Plagiarism detection using information retrieval and similarity measures based on image processing techniques

Marta R. Costa-jussà, Rafael E. Banchs, Jens Grivolla and Joan Codina

Barcelona Media – Innovation Center
Av Diagonal 177, 9th floor, 08018 Barcelona
{marta.ruiz, rafael.banchs, jens.grivolla,
joan.codina}@barcelonamedia.org

Abstract. This paper describes the Barcelona Media Innovation Center participation in the 2nd International Competition on Plagiarism Detection. Particularly, our system focused on the external plagiarism detection task, which assumes the source documents are available. We present a two-step a approach.
In the first step of our method, we build an information retrieval system based on Solr/Lucene, segmenting both suspicious and source documents into smaller texts. We perform a search based on bag-of-words which provides a first selection of potentially plagiarized texts.
In the second step, each promising pair is further investigated. We implemented a sliding window approach that computes cosine distances between overlapping text segments from both the source and suspicious documents on a pair wise basis. As a result, a similarity matrix between text segments is obtained, which is smoothed by means of low-pass 2-D filtering. From the smoothed similarity matrix, plagiarized segments are identified by using image processing techniques. Our results were placed in the middle of the official ranking, which considered together two types of plagiarism: intrinsic and external.

1 Introduction

Plagiarism can be defined as *use or close imitation of the language and thoughts of another author and the representation of them as one's own original work*[1]. The challenge of automatic identification of plagiarism can be divided into: external and internal plagiarism detection. In the case of external plagiarism a corpus of potential source documents is available. We focus exclusively on the external plagiarism. This type of plagiarism has been investigated in many works, and some of the most recent research can be found in the context of the 1st International Competition on Plagiarism Detection [1]. Additionally, several commercial systems can be found in the web [2]. This paper reports our first experience with the plagiarism challenge. We build a two-step plagiarism detection system which uses information retrieval and similarity measures techniques.

[1] http://en.wikipedia.org/wiki/Plagiarism

We found the computational cost of finding plagiarism within a large corpus of documents to be one of the major difficulties in this task. It is therefore important to efficiently preselect potential plagiarism candidates ahead of the more costly in-depth matching. We have not been able to apply the full processing to all documents in the test collection in the available time, even though we did a rather strict selection of potentially plagiarized sections in our first processing step (at the expense of lowered recall). Our ability to optimize various parameters was also hindered by the computational difficulty of executing representative test runs.

The rest of the paper is structured as follows. Section 2 describes the plagiarism detection procedure which included two steps and postprocessing. Section 3 reports results for each step of the procedure in the test collection in terms of recall, precision and granularity, as well as the results of the official evaluation. Finally, section 4 concludes.

2 Plagiarism detection procedure

In this section we describe in detail our plagiarism detection procedure. It consists of two steps and postprocessing. Parameters were adjusted through informal preliminary experiments on small subsets of the available data. In the future, we will have to perform further experiments to adjust the parameters for all the steps in a more systematic way.

2.1 1st step

We segment both the source and suspicious documents in sub-documents of 100 words, with an overlap of 50%. We choose 100 words because the shortest plagiarism contains a minimum of 50 words. All sub-documents are lowercased, stemmed and tokenized. Stopwords are removed, considering as stopwords the 80 more frequents in the entire collection. We index both source and suspicious documents into a single index (in $\sim$9 hours[2]) and the resulting test collection contains $\sim$18M documents.

All suspicious documents are used as queries. We consider short (less than 1000 lines) and long (more than 1000 lines) documents. We perform a query using the N most "interesting" terms and force a matching of 35%. Interesting terms are those that have a higher *tf-idf* score. The N is set to 30 for short documents and to 20 for long documents. The N choice is limited due to time constraints. It is not feasible to perform a query taking all the terms into account, as each query would be very expensive computationally. The choice of 20 and 30 allows a trade-off between the detection of possible plagiarism candidates and computational cost. For each document segment used as a query, the top ranked match is considered as a plagiarism candidate. This search is illustrated in Figure 1.

The information retrieval system was implemented by using Solr, which is an open-source search server based on the Apache-Lucene search library[3]. In the test collection, this phase reports 175k possible cases of plagiarism out of $\sim$10M suspicious fragments of 100 words. Consecutive suspicious fragments coming from consecutive source fragments count as a single plagiarism. This first step was computed in less than two weeks.

[2] All experiments were processed on a single machine: Intel(R) Xeon(R) CPU E335 2GHz, 1 processor, 4 cpu cores, 4M cache and 28G RAM.

[3] http://lucene.apache.org/solr/tutorial.html

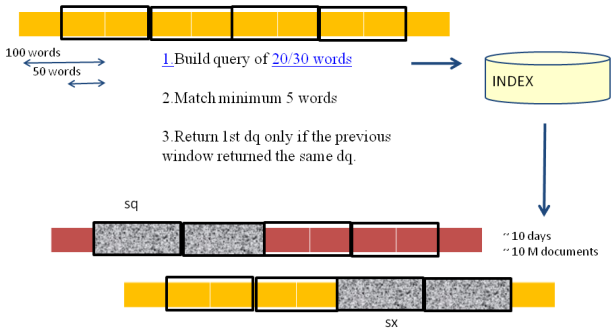

Fig. 1. First step procedure: search.

2.2 2nd step

The starting point of this 2nd step is the set of plagiarism candidates from the 1st step.

