Source Retrieval Based on Learning to Rank and Text Alignment Based on Plagiarism Type Recognition for Plagiarism Detection

Notebook for PAN at CLEF 2014

Kong Leilei[1,2], Han Yong[1], Han Zhongyuan[1,3],

Yu Haihao[1], Wang Qibo[1], Zhang Tinglei[1], Qi Haoliang[1]

[1]Heilongjiang Institute of Technology, China
[2]Harbin Engineering University, China
[3]Harbin Institute of Technology, China

kongleilei1979@gmail.com

Abstract. This paper regards the query keywords selection problem in source retrieval as learning a ranking model to choose the method of keywords extraction over suspicious document segments. Four basic methods are used in our ranking function: BM25, TFIDF, TF and EW. Then, a ranking model based on Ranking SVM is proposed to rank the query keywords group which is contributed to get the higher evaluation measure F. In our ranking model, achieving the best performance measure F of source retrieval is used as the target of learning to rank.

In text alignment, a novel method based on the plagiarism type recognition model is proposed. This approach employs the distinct strategies to detect the plagiarism text according the different plagiarism type. The plagiarism type recognition model is based on logical regression model. The experimental results on PAN 2014 plagiarism detection corpus indicate the efficiency of the proposed methods.

Keywords: plagiarism detection; source retrieval; text alignment, ranking model; plagiarism type recognition

1 Source Retrieval Based on Learning to Rank

For the task of source retrieval, the target is to retrieve all plagiarized sources while minimizing retrieval costs. It has become standard for plagiarism detection to retrieve plagiarism sources with query keywords selected from suspicious document. Extracting key phrases from a text document to be used as queries is a challenging problem in the source retrieval task for the plagiarism detection. In this year's work, we focus on the task of selecting query keywords for source retrieval.

There have been many efforts toward keywords extraction for text domain. In contrast, there is less work on query keywords extraction for source retrieval in plagiarism detection. The methods based on machine learning are still little used.

This year, we aim at improving the evaluation measure F. Our method regards the keywords extraction as a ranking problem for improving the evaluation measure F which is defined in [1]. A ranking model is learned to rank the keywords which are selected by various keywords extraction methods. The ranking model is used to decide which group of the keywords helps greatly to improve measure F. At the same time, the ranking model can incorporate more features of query keywords to describe the keywords in more aspects.

The train cases for learning the ranking model are constructed by using the corpus of PAN@CLEF2012 detailed comparison task [2]. A ranking model based on Ranking SVM [3] is trained to selected the better query keywords group which extracted by some keywords extraction methods. The better query keywords group means that they are more conducive for getting the higher performance measure F. The basic candidate keywords extraction methods include BM25, TFIDF, TF and EW [4]. The Ranking SVM is used as the learning to rank algorism. Some statistic features, such as TF, TFIDF, BM25, is used to describe the keywords.

During the test period, the suspicious document is first partitioned into text segments that are made up of 5 sentences. Then, we use the basic keywords extraction methods to select query keywords. Furthermore, the features of each query keyword which selected by different basic keywords extraction methods are computed. Lastly, the keywords ranking model is used to choose the better query keywords group for a text segment of suspicious document.

In the procedure of source retrieval, we used the ChatNoir [5] search engine API. Queries are constructed by combining each non-overlapping k keywords which selected by ranking model, where k = 10, in order to create a set of queries for each segment. Only the top 3 results are downloaded. Then, each segment is regarded as a query and retrieved in the index which constructed by all the downloaded documents. Each top 1 result is reported as the final result. In addition, a voting method which needs no downloading documents is used in our method. If k queries retrieved the same result, that result will be regarded as the final result either. We set k=6.

The test result on the source retrieval test corpus2 is shown in Table 1.

Table 1. Evaluations on pan14-source-retrieval-test-corpus2-2014-05-14

Total workload	Queries	83.5
Time to 1st Detection	Queries	85.7
	Downloads	24.9
Retrieved Sources	Precision	0.07568
	Recall	0.48203

2 Text Alignment Based on Plagiarism Type Recognition

The objective of text alignment of plagiarism detection is searching the plagiarism suspicious fragment in suspicious document together with its source.

The plagiarism can be divided into many categories according the different plagiarism means [6]. However, the existing methods do not distinguish the plagiarism types and they detect the plagiarism cases which belong to the different plagiarism types by using the same method, which result in the difficulty of finding a balance among the different plagiarism type. If we can identify the plagiarism types before we align the plagiarism text, we can deal the different plagiarism types with different methods to improve the performance.

From this perspective, we proposed a novel method based on plagiarism type recognition for this year's text alignment task. During the training period, the golden standards of detailed comparison training corpus of PAN@CLEF 2012 are used as the training corpus to train the Plagiarism Type Recognition Model. This model is based on Logistic Regression model. The plagiarism types are grouped into two categories: obfuscation and no-obfuscation. The main lexical features of plagiarism text include Dice Coefficient, Jaro Distance, Jaccard Coefficient, Levenshtein Distance, Manhattan Distance and Ngram Distance.

During the test period, the suspicious document and the source document are compared to take the original plagiarism fragments by the method we developed for PAN 2013 [7], and then the Plagiarism Type Recognition Model is used to recognize the plagiarism types. Finally, the pair of suspicious and source document is compared again by the method we proposed in the text alignment task in [7]. The only difference is that the parameters are revised according to the different plagiarism types.

Table 2 shows the results of PAN@CLEF2014 Text Alignment subtask on test corpus 2 and test corpus 3.

Table 2. Evaluations on pan14-text-alignment-test-corpus

Sub-Corpus	Plagdet Score	Recall	Precision	Granularity
pan14-text-alignment-test-corpus2-2014-05-09	0.82161	0.80746	0.84006	1.00309
pan14-text-alignment-test-corpus3-2014-05-14	0.83514	0.84156	0.82882	1.00000

3 Conclusions

In this paper, we describe the approaches we used in the subtask of Source Retrieval and Text Alignment for PAN@CLEF 2014.

In the sub-task of Source Retrieval, we applied a method based on learning to rank. We design a model based on Ranking SVM to select the keywords groups which extracted by the different keywords extraction methods that can get a better performance on the evaluation measure F.

In the sub-task of Text Alignment, we designed a method based on Logistic Regression model to identify the different plagiarism types. The text alignment algorisms with different parameters are used to the detailed comparison to detect the plagiarism with various plagiarism types.

We feel this is more of a beginning than an end to develop our two methods. More features and keywords extraction approach will be used in our query keywords extraction ranking model. And the Plagiarism Type Recognition Model will be trained to identify more kinds of plagiarism types.

Acknowledgments

This work is supported by National Natural Science Foundation of China(61370170), National Social Science Fund Project - Youth Project of China(14CTQ032) and Heilongjiang Province Educational Committee Science Foundation(12541649, 12541677).

Remark

This work was done in Heilongjiang Institute of Technology.

Reference

1. Potthast M, Hagen M, Gollub T, et al. Overview of the 5th Overview of the 5th International Competition on. CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September,Valencia, Spain.
2. http://pan.webis.de/
3. Joachims T. Optimizing search engines using clickthrough data. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 133-142.
4. Lee Gillam. Guess Again and See if They Line Up: Surrey's Runs at Plagiarism Detection—Notebook for PAN at CLEF 2013. Working Notes Papers of the CLEF 2013 Evaluation Labs, September 2013.
5. Potthast M, Hagen M, Stein B, et al. ChatNoir: a search engine for the ClueWeb09 corpus. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012: 1004-1004.
6. Alzahrani S M, Salim N, Abraham A. Understanding plagiarism linguistic patterns, textual features, and detection methods. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2012, 42(2): 133-149.
7. Leilei Kong, Haoliang Qi, Cuixia Du, Mingxing Wang, and Zhongyuan Han. Approaches for Source Retrieval and Text Alignment of Plagiarism Detection—Notebook for PAN at CLEF 2013. Working Notes Papers of the CLEF 2013 Evaluation Labs, September 2013.