minimum length for a merged section, any shorter than this are discarded.

3 Evaluation

3.1 System Development

The system was developed using the test corpus from the 2010 PAN competition (PAN-PC-10) [7].

Candidate Document Selection (Section 2.2): Results of the candidate document selection stage for different types of obfuscations are shown in Table 1. The top 10 documents were selected, retrieving any more leads to difficulties in processing the documents in the detailed analysis stage.

The recall for source documents is over 0.55 over the whole corpus but the type of obfuscation effects the recall. The best recall is obtained for simulated obfuscation but the results for the none, low and high types are much lower. We analysed the number of source documents were used to plagiarise each suspicious document and found the results varied by obfuscation type. All the suspicious documents that used simulated plagiarism were derived from fewer than 10 source documents. However, more documents were used for other types of obfuscation and 23.86% of none, 30.72% of low

Obfuscation	Precision	Recall	F1 Measure
Entire corpus	0.3135	0.5558	0.3301
None	0.3731	0.4967	0.3660
Low	0.3680	0.5010	0.3629
High	0.3558	0.4706	0.3442
Simulated	0.1111	0.6804	0.1857

Table 1. Performance for top 10 candidate documents using the PAN-PC-10 corpus

and 31.11% of high documents were plagiarised using 10 or more source documents. A large number of source documents adversely effects our approach to candidate document selection and effects performance for the none, low and high obfuscation types.

Detailed Analysis Stage (Section 2.3): Parameters for Greedy String Tiling algorithm were selected using a small number of documents from the PAN-PC-10 corpus. The values of α_{length}, α_{merge} and $\alpha_{discard}$ were varied and it was observed that a change in one parameter's value effects the system's performance. The best performance was observed with $\alpha_{length} > 5$, $\alpha_{merge} \leq 35$ characters and $\alpha_{discard} \leq 230$ characters. The computational effort required by Greedy String Tiling made it difficult to tune all the parameters on a large dataset.

MML	PlagDet Score	Recall	Precision	Granularity
2	0.0492	0.1285	0.0571	2.0480
3	**0.2201**	**0.2928**	**0.8479**	**2.9371**
4	0.1969	0.2848	0.9472	3.6719
5	0.1764	0.3884	0.8750	7.2786

Table 2. Performance of proposed system for detailed analysis stage on subset of 60 documents randomly selected from PAN-PC-10 corpus

A set of 60 documents was created by randomly selecting 15 suspicious documents for each obfuscation type: none, low, high and simulated. These were used to extract suspicious-source section pairs by choosing 10 candidate documents for each suspicious document. Table 2 shows the results when length of *mml* was changed from 2 to 5 using the best parameter values observed for a very small set of documents. Best results are obtained with *mml* = 3, indicating that a minimum match of three words gives good performance. The system gets good precision but recall is low and granularity is high. The reason for low recall is that only 10 documents were selected as candidates. GST algorithm can only detect exact matches and fails to detect paraphrasing. Cases of plagiarism created with high obfuscation might be increasing the granularity while merging exact matches.

3.2 System Performance

Our system achieved an overall performance of 0.0804, precision of 0.2780, recall of 0.0885 and granularity of 2.18 in the formal evaluation.

Document level performance before and after Detailed Analysis						
	Top 10 Candidate Documents			After Detailed Analysis		
Obfuscation	Precision	Recall	F1	Precision	Recall	F1
Entire corpus	0.1313	0.5596	0.195	0.3316	0.2827	0.3052
None	0.1807	0.7280	0.2895	0.6808	0.7280	0.7036
Low	0.1642	0.6890	0.2652	0.6547	0.5803	0.6153
High	0.1091	0.5223	0.1805	0.0643	0.0422	0.0510
Simulated	0.2648	0.1675	0.2052	0.5361	0.0859	0.1481

Table 3. Document level performance using top 10 candidate documents before and after detailed analysis on PAN-PC-11 test corpus

Our analysis shows that the test corpus [7] contains 555 (10%) documents plagiarised with translated, 105 (1.89%) with simulated, 2404 (43.33%) with high, 2369 (42.70%) with low and 114 (2.05%) with none obfuscation.

