Towards Document Plagiarism Detection based on the Relevance and Fragmentation of the Reused Text

Fernando Sánchez-Vega[1], Luis Villaseñor-Pineda[1],
Manuel Montes-y-Gómez[1,3] and Paolo Rosso[2]

[1] Laboratory of Language Technologies, Department of Computational Sciences,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
{fer.callotl, mmontesg, villasen }@inaoep.mx

[2] Natural Language Engineering Lab, ELiRF, DSIC,
Universidad Politécnica de Valencia, Spain.
prosso@dsic.upv.es

[3] Department of Computer and Information Sciences,
University of Alabama at Birmingham.

Abstract. Traditionally, External Plagiarism Detection has been carried out by determining and measuring the similar sections between a given pair of documents, known as source and suspicious documents. One of the main difficulties of this task resides on the fact that not all similar text sections are examples of plagiarism, since thematic coincidences also tend to produce portions of common text. In order to face this problem in this paper we propose to represent the common (possibly reused) text by means of a set of features that denote its relevance and fragmentation. This new representation, used in conjunction with supervised learning algorithms, provides more elements for the automatic detection of document plagiarism; in particular, our experimental results show that it clearly outperformed the accuracy results achieved by traditional n-gram based approaches.

1 Introduction

Plagiarism is regarded as intellectual theft; it consists in using the words (and ideas) of others and presenting them as your own. Nowadays, due to current technologies for creating and disseminating electronic information, it is very simple to compose a new document by copying sections from different sources extracted from the Web. This situation has caused the growing of the plagiarism phenomenon, and, at the same time, it has motivated the development of tools for its automatic detection.

From a general point of view document plagiarism detection divides in two major problems, intrinsic and external plagiarism detection [8]. The former aims to determine plagiarized sections by analyzing style changes within the document of interest, whereas, the latter tries to discriminate plagiarized from non-plagiarized documents by determining the reused text sections from a reference collection.

Regarding external plagiarism detection, its main concern involves finding similarities between any two documents which are more than just coincidence and

more likely to be result of copying [3]. This is a very complex task since reused text is commonly modified with the aim of hide or camouflage the plagiarism. To date, most approaches have only partially addressed this issue by measuring the lexical and structural similarity of documents by means of different kinds of features such as single words [4, 11], fixed length substrings (known as n-grams) [1, 4, 6, 7], and variable-length substrings [4, 2]. The main drawback of these approaches is that they carry out their decision/classification based on one single value/feature, namely, the degree of overlap between the suspicious and source documents. Due to this strategy, they are affected by the thematic correspondence of the documents, which implies the existence of common domain-specific word sequences, and, therefore, causes an overestimation of their overlap [3].

In order to face the above problem we propose to consider more information into the classification process of the documents. Our idea is to characterize the common (possibly reused) text by its relevance and fragmentation. In particular, we consider a set of features that denote the frequency of occurrence of common sequences as well as their length distribution. Our assumption is that the larger and the less frequent the common sequences the greater the evidence of plagiarism. In other words, we consider that frequent common sequences tend to correspond to domain specific terminology, and that small common sequences may be co-incidental, and, therefore, they are not a clear signal of plagiarism.

The experimental evaluation of the proposed approach was carried out on a subset of the METER corpus [5]. In particular, we model the document plagiarism detection as a classification problem, and, therefore, our goal was to show that using the proposed set of features, which better describe the particularities of the common sequences, it is possible to achieve a greater discrimination performance between plagiarized and non-plagiarized documents than only considering the general degree of overlap.

The rest of the paper is organized as follows. Section 2 describes the proposed representation of the common text. First, it formally defines the set of common sequences between two given documents. Then, it introduces the set of relevance and fragmentation features used to characterize the common text. Section 3 presents the experiments. It describes the experimental configuration and shows the results from the classification of 253 pairs of suspicious and source documents. The achieved results are encouraging; they indicate that the proposed approach outperformed by more than 7% the accuracy of the current approaches mentioned above. Finally, Section 4 depicts our conclusions and formulates some future work ideas.

