Approaches for Candidate Document Retrieval and Detailed Comparison of Plagiarism Detection

Notebook for PAN at CLEF 2012

Kong Leilei1, Qi Haoliang1, Wang Shuai1, Du Cuixia2,Wang Suhong2 and Han Yong1

[1]Heilongjiang Institute of Technology, China
[2]Harbin Engineering University, China
sevenkll@hotmail.com

Abstract. In this paper we report on our plagiarism detection system which is used to process the PAN plagiarism corpus for the tasks of Candidate Document Retrieval and Detailed Comparison. To retrieve the plagiarism candidate document by using ChatNoir API, a method based on tf*idf to extract the keywords of suspicious documents as queries is proposed. An Lucene ranking method is used for plagiarism candidate document reduction. And a detailed comparison algorithm to get the web pages that are actually sources for plagiarized passages is applied. To extract all plagiarism passages from the suspicious document and their corresponding source passages from the source document, a plagiarism detection method combined with semantic similarity and structure similarity is proposed. Semantic similarity is calculated by Vector Space Model while structure similarity is calculated by our own method. We use information retrieval to get candidate pairs of sentences from suspicious document and potential source document. A method which is called Bilateral Alternating Sorting is applied to merge pairs of sentences. Those plagiarism candidate result pairs are screened in post-processing.

Keywords: plagiarism detection, tf*idf, semantic similarity, structure similarity, Vector Space Model

1 Introduction

The rapid development of network technology, including large numbers of search engines, document repositories, translation software systems, not only provides people with the various knowledge acquisition channel, but also opens the door for text plagiarism. Plagiarism generally refers to the illegitimate use of someone else's information, text, ideas, etc. without proper reference to the original source of these borrowings [1]. Plagiarism and its automatic retrieval have attracted considerable attention from research and industry: various papers have been published on the topic, and many commercial software systems are being developed [2]. It becomes more important to determine the originality of the text. The research on copying text recognition has become an urgent need to address the problem.

In recent years, many well-known organizations carried out evaluation, international competitions and conferences focused on the plagiarism detection. And PAN[3] is one of them. PAN@CLEF [3]offers a controlled evaluation environment to evaluation the algorithm or system for plagiarism detection. This year, we focused on the plagiarism detection evaluation of PAN@CLEF2012, which included two sub-tasks: Candidate Document Retrieval and Detailed Comparison. We spent six months to research the problems and submitted our results of the two sub-tasks. Our team obtained the first place for the Detailed Comparison sub-task.

In this paper, we introduce a method for Candidate Document Retrieval and Detailed Comparison sub-tasks. In the Candidate Document Retrieval sub-task, a method based on tf*idf to extract the keywords of suspicious documents as queries was proposed. A scoring method was used to plagiarism candidate document ranking. In the Detailed Comparison sub-task, a plagiarism detection method based on Vector Space Model(VSM) and Overlapping Measure Model at the sentence level was presented. Bilateral Alternating Sorting was designed to merge the pairs of plagiarism sentences, and

those plagiarism candidate result pairs were screened in post-processing.

The rest of the paper is organized as follows. Section 2 describes the overview of related work. Then, Section 3 and Section 4 describe the method for Candidate Document Retrieval sub-task and Detailed Comparison sub-task while Section 5 includes the evaluation results. Finally, Section 6 discusses the main points of this study and proposes future work directions.

2 Overview of Related Work

Most often, copy is an open copying, and plagiarism mainly refers to plagiarize other's language, charts, formulas or research ideas and then edited, pieced together, modified and added to their own papers, writings, project applications, data files, computer code material, and so on, and set him/herself up as the author. The research of formal language text plagiarism starts earlier for its strict formal syntax, clear semantics of expression, easy analysis and processing. Since Ottenstein [4] put forward attribute counting to detect the program copying, there emerged a lot of the formalization text copying recognition system. Natural language text has no formal syntax constraints and its semantics has the ambiguity, so it is more difficult to carry out plagiarism identification. The research on natural language text copy detection began in the 1990s, and has made great progress since Richard used keyword matching algorithm to develop the WordCheck[5].

