Authorship Segmentation for Retrieving Source Documents in Plagiarism Detection

Sidik Soleman[1,a] Atsushi Fujii[1]

Abstract: Plagiarism is an act of using another person's words or ideas without giving credit to that person. Plagiarism detection is a task to determine whether an input document was made through plagiarism and present one or more source documents as evidence. Because a plagiarist often uses more than one source document by different authors to produce a single document, it is useful to segment the input document based on the authorship attribution. We propose a text segmentation method that uses a number of authorship attribution features, such as distributions of characters, words, and parts of speech, and divides the input document where the writing style is substantially different. We experimentally show the accuracy of our segmentation method and its effectiveness for retrieving source documents.

1. Introduction

As the Internet has become more prevalent, its misuse has a more significant impact on our society. Plagiarism, an act of using another person's words or ideas without giving credit to that person[*1], is a serious problem in industry and academia. By definition, plagiarism violates intellectual property rights and discourages creativity and innovation in the industry. Plagiarism also invalidates a traditional education methodology, namely school assignments. A survey showed that 36% of undergraduate students have performed paraphrasing or copying plagiarism to cheat on written assignments [5] .

Due to an explosively growing number of documents on the Internet, fully automated or interactive systems to detect plagiarism have recently been in high demand. Automated systems are requested to determine whether an input document has been made through plagiarism or not, while interactive systems are requested to present one or more candidates of source documents facilitating user's decision making.

The above background has motivated researchers and vendors to develop methodologies and commercial systems for the plagiarism detection (PD). However, existing methods and systems for PD are still associated with a large number of false positives, where original work was determined as plagiarism, and false negatives, where plagiarism was found innocent [13]. For an interactive system, false negatives are more problematic because users cannot investigate them. However, the number of false positives should also be small to reduce users' effort to verify whether the detected documents are associated with plagiarism.

Because a plagiarist often use more than one source document to produce a single document, it is important to divide an input document into fragments each of which corresponds to a single source document. Hence, searching for source document using its corresponding fragment could reduce both false negatives and false positives. Whereas existing methods for text segmentation [9] detect changes of the topic, it is also important to detect changes of the authorship, assuming different source documents were written by different authors. In this paper, we propose a method for segmenting document in PD that uses features of authorship attribution.

Examples 1 and 2 illustrate topic and authorship boundaries, respectively. These examples are excerpts from two input documents in a dataset to evaluate PD systems [8]. Example 1 is about family of Barack Obama, and the segments before and after the boundary describe his wife and his childhood, respectively. Thus, the distributions of content words, such as "wife" and "child", are different depending on the segment. However, in Example 2, which is about iron, the use of personal pronouns, such as "I" and "you", is different depending on the segment.

Example 1 for topic boundary:

Described as a real life Claire Huxtable, many observers commend Michelle's ability to juggle motherhood, a galant career as a lawyer, and wife to a junior United States Senator. In order to do this, she and the kids follow a strict schedule that even allocates time for play. While serving as an executive for the University of Chicago Hospital, her children attend a school connected to the facility. Before the 2008 Presidential campaign began, the entire Obama family congregated together at the Trinity United Church of Christ each Sunday. In January 2008, Michelle took a leave of absence from her post to open her date book for a more active schedule campaigning for her husband's presidency.
<boundary>

Extended Family: Maternal Relations

The maternal relations of Barack Obama can be traced back generations. His mother was Ann Dunham who, after separat-

[1] Tokyo Institute of Technology, 2–12–1 Ookayama, Meguro–ku, Tokyo 152–8550, Japan
[a] soleman.s.aa@m.titech.ac.jp

[*1] http://www.merriam-webster.com/dictionary/plagiarism

ing from his father, married Lolo Soetoro. The two of them had another child, his half-sister Maya Soetoro. The maternal grandparents of Barack Obama are Madelyn Dunham and Stanley Dunham. Mother: Stanley Ann Dunham Soetoro

Example 2 for authorship boundary:

Iron For Your Body. Looking so pale? Don't rejoice; You aren't going to turn into an Edward Cullen yet. In fact, were I you, I would be concerned. Paleness is one of the indicators that you might lack in something important: iron.