2 A New Representation of the Common Text

Generally, as stated above, common word sequences between the suspicious and source documents are considered the primary evidence of plagiarism. Nevertheless, using their presence as unique indicator of plagiarism is too risky, since thematic coincidences also tend to produce portions of common text (i.e., false positives). In addition, even a minor modification to hide the plagiarism will avoid the identification of the corresponding sequences, generating false negatives. In order to handle these problems we propose using a set of features that describe diverse

characteristics of the common sequences, and, therefore, that facilitate the recognition of sequences denoting the reused (plagiarized) text.

Before introducing the proposed set of features we present the definition of a common sequence. Assuming that D_S and D_R are two documents, the suspicious and source (reference) documents respectively, and that each document is a sequence of words, where w_i^S and w_i^R are the ith words of D_S and D_R respectively, then:

Definition 1. The word sequence $\langle w_i^S, w_{i+1}^S, \dots, w_{j-1}^S, w_j^S \rangle$ contained in D_S is a common sequence between D_S and D_R if and only if there exist at least one sequence $\langle w_{i+x}^R, w_{i+x+1}^R, \dots, w_{j+x-1}^R, w_{j+x}^R \rangle$ in D_R, such that:

$$\forall_{i \le k \le j} \quad w_k^S = w_{k+x}^R$$

$$w_{i-1}^S \neq w_{i+x-1}^R$$

$$w_{j+1}^S \neq w_{j+x+1}^R$$

In order to learn to discriminate between plagiarized and non-plagiarized documents, we propose to characterize the set of common sequences (denoted by Ψ) by two main kinds of features, namely, relevance and fragmentation features. The next formula shows the proposed representation of Ψ.

$$\Psi = \langle f_1^{rel}, f_2^{rel}, \dots, f_m^{rel}, f_1^{frg}, f_2^{frg}, \dots, f_m^{frg} \rangle$$

As noticed, we represent the set of common sequences by $2 \times m$ features, where each feature f_i^{rel} and f_i^{frg} indicate the relevance and fragmentation of the sequences of length *i* respectively. Cases of particular interest are the f_m features, which indicate the values of all sequences with length equal or greater than *m* (a user-defined value). Their purpose is to deal with the data sparseness and to allow taking advantage of the occurrence of discriminative but very rare longer sequences.

Following we define both kind of features. For the sake of simplicity we first describe fragmentation features and afterward relevance features.

Fragmentation features. By means of these features we aim to find a relation between the length and quantity of common sequences and the plagiarism. These features are based on two basic assumptions. On the one hand, we consider that the longer the sequences the greater the evidence of plagiarism, and, on the other hand, based on the fact that long sequences are very rare, we consider that the more the common sequences the greater the evidence of plagiarism.

According to these basic assumptions we compute the value of the f_i^{frg} feature by adding the lengths of all common sequences of length equal to *i* as described in the following formula:

$$f_i^{frg} = \sum\nolimits_{\{seq_j : seq_j \in \Psi \ \wedge |seq_j| = i\}} |seq_j|$$

The definition of the agglomerative feature f_m^{frg} is as stated below:

$$f_m^{frg} = \sum_{\{seq_j:\, seq_j \in \Psi \wedge |seq_j| \geq m\}} |seq_j|$$

Relevance features. This second group of features aims to qualify the sequences by their words. That is, they aim to determine the relevance of the sequences with respect to the thematic content of both documents. The idea behind these features is that frequent words/sequences are related to the topic of the documents, and not necessarily are a clear signal of plagiarism. On the contrary, they are supported on the intuition that plagiarism is a planned action, and, therefore, that plagiarized sections (sequences) are not exhaustively used.

In particular we measure the relevance of a given common sequence $seq_i \in \Psi$ by the following formula:

$$relevance(seq_i) = \frac{1}{e^{occ(seq_i, D_S)-1}} \times \prod_{k=1}^{|seq_i|} \frac{2}{occ\left(w_k^{seq_i}, D_S\right) + occ\left(w_k^{seq_i}, D_R\right)}$$

where $occ(seq_i, D)$ indicates the occurrences of the common sequence seq_i in document D, and $occ(w_k, D)$ indicates the times word w_k occurs in D.

This measure of relevance has two components, the first one evaluates how frequent is the given sequence in the suspicious document, strongly penalizing frequent sequences because they have more probability of being idiomatic or domain specific expressions. On the other hand, the second component castigates the sequences formed by words that are frequent in both documents. As noticed, this formula reaches its greatest value (relevance = 1), when the common sequence (and all their inner words) appear exclusively once in both documents, indicating that it has great chance for being a deliberate copy.

