External Plagiarism Detection: N-Gram Approach using Named Entity Recognizer

Lab Report for PAN at CLEF 2010

Gupta Parth, Rao Sameeer, and Prasenjit Majumdar

DA-IICT, Gandhinagar, India.
gupta_parth@daiict.ac.in, rao_sameer@daiict.ac.in, p_majumder@daiict.ac.in

Abstract We tried Named Entity features of source documents to identify its suspicious counter part. A three stage identification method was adopted to understand the impact of NEs in plagiarism. Results along with a brief analysis are given in this note.

1 Introduction

Plagiarism detection in the text documents is a challenging retrieval task. This is the first time; CLEF is offering this task of Plagiarism Detection. We have developed the system for the external plagiarism detection, in which plagiarized chunks need to be found from given large source documents collection. This is our first participation in CLEF.

2 External Plagiarism Detection

2.1 Approach:

Corpus for the task is PAN-PC-10 [3]. In our hybrid model the central idea was to identify n-gram overlaps between two documents, suspicious and source. A set of "suspected" n-grams was used to query the source doc database. Top 5 source docs thus obtained were considered as the most likely source of plagiarism. Our algorithm can be divided into 3 stages:

Creation of suspicious queries: The suspicious docs are tagged with Lingpipe NE Tagger [1]. Then, we take non-overlapping n-grams (n=9) which contain at least one Named Entity (NE). The hypothesis behind is "Information lies in and around NE".

Find Candidate Documents: The source docs are indexed using Indri [2]. Now the suspicious document with Named Entity n-grams is passed as a query to this index and the top 5 relevant source docs are obtained.

Detection Algorithm: We compared overlapping n-grams (n=7) of suspicious docs to those of source docs. If they match, were marked as plagiarized chunks. Then all those chunks less than 500 chars apart were merged.

3 Evaluation

For comparing n-grams overlaps exact match is used, so we doubt the performance of the algorithm in simulated and translated cases, though it identifies plagiarism cases like POS reserving reorder, semantic word replacement and random text operations. But quality of detection deteriorates as obfuscation increases. Hence we paid that toll in precision. Our detection results are depicted below:

Rank: 14 Plagdet Score: 0.2034 Recall: 0.1446 Precion: 0.4983 Granularity:1.1465

The one of the reasons behind low recall is, we considered only top 5 candidate docs while we found in annotations that many of the docs are plagiarized from more than 5 or even more than 10 docs. Hence, experiments with 500 suspicious documents considering top 50 candidate documents were performed and corresponding results are depicted below:

Plagdet Score: 0.2356 Recall: 0.1792 Precion: 0.4830 Granularity: 1.1576

4 Conclusion

Using NE as a feature helps to identify candidate documents. This encourages, to increase the recall by considering a few other features along with NE. Analysis of the actual annotations proves that the parameter tuning is an important aspect which we will reinvestigate. Also we need to consider more candidate documents.

References

1. Alias-i. 2008. LingPipe 4.0.0. http://alias-i.com/lingpipe
2. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: *Indri: a language-model based search engine for complex queries*. Technical report, University of Massachusetts (2005)
3. Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. *Overview of the 2nd International Benchmarking Workshop on Plagiarism Detection*. In Benno Stein, Paolo Rosso, Efstathios Stamatatos, and Moshe Koppel, editors, Proceedings of PAN at CLEF 2010: Uncovering Plagiarism, Authorship, and Social Software Misuse, September 2010.
4. Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso, *Overview of the 1st International Competition on Plagiarism Detection*, Proceedings of PAN'09.
5. Cristian Grozea, Christian Gehl, and Marius Popescu, *ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection*, Proceedings of PAN'09.
6. Jan Kasprzak, Michal Brandejs, and Miroslav Kipa, *Finding Plagiarism by Evaluating Document Similarities*, Proceedings of PAN'09.