<boundary>

What is Iron? Iron, one of the most abundant metals on Earth, is essential to most life forms and to normal human physiology. Iron is an integral part of many proteins and enzymes that maintain good health. In humans, iron is an essential component of proteins involved in oxygen transport. It is also essential for the regulation of cell growth and differentiation.

2. Related Work

The task for PD can be formulated as a combination of information retrieval (IR) and natural language processing (NLP). From an IR point of view, the task is to search a document collection for potential source documents with respect to a query document. From an NLP point of view, the task is to align a query and each of the potential source documents to obtain evidence for plagiarism. These tasks share a chicken and egg relationship because the source retrieval is needed to determine the degree to which a query and a candidate of the source documents can be aligned to each other and the text alignment requires the predetermined candidates of source documents.

In the PAN workshop[*2], these two tasks have been explored independently assuming the source retrieval is performed prior to the text alignment. Thus, the source retrieval can be seen as the first step for PD. Because a query document consists of fragments that were copied or paraphrased from more than one source document, the task can be similar to the associative retrieval [11], in which a query document is segmented to search a collection for the documents related to each segment and all the retrieved documents are combined into a single ranked list.

In the source retrieval, Hagen et al. [3] reported that many proposed methods only segment a query document with a predetermined number of words or sentences and did not considered changes of authorship.

Graham et al. [2] proposed a method for authorship-based text segmentation that models changes of the authorship by means of stylistic inconsistencies. However their method has been evaluated with pseudo-plagiarism documents each of which is a concatenation of existing documents, whereas the content of source documents are often edited rather than used as verbatim copies.

3. Retrieving Source Document in PD

For retrieving source documents, we adopt the technique used for associative patent retrieval consisting of segmentation, query generation, document retrieval, and result integration process [11]. Figure 1 illustrates an overview of our adopted method.

First, segmentation process divides query document into several fragments. For this process, we use text segmentation based on authorship attribution. Second, query generation process produces a number of queries from each fragment. A query has length of 10 words due to the limitation of document retrieval module that we use. Top 15 words are selected and sorted based on its word frequency in the current fragment and IDF (Inverted Document Frequency) in a document collection. The queries are 10-gram sequences generated from the sorted top 15 words. We select only top 15 words because we observe that generally these words are the most effective for query. For example, using the text before the boundary in the Example 1 in Section 1, the generated queries are:

- united michelle chicago states obama senator together schedule huxtable galant
- michelle chicago states obama senator together schedule huxtable galant congregated
- chicago states obama senator together schedule huxtable galant congregated malia
- states obama senator together schedule huxtable galant congregated malia allocates
- obama senator together schedule huxtable galant congregated malia allocates lawyer
- senator together schedule huxtable galant congregated malia allocates lawyer date

Third, each query is submitted to the document retrieval module that employs a publicly available search engine[*3] for document collection which is the ClueWeb09[*4] [7]. We use BM25 method for document scoring.

Finally, result integration combines all the documents lists to produce a single document list as follows. First, the document lists of the queries generated from a fragment are merged and re-ranked based on the accumulative document score and top M documents are selected. Therefore, the number of generated document lists becomes equal to the number of fragments. Second, these generated document lists are merged and re-ranked using the same method in the first step and top N documents are selected as the final document list. According to our experiment, the optimal values for M and N are 64 and 300, respectively.

3.1 Text Segmentation Based on Authorship Attribution

We formulate text segmentation as a classification task that determines whether or not each of the sentence boundaries is the boundary where segments from different source documents are concatenated. To perform classification for each sentence boundary, we follow this process:

(1) K sentences before and after the sentence boundary are extracted as a pair of segments. According to our experiment, the optimal value for K is 5.

(2) For each segment, a vector consisting of the values for the authorship attribution feature is produced.

(3) The similarity between two segments is computed as the cosine of the angle between their corresponding vectors.

(4) Classification is performed using these cosine similarity

[*2] http://pan.webis.de/

[*3] http://chatnoir.webis.de

[*4] http://www.lemurproject.org/clueweb09.php/

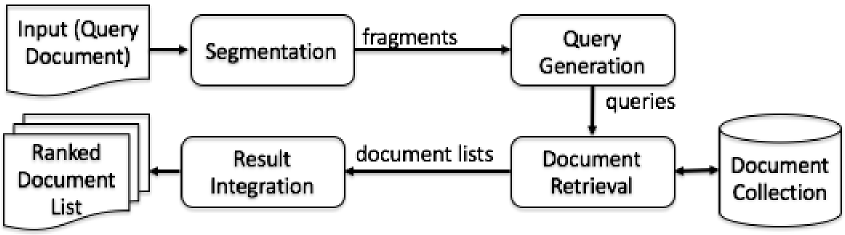

Fig. 1 Overview of the method for retrieving source documents in PD

scores. For the classification, we employ Support Vector Machine (SVM)[*5] using Radial Basis Function as kernel function.

In addition, we explore the following notable authorship attribution features mentioned in [4] [10] for this task:

(1) **Part of Speech Tag (PT)**: This feature compares the distribution of part of speech tag because people may use different POS tag distribution in their texts. We could observe this tendency from the Example 2 in Section 1 where the segment before the boundary uses more personal pronouns than the segment after the boundary. A POS tagger is used to extract this feature [12].

(2) **Noun Phrase Syntax (NS)**: This feature is related to the syntactic of sentence focusing on noun phrase syntax because there are many possible structures of noun phrases. People may have tendency to use certain syntax to construct noun phrase. For example, using the Example 2 in Section 1, the segment before and after the boundary are sharing no noun phrase syntax.

(3) **Character Trigram (CT)**: Character trigram is reported having good performance for authorship attribution [10] although it may be similar to word based feature.

(4) **Word POS tag (WP)**: Word could have multiple grammatical function in sentences. People may have tendency to use certain word with certain grammatical function consistently. Therefore, in this feature, we combine the word and its POS tag as single feature variable.

(5) **Noun Phrase (NP)**: This feature compare the distribution of noun phrase. This feature undirectly compares topic of texts focusing only for noun phrase. For example, both Example 1 and 2 in Section 1 are rarely sharing noun phrases.

(6) **Function Word (FW)**: This feature contains words that are topic-independent such as "the" or "of". Someone possibly has tendency to use one of function words among the other words that have similar meaning, for example: the expression "at office" or "in office". We determine function word based on POS tag for the following label: conjunction, preposition, predeterminer, and determiner. For example, using Example 2 in Section 1, the segment after the boundary uses more function word "of" that the segment before the boundary.

4. Experiments

Since we proposed the use of text segmentation based on authorship attribution for retrieving source document in PD, we conducted the experiment to investigate whether text segmentation based on authorship attribution improves the performance of source retrieval or not, compared with topically text segmentation and no segmentation. We also investigated the best combination of feature types that optimizes the peformance of the source retrieval.

We used the dataset[*6] for the source retrieval task produced in the PAN workshop [8]. This dataset consists of query documents and a list of source documents for each query document.

There are 97 query documents that were manually produced by simulating plagiarism. Given a topic, such as **"Obama's family"**, a writer who was in charge of that topic searched the document collection for source documents and edited them to make a query document as natural as possible. The document collection is the ClueWeb09[*7], consisting of approximately 500 million Web pages. The statistics per query document are as follows.