The core problem of text plagiarism detection is to determine whether the plagiarism exists and how to measure the similarity degree. For text similarity problem, many researchers have put forth some effective detection methods. It mainly includes (1) Similarity calculation method based on statistics, such as [6] and [7]. It needs the support of large-scale corpus, the long training process and has some limitations. (2) Similarity calculation method based on semantic comprehension: it neither needs the support of large-scale corpus nor the long training process. It has a high precision but mostly is limited in the scope of words or sentences. Specific methods includes similarity calculation with Wordnet [8], similarity calculation with TongYiCi CiLin [9], similarity calculation with semantic sequence kernel [10], etc.

Finger Printing and Word Frequency are the mainly method to recognize the plagiarism. Finger Printing is fast, simple and effective, suitable for large-scale computing. Word Frequency method first statistics the number of each word in the document to constitute the feature vector of the document, then use vector dot product, cosine law, the correlative frequency model, etc. to measure the similarity of two documents. Word Frequency statistical method has a high precision but its speed in not fast than Finger Printing technology.

Although a lot of plagiarism detection systems are better to complete the simple text copy detection, detection for English has also made some achievements, there are also some questions and works which have not received much attention yet.

First, plagiarism detection filed has not effective techniques to filter the plagiarism source in massive data corpus. The corpus of reference documentation is limited to the data of several G in the existing plagiarism detection system. However, over time, reference documentation set of documents to be detected grows increasingly large and source document set is not limited to be a few G magnitude data. Massive data processing and the growing amount of data increase the difficulty of plagiarism detection. Existing methods are powerless in the face of dealing with in the data of several T. Timely and effective in a limited time to find suspicious from the source document to be detected is the key to the establishment of effective plagiarism detection system.

Second, the performance of existing systems which are based on matching and statistics techniques is not satisfactory for the practical application. They still have the wrong check, leakage check, especially non-straightforward copy which has a low distinction degree, and has difficult to achieve

accurate identification of plagiarism. And they can not deal with the complex text plagiarism, especially for two articles which have the same meaning and converting the writing method.

3 Candidate Document Retrieval

Given a suspicious document and a web search engine, the task is to retrieve a set of candidate source documents that may have served as an original to plagiarize from[3]. To retrieve the plagiarism candidate document by using ChatNoir API, we apply a method based on tf*idf to extract the keywords of suspicious documents as queries. When using ChatNoir API gets copy source of suspicious document, we use an improved Lucene[23] scoring method to reduce the plagiarism candidate document. Finally, a detailed comparison algorithm to get the web pages that are actually sources for plagiarized passages is applied. The detailed method is described as follows.

3.1 Getting Query

First, each suspicious document s is preprocessed, including stemming, removing stop words and replacing figures. The queries for every s are coming from the top queryGroup*queryLength terms which are the top n sorted by tf*idf values from high to low in each paragraph of s, where queryGroup is the group number of queries and queryLength is the term number of each query group. In testing phase, the queries we used is 2*5.

3.2 Retrieving

ChatNoir API is applied to retrieve the plagiarism candidate document for each query group. Then, put retrieved top n results in the result set of plagiarism candidate document and use ChatNoir API to download them, where n=10.

3.3 Getting Sources for Plagiarized Passages

The plagiarism candidate document is preprocessed into non-overlapping plaintext passages. We index them and use query of 2*5 for each passage of each suspicious document to retrieve by using an Lucene scoring method for filtering the plagiarism candidate in [11].

And we take top n of retrieving results of each query as results. Last, we use an algorithm to get the web pages that are actually sources for plagiarized passages which will be described in following PAN Detailed Comparison Task.