This 2nd step is similar to the dotplot technique [3]. For each pair of candidates, we implemented a sliding window approach that computes cosine distances between overlapping text segments from both the source and suspicious documents in a pair wise basis, using a sliding window of 50 words. As a result, a similarity matrix between text segments is obtained. If we represent this similarity matrix as an image, it results in a very noisy image. That is why we smooth it by using a low-pass filter, in particular a 2-dimensional hamming window of $5x5$. Then, we choose a threshold above which we consider the text may be plagiarized.

For each fragment above this threshold, we extract the begining and the end positions (see Figure 2). Finally, we compute a measure based on word matching: $WM = \frac{N_{coincidences}}{\sqrt{N_{source}N_{suspicious}}}$ where $N_{coincidences}$ is the number of words equal in the source and suspicious text fragments; N_{source} is the number of words in the source text fragment and $N_{suspicious}$ is the number of words in the suspicious text fragment. The final candidates are reported as detected plagiarisms only if the WM is over 0.3.

This phase is computationally expensive. The 175k cases of plagiarism reported from first step are computed in 4 weeks using 4 parallel processes on our machine.

2.3 Postprocessing

Finally, we observed that our plagiarism detection reported overlapped plagiarism sections from the same source document. Therefore, we performed a postprocessing which compacted all overlapped sections and additionally, all sections which were separated less than 5000 characters. This step is intended to approach granularity to one. There is a tradeoff between precision and granularity, and the postprocessing parameters would need to be adjusted depending on the specific objective measure.

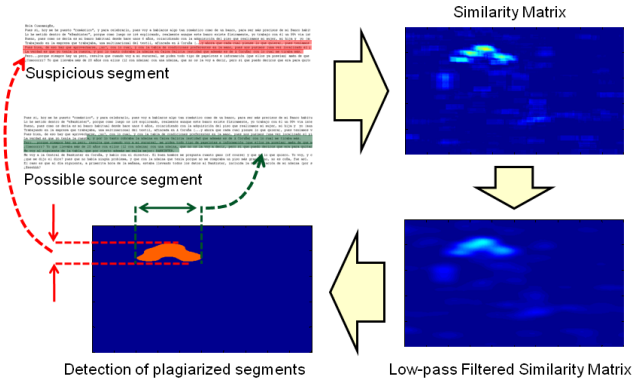

Fig. 2. Second step procedure

3 Experiments

We initially performed experiments with the plagiarism collection proposed in the 1st International Competition on Plagiarism Detection [1]. We used small subsets of this collection to conduct a trial-and-error training of the different parameters for the first and second processing step.

Here we report results using only the external test part of this year's plagiarism collection (as described in [4]). Table 1 reports the results when using each phase of the process, as well as the official evaluation results.

System	Overall	Recall	Precision	Granularity
First step	0.2268	0.4633	0.1523	1.0149
Second step	0.0905	0.0499	0.4850	1.0
Official submission	0.2130 (12)	0.2987 (9)	0.1682 (17)	1.0142 (5)

Table 1. Results of first and second step (both include the same postprocessing), as well as the Official Evaluation Results (including intrinsic plagiarism). The relative position in the overall ranking is in parentheses.

Regarding the first step, it would be interesting to test if retrieving more than one document could improve the recall. Additionally, notice that external plagiarisms were created by using the following offuscations strategies: random text operations, semantic word variation, POS-preserving word shuffling and translation [4]. Our procedure uses bag-of-words, which implies that is not able to detect offuscations based on semantic word variation or translation. Regarding the second step, there are many parameters that were optimized informally using only very small subsets of the training data, which could explain such a low performance. While precision increases as expected, recall drops dramatically. Using an optimized set of parameters it should be possible to increase precision while avoiding a great loss regarding recall.

We were not able to finish the task on time for the official submission. We were able to process 100% of the documents with the first step, but only 26% of the documents with the second step. Therefore, our submission consisted of the second step output when available, and the first step output otherwise. In all cases, we performed the post-process. Our results were ranked in the 12th position out of 18. Notice that the lowest measure is the precision. The low score in precision is due to the fact that we were not able to finish the second step of our plagiarism detection procedure on time.

4 Conclusions

This paper presented a two-step system for plagiarism detection based on information retrieval and sentence similarity. The first step uses a standard information retrieval system based on bag-of-words. The second step computes a similarity matrix among the first step candidates and it uses image processing techniques to filter out non-plagiarized texts. We report results ranked towards the middle of the field of participants in the 2nd International Competition on Plagiarism Detection.

This work is our first experience in plagiarism detection. There are many aspects of our methodology that can be further investigated. Specially, we should focus on improving the recall, for example, by using part-of-speech in the first step. Reducing the second step's computational complexity would allow to deal with a larger number of plagiarism candidates and thus be less restrictive in the initial selection. Optimization of all parameters should be done systematically and their effect on the different performance measures analyzed in more detail.

Acknowledgements

This work has been partially funded by the Spanish Department of Education and Science through the *Juan de la Cierva* fellowship program and the Spanish Government under the BUCEADOR project (TEC2009-14094-C04-01). The authors also want to thank the Barcelona Media Innovation Centre for its support and permission to publish this research.

References

1. Potthast, M., Stein, B., Eiselt, A., no, A.B.C., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), CEUR-WS.org (September 2009) 1–9
2. Maurer, H., Kappe, F., Zaka, B.: Plagiarism - a Survey . Journal of Universal Computer Science **12**(8) (2006) 1050–1084
3. Helfman, J.: Dotplot: a program for exploring self-similarity in millions of lines of text and code. Journal of Computational and Graphical Statistics **2**(2) (1993) 153–174
4. 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) (to appear), Beijing, China, Association for Computational Linguistics (August 2010)