Candidate Document Selection (Section 2.2): Table 3 shows the document level results on the test corpus for the top 10 candidate documents before and after applying the detailed analysis stage. Performance is shown both for the entire corpus and for each type of obfuscation. After the candidate document selection stage the recall score for the entire corpus is 0.5596 (for the top 10 candidate documents). This figure drops to 0.2827 after the detailed analysis stage. The drop in recall varies by obfuscation type. There is no reduction for the none obfuscation but the difference increases for more obfuscated texts. Large reductions in recall are observed for the high and simulated types.

Detailed Analysis Stage (Section 2.3): The GST algorithm is best suited to detect verbatim copy and fails to detect rewritten text. However, a major portion of test corpus is composed of low, high and simulated obfuscations. Therefore, the sequence alignment algorithm did not work correctly to align a pair of suspicious and candidate document. The poor performance of the GST algorithm for these types of obfuscation explains why the recall decreases so severely for these types after the detailed analysis stage.

3.3 Sources of Error

Detailed Analysis Stage (Section 2.3): The system's performance for this stage is worst. Several parameters α_{length}, α_{merge} and $\alpha_{discard}$ and mml were set using a small set of documents due to processing limitations. Using a larger corpus may lead to more suitable values for these parameters being identified. In addition, GST fails to align rewritten text (cases of simulated, low and high obfuscations). If the algorithm is adapted to identify rewritten text then it will also improve the overall performance.

4 Conclusion

This paper described the University of Sheffield entry to the 3rd international PAN plagiarism detection competition which attempted to identify monolingual external plagiarism. Our system did not attempt to identify translation/multilingual plagiarism. A three

stage approach was used: preprocessing and indexing, candidate selection and detailed analysis. The proposed approach achieved an overall performance of 0.0804, precision of 0.2780, recall of 0.0885 and granularity of 2.18 in the formal evaluation in which monolingual and multilingual external plagiarism cases were evaluated together.

The main source of error occurred in the detailed analysis stage. The approach used did not perform well for the low, high and simulated classes of obfuscation. In future, we plan to adapt the GST algorithm to identify correspondences between texts. However, care must be taken to ensure that the algorithm does not become too complex to be applied to the large amounts of data in the PAN corpus.

Acknowledgements

Rao Muhammad Adeel Nawab thanks the COMSATS Institute of Information Technology, Islamabad, Pakistan for funding this work under the Faculty Development Program (FDP).

References

1. Boisvert, R., Irwin, M.: Plagiarism on the rise. In: Communications of the ACM. vol. 49, pp. 23–24 (2006)
2. Loper, E., Bird, S.: NLTK: The Natural Language ToolKit. In: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics. pp. 63–70 (2002)
3. McCabe, D., Butterfield, K., Trevino, L.: Academic Dishonesty in Graduate Business Programs: Prevalence, Causes, and Proposed Action. Academy of Management Learning and Education 5(3), 1–294 (2006)
4. Muller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF - Experimental Evaluation of Visual Information Retrieval. Springer (2010)
5. Nawab, R., Stevenson, M., Clough, P.: University of Sheffield. In: Proceedings of the 4th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse. Lab Report for PAN at CLEF 2010 (2010)
6. Ounis, I., Amati, G., V., P., He, B., Macdonald, C., Johnson: Terrier Information Retrieval Platform. In: Proceedings of the 27th European Conference on IR Research. pp. 517–519. Springer (2005)
7. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). pp. 997–1005 (2010)
8. Potthast, M., Stein, B., Eiselt, A., Cedeño, A., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Proceedings of the CLEFŠ10 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. Padua, Italy (2010)
9. Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E.: 3rd PAN Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. In: 25th Annual Conference of the Spanish Society for Natural Language Processing (SEPLN). pp. 1–77 (2009)