- length: approximately 5 000 words (300 sentences)
- the number of source documents: 13
- the number of source boundaries: 26

Example topics of the query documents are given below:

(1) Barack Obama's family
(2) Mitchell College
(3) Air travel information
(4) Real estate valuation
(5) Cheap internet
(6) GMAT (Graduate Management Admission Test)
(7) Dinosaurs
(8) ESPN (Entertainment and Sport Programming Network)
(9) Kiwi
(10)Iron for your body

For the evaluation, we prepared the following methods:

(1) No segmentation: this method performs the source retrieval in PD without segmenting the query document.

(2) Word unigram (WU) feature: this method performs the source retrieval in PD with segmenting the query documents using the same process of our segmentation method, but employing word unigram as feature. It is natural to use this feature for segmenting text based on topic.

(3) Ideal segmentation: the source retrieval is performed using manually segmented query documents based on the source boundaries as ideal condition.

[*5] We use SVM library from http://scikit-learn.org/stable/modules/svm.html#classification

[*6] http://www.uni-weimar.de/medien/webis/corpora/corpus-pan-labs-09-today/pan-14/pan14-data/pan14-source-retrieval-training-corpus-2014-12-01.zip

[*7] http://www.lemurproject.org/clueweb09.php/

To measure the effectiveness of our method regarding the source retrieval, we use the following evaluation methods:

(1) Mean Average Precision (MAP)
(2) Precision (P)
(3) Recall (R)
(4) F-measure (F)

In addition, we also evaluate our method regarding the effectiveness of segmenting text using the following methods:

(1) Boundary similarity (BS): This method classifies the type of error of boundary whether it is total or near miss boundary and gives penalty according to this error type [1]. For this method, the greater the score, the better the performance.
(2) Window diff (WD): This method gives penalty when the number of boundary is not matched between two n-size sliding windows [6]. For this method, the lower the score, the better the performance.

Table 1 shows our evaluation results. We rescale the range of the scores from between 0 to 1 to 0 to 100. According to the evaluation results on text segmentation, if we compare our methods to WU, for evaluation results from WD, our proposed methods perform worse than WU. However, for evaluation results using BS, majority of our proposed methods perform better than WU. Additionally, all our proposed methods that use WU as feature combination perform better than WU. Because there is no consistency between WD and BS, it is difficult to investigate which method performs better.

To answer our question whether the text segmentation improves the source retrieval or not, we compare the score of MAP, P, R, and F of the method without segmentation to the other methods that use text segmentation. All methods using the text segmentation are always better than the method without text segmentation. It means that the text segmentation improves the performance of the source retrieval in PD.

When comparing the score of P, R, and F of WU with methods that combine WU with authorship attribution features, it shows that the better performance is achieved by methods using combination. This indicates that using authorship attribution features in text segmentation could improve the topically text segmentation that uses word unigram as feature.

Among the proposed methods and WU, we used two-tailed paired t-test for statistical testing and found that the difference between WU and each the proposed method in MAP was not statistically significant at the 0.1% level. We discover that the source documents do not have good rank in document list because of being overwhelmed by near-duplicate documents or documents that have similar topic with the source documents.

In the source retrieval result, the best performance is achieved by method that combine function word, noun phrase syntax, and noun phrase (FW-NS-NP). In addition, we used two-tailed paired t-test for statistical testing and found the difference between FW-NS-NP and WU in the F score was statistically significant at the 0.1% level. Moreover, this method also improves the source retrieval results of 56 query documents out of 97 documents.

5. Conclusion

We proposed a segmentation method that uses the following authorship attribution features: distribution of part of speech tag, noun phrase, noun phrase syntax, function word, word and its part of speech tag, and character trigram and divides the query document where the writing style is substantially different. We evaluate our method in the terms of text segmentation and the source retrieval performance.

Our evaluation showed that the authorship segmentation improves the source retrieval with the best performance is achieved by method using combination of function word, noun phrase syntax, and noun phrase. In the future, there is still room for improvement, for example by introducing the other authorship attribution features, applying segment weighting methods to distinguish segments that contains the source document or not, or using method that detects near-duplication.