4 Detailed Comparisons

Given a pair of suspicious document and potential source document, the task is to extract all plagiarized passages from the suspicious document and their corresponding source passages from the source document [3]. Firstly, the suspicious documents and plagiarism candidate source documents are pre-processed. We apply a plagiarism detection method combined with semantic similarity and structure similarity to extract all plagiarism passages from the suspicious document and their corresponding source passages from the source document. Semantic similarity is calculated by Vector Space Model while structure similarity is calculated by an Overlapping Measure Model which will be described as follows. We use a method of information retrieval to get candidate pairs of sentences from suspicious document and potential source document and a merge algorithm which is called Bilateral

Alternating Sorting is applied to merge pairs of sentences. Finally, those plagiarism candidate result pairs are screened in post-processing. This method is described in detail in the following parts.

4.1 Pre-processing

In the pre-processing part, the suspicious documents and plagiarism candidate source documents will be processed in some ways, including removal of special characters and whitespace, case transformation, removal of stopwords and stemming.

4.2 Detailed Comparison

Since the passage is the smallest unit that an author expresses an independent and complete view and the sentence is the basic structure of one passage, we choose sentences as the chunks. The following is the steps of processing the candidate plagiarism passages detection.

Step1: Suspicious documents and source documents are divided according to the sentence. After that we index all the sentences in the source documents. Each sentence in the suspicious documents will be retrieved in the index as a query. This kind of process is called sentence similarity retrieval. We regard suspicious passage S and reference passage R in source document as pairs of plagiarism candidate sentence which their cosine distance is greater than t1 to get semantic similarity, as shown in formula 1:

$$Sim(\mathrm{S},\mathrm{R}) = \cos\theta = \frac{\sum_{k=1}^{n} w_{Sk} * w_{Rk}}{\sqrt{(\sum_{k=1}^{n} w_{Sk}{}^{2})(\sum_{k=1}^{n} w_{Rk}{}^{2})}} > t1 \quad (1)$$

where Sim(S,R) is the similarity degree of S and R, Θ is document vector angel, WSk and WRk are the weight of S and R respectively,t1 is threshold. We used t1=0.42.

Step2: We screen plagiarism candidate sentence to get structure similarity by using formula 2.

$$T = \frac{2 * \sum_{t \in I_S \cap I_R} Min(\mathrm{N}_{I_S}(t), \mathrm{N}_{I_R}(t))}{|I_S| + |I_R|} > t2 \quad (2)$$

where NIs(t) and NIR(t) are the number of the terms which are overlapping in the suspicious sentence and reference sentence, Min(NIs(t),NIR(t)) is the smallest one of NIs(t) and NIR(t) , t2 is the threshold. We used t2=0.32.

Those sentence pairs which are not only in line with formula 1 but also formula 2 will be regarded as the plagiarism candidate sentence pairs.

Step 3: Merge the scattered plagiarism candidate sentence pairs which are got by above process method. This process is the recovery of a complete plagiarism case. We design a Bilateral Alternating Sorting algorithm to merge the suspicious sentence and reference sentence which guarantee the suspicious sentence and reference sentence are adjacent. Because of patent pending, this method is inconvenient stated here. The passage pairs after merging are called the candidate result pairs.

4.3 Post-processing

We use formula 2 to screen candidate result pairs which perhaps the non suspicious passage pairs in post-processing phase .We used t3=0.30 this time.

5 Results

The results of Candidate Document Retrieval sub-task and the Detailed Comparison sub-task are summarized in Table 1 and Table 2.

Table 1 Results of Candidate Document sub-task

Team	Reported Sources		Downloaded Sources		Retrieved Sources	
	Precision	Recall	Precision	Recall	Precision	Recall
Gillam et al. University of Surrey, UK	0.6266	0.2493	0.0182	0.5567	0.0182	0.5567
Jayapal University of Sheffield, UK	0.6582	0.2775	0.0709	0.4342	0.0698	0.4342
Kong Leilei Heilongjiang Institute of Technology, China	0.5720	0.2351	0.0178	0.3742	0.0141	0.3788
Palkovskii et al. Zhytomyr State University, Ukraine	0.4349	0.1203	0.0025	0.2133	0.0024	0.2133
Suchomel et al. Masaryk University, Czech Republic	0.5177	0.2087	0.0813	0.3513	0.0094	0.4519

Table 2 Results of Detailed Comparison sub-task

Detailed Comparison Task						
Rank	Team	PlagDet	Precision	Recall	Granularity	Runtime*[Seconds/Pair]
1	Kong Leilei Heilongjiang Institute of Technology, China	0.7386159	0.8249708	0.6782238	1.0109503	5.9187108
2	Kasprzak et al. Masaryk University, Czech Republic	0.6826726	0.8931670	0.5524708	1.0000000	5.3679195
3	Grozea et al. Fraunhofer Institute FIRST, Germany	0.6787810	0.7747815	0.6351092	1.0396952	4.8279920
4	Oberreuter Universidad de Chile, Chile	0.6735574	0.8673093	0.5553130	1.0073026	2.5899274
5	Rodríguez Torrejón et al. Universidad de Huelva, Spain	0.6252024	0.8344227	0.5004208	1.0009596	0.1900923

6 Conclusions

Our method is evaluated by PAN2012@CLEF and compared with the other plagiarism detection systems. The evaluation results of our method in the competition were excellent. With the PAN-09, PAN-10, PAN-11 and PAN-12 corpora, our method showed a great advantage and produced a high performance. Results show that our system's overall performance, especially the recall is higher than most of the other methods for most kinds of plagiarism cases. The plagiarism detection method we proposed is flexible and scalable, the time limit is reasonable. In our case we only needed one mainstream server to run the complex plagiarism detection system. Furthermore, we will aim to determine the threshold boundary more reasonably. The synonym replacement and the translation would also be introduced into the plagiarism system. We will work on a better Candidate Document Retrieval method.

Acknowledgements. This work is supported by NSF of China(60970057,60903083).

Remark: This work was done in Heilongjiang Institute of Technology.

Reference

1. Maurer, H., Kappe, F., Zaka, B.: Plagiarism-A Survey. Journal of Universal Computer Science 12(8), 1050–1084 (2006)
2. Martin Potthast, Benno Stein, Andreas Eiselt , Bauhaus-universität Weimar, Alberto Barrón-cedeño , Paolo Rosso. Overview of the 1st International Competition on Plagiarism Detection. SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), CEUR-WS.org
3. http://pan.webis.de/
4. Ottenstein K J. An Algorithmic Approach to the Detection and Prevention of Plagiarism .ACM SIGCSE Bulletin.1976,8(4):30-41
5. Clough P. Plagiarism in Natural and Programming Languages: an Overview of Current Tools and Technologies. Research Memoranda:CS-00-05, Department of Computer Science, University of Shefield, 2000
6. PAN Qian Hong, WANG Ju, SHI Zhong Zhi .Text Similarity Computing Based on Attribute Theory. Chinese Journal of Computers, 1999, 22(6) : 651-655
7. Song Qin-Bao, Yang Xiang-Rong. A Detection Algorithm for the Illegal Coping and Distributing of Digital Goods. Chinese Journal of Computers, 2002, 25(11):1206-1211
8. Agire E, Rigau G. A Proposal for Word Sense Disambiguation Using Conceptual Distance. International Conference on Recent Advances in Natural Langu age Processing, Velingrad, 1995:258-264
9. Che Wanxiang, Liu Ting, Qin Bing, Li Sheng. Chinese Sentences Similarity Computation Oriented the Searching in Bilingual Sentence Pairs. The Seventh Joint Academic Conference for Conputational Lingguistics. Beijing: Tsinghua University Press.2003:520-526
10. Bao Junpeng, Shen Junyi, Liu Xiaodong, Liu Haiyan, Zhang Xiaodi. Document Copy Detection Based on Kernel Method. Proceedings of 2003 IEEE International Conference on Natural La nguage Processing an d Knowledge Engineering, Beijing, Oct.2003:250-256
11. http://lucene.apache.org/