Based on the definition of the relevance of a sequence, relevance features are computed as follows:

$$f_i^{rel} = \sum_{\{seq_j:\, seq_j \in \Psi \wedge |seq_j| = i\}} relevance\left(seq_j\right)$$

The definition of the agglomerative feature f_m^{rel} is as follows:

$$f_m^{frg} = \sum_{\{seq_j:\, seq_j \in \Psi \wedge |seq_j| \geq m\}} relevance\left(seq_j\right)$$

3 Experimental Evaluation

3.1 The corpus

For the experiments we used a subset of the METER corpus[1] [5]; a corpus specially designed to evaluate text reuse in the journalism domain. It consists of annotated examples of related newspaper texts collected from the British Press Association (PA) and nine British newspapers that subscribe to the PA newswire service.

In the METER corpus news from the PA are considered as the source documents and the corresponding notes from the newspapers are regarded as the suspicious documents. In particular, we only used the subset of news (suspicious document) that has only one single related note (source documents). That is, we considered a subset of 253 pairs of source-suspicious documents.

In this corpus each suspicious document (note from a newspaper) is annotated with one of three general classes indicating its derivation degree with respect to the corresponding PA news: wholly-derived, partially-derived and non-derived. For our experiments we considered wholly and partially derived documents as examples of plagiarism and non-derived documents as examples of non-plagiarism, modeling in this way the plagiarism detection task as a two-class classification problem. In particular, the formed evaluation corpus consists of 181 instances of plagiarism and 72 of non-plagiarism.

3.2 Evaluation

For the evaluation of the proposed approach, as well as for the evaluation of the baseline methods, we employed the Naïve Bayes classification algorithm as implemented by Weka [10], and applied a 10 cross-fold validation strategy. In all cases we preprocessed the documents by substituting punctuation marks by a generic label, but we did not eliminate stopwords nor apply any stemming procedure.

The evaluation of results was carried out mainly by means of the classification accuracy, which indicates the overall percentage of documents correctly classified as plagiarized and non-plagiarized. Additionally, due to the class imbalance, we also present the averaged F_1 measure as it was used in the work by [4], which indicates the average of the F_1 scores across the two classes.

3.3 Selection of the *m* value

As we explained in Section 2, we propose representing the set of common sequences between the suspicious and source documents by a vector of $2 \times m$ features. In this vector, each feature indicates the relevance or fragmentation of the sequences of a particular length, except for the *m*-features which integrate information from all sequences with length greater than *m*.

In order to determine an appropriate value of *m* for our experiments we evaluated the information gain (IG) [9] of each obtained feature. Given that we extracted

[1] www.dcs.shef.ac.uk/nlp/funded/meter.html

common sequences of lengths varying from 1 to 61, we initially constructed a representation of 122 features. Then, for each one of the 10 folds, we evaluated the information gain of these features, and, finally, we decided preserving those having an averaged-IG greater than 0.1. Following this procedure we established $m = 4$ for the experiments reported in this paper. As a reference, Table 1 shows the obtained averaged IG values as well as their standard deviation (for the 10 different folds) for the first five features, which correspond to sequences of lengths from 1 to 5.

Table 1. IG of the first five features of the proposed representation

Length of sequences	Average IG	Standard Deviation
1	0.382	0.024
2	0.288	0.025
3	0.125	0.026
***m* = 4**	0.037	0.025
5	0.006	0.017

3.4 Baseline results

Table 2 shows some baseline results corresponding to current approaches for document plagiarism detection. For these results, the classification was carried out using different features denoting the percentage of overlap between the suspicious and source documents. In particular, for the first experiment we measured this overlap by means of the common words (unigrams); for the second experiment we represented the overlap by three features corresponding to the percentage of common unigrams, bigrams and trigrams respectively, and, for the third experiment we considered as single feature the percentage of common words extracted from the common sequences.

As noticed, all results are very similar being the one based on the percentage of common unigrams the best. This result indicates that the used corpus has a great level of modification (rewritten), and, therefore, that the insertion of words for cutting long sequences may be high. On the other hand, this result was worrying (for us), since it indicates that structural information (not captured by unigrams) is not needed, and, in contrast to this conjecture, our approach aims to take advantage of this kind of information.