References

[1] Fournier C.: Evaluating text segmentation using boundary edit distance, *Proc. Intl. Conf. ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1702–1712, Vol. 1 (2013).

[2] Graham N. and Hirst G. and Marthi B.: Segmenting documents by stylistic character, *Journal of Natural Language Engineering*, No. 4, Vol. 11, pp. 397–415 (2005).

[3] Hagen, Matthias and Potthast, Martin and Stein, Benno: Source retrieval for plagiarism detection from large web corpora: recent approaches, *Working Notes Papers of the CLEF*, pp. 1613–0073 (2015).

[4] Juola P.: Authorship attribution, *Journal of Foundations and Trends in Information Retrieval*, No. 3, Vol. 1, pp. 233–334 (2006).

[5] McCabe, Donald L: Cheating among college and university students: A North American perspective, *International Journal for Educational Integrity*, No. 1, Vol. 1, pp. 1–11 (2005).

[6] Pevzner L. and Hearst M.A.: A critique and improvement of an evaluation metric for text segmentation, *Journal of Computational Linguistics*, No. 1, Vol. 28, pp. 19–36 (2002).

[7] Potthast M. and Hagen M. and Stein B. and Graßegger J. and Michel M. and Tippmann M. and Welsch C.: ChatNoir: A search engine for the ClueWeb09 corpus, *Proc. of the Intl. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1004 (2012).

[8] Potthast M. and Hagen M. and Völske M. and Stein B.: Crowdsourcing interaction logs to understand text reuse from the web, *Proc. of the Conf. 51st Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 1212–1221 (2013).

[9] Reynar, Jeffrey C.: Topic Segmentation: Algorithms and Applications, *PhD thesis*, University of Pennsylvania, Philadelphia, PA, USA (1998).

[10] Stamatatos E.: A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology*, No. 3, Vol. 60, pp. 538–556 (2009).

[11] Takaki T. and Fujii A. and Ishikawa T.: Associative document retrieval by query subtopic analysis and its application to invalidity patent search, *Proc. of Intl. Conf. on Information and Knowledge Management*, pp. 399–405 (2004).

[12] Toutanova, Kristina and Klein, Dan and Manning, Christopher D and Singer, Yoram: Feature-rich part-of-speech tagging with a cyclic dependency network, *Proc. of the 2003 Conf. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Vol. 1, pp. 173–180 (2003).

[13] Weber-Wulff, Debora and Möller, Christopher and Touras, Jannis and Zincke, Elin: Plagiarism Detection Software Test 2013, available from ⟨http://plagiat.htw-berlin.de/wp-content/uploads/Testbericht-2013-color.pdf⟩ (accessed 2015-12-25).

Table 1 Effectiveness of segmenting query documents and source retrieval

Method	Segmentation		Source retrieval (top 300 docs)			
	BS	WD	MAP	P	R	F
Baseline						
w/o segmentation	0.00	100.00	1.19	0.49	5.25	0.09
WU	12.98	50.60	14.79	4.82	50.27	1.04
Combination of feature types						
WU-PT	13.18	51.27	14.60	4.87	52.09	1.09
WU-NS	13.23	52.85	14.77	4.90	52.81	1.09
WU-CT	13.75	49.36	13.81	4.98	54.35	1.13
WU-WP	13.45	50.08	14.25	5.04	53.89	1.13
WU-NP	13.28	53.21	15.49	5.22	55.08	1.13
WU-FW	13.38	52.38	14.96	5.20	54.42	1.13
FW-NS-NP-WP	13.19	54.85	16.36	5.55	58.37	1.22
WU-PT-FW-NS-CT-NP-WP	13.24	57.01	16.36	5.43	58.26	1.23
PT-FW-NS-NP	12.85	53.68	15.70	5.49	58.45	1.23
FW-NS-NP	12.38	59.17	16.97	5.69	58.98	1.23
Ideal situation						
Manual segmentation	100.00	0.00	13.55	5.62	60.17	1.31