Table 2. Baseline results: based on the proportion of common n-grams and sequences

Kind of features	Number of features	Accuracy	F_1 measure
Unigrams	1	73.12%	0.655
{1,2,3}-grams	3	70.75%	0.6885
Common sequences (length ≥ 2)	1	72.72%	0.677

3.5 Results of the proposed approach

Table 3 shows the results from the proposed approach. The first two rows indicate the results achieved by the relevance and fragmentation features respectively, whereas,

the last row presents the results obtained by their combination. Results from this table indicate that:

- Relevance and fragmentation features are both relevant for the task of plagiarism detection. In particular, fragmentation features showed to be very appropriate, outperforming the classification accuracy of current methods; whereas, relevance features only obtained comparable results.
- Relevance and fragmentation features are complementary; their combined usage allowed obtaining a better result than their individual applications.
- Results of the proposed approach, based on the combination of relevance and fragmentation features, improved by more than 7% the accuracy of the reference methods, and by more than 2% their averaged F_1 measure.

Table 3. Results of the proposed approach based on relevance and fragmentation features from the common sequences

Kind of features	Number of features	Accuracy	F_1 measure
Fragmentation features f_i^{frg} ($m = 4$)	4	77.07%	0.6755
Relevance features f_i^{rel} ($m = 4$)	4	73.91%	0.606
All features f_i^{frg} and f_i^{rel} ($m = 4$)	8	78.26%	0.7045

4 Conclusions and Future Work

This paper describes the first ideas of a new approach for external plagiarism detection. This approach is based on the characterization of the common (possible reused) text between the source and suspicious documents by its relevance and fragmentation. In particular, it considers a set of features that denote the frequency of occurrence of the common sequences as well as their length distribution. The main assumption is that the larger and the less frequent the common sequences the greater the evidence of plagiarism.

Experimental results on a subset of 253 pairs of source-suspicious documents from the METER corpus are encouraging; they indicated that the proposed features are appropriate for the plagiarism detection task and that they provide relevant elements for a classifier to discriminate between plagiarized and non plagiarized documents. In particular, the achieved accuracy results outperformed by more than 7% the results from other current methods based on the use of one single feature describing the degree of overlap between the documents.

As future work we plan to investigate more features describing the common text between the source and suspicious documents. For instance, we consider incorporating some features that describe the density of the common sequences in the suspicious document as well as features that capture their relative order in both documents. In addition we plan to improve the evaluation of the relevance of single words by computing statistics from the Web or other external but thematically related corpus.

Acknowledgments. This work was done under partial support of CONACYT (Project grants 83459, 82050, 106013 and scholarship 258345), and the research work of the

last author is partially funded by CONACYT-Mexico and the MICINN project TEXTENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i). In addition, we thank Paul Clough for his help by providing us the METER corpus.

References

1. Barrón-Cedeño, A., Rosso, P.: On Automatic Plagiarism Detection Based on n-grams Comparison. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (ECIR), Berlin, Heidelberg, 2009.
2. Basile, C., Benedetto, D., Caglioti, E., Cristadoro, G., Degli Esposti, M.: A Plagiarism Detection Procedure in Three Steps: Selection, Matches and "Squares". In: Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09), pp. 1-9, Donostia-San Sebastian, Spain, September 2009.
3. Clough. P.: Old and new challenges in automatic plagiarism detection. In: National Plagiarism Advisory Service, 76, 2003.
4. Clough, P., Gaizauskas, R., Piao, S., Wilks, Y.: METER: Measuring Text Reuse. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002.
5. Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., Piao, S.: The meter corpus: A corpus for analysing journalistic text reuse. In: Proceedings of the Corpus Linguistics 2001 Conference, 2001.
6. Grozea, C., Gehl, C., Popescu. M.: ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In: Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09), pp. 1-9, Donostia-San Sebastian, Spain, September 2009.
7. Kasprzak, J., Brandejs, M., Křipač, M.: Finding Plagiarism by Evaluating Document Similarities. In: Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09), pp. 1-9, Donostia-San Sebastian, Spain, September 2009.
8. Potthast M., Stein, B., Eiselt A., Barrón-Cedeño A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09), pp. 1-9, Donostia-San Sebastian, Spain, September 2009.
9. Sebastiani, F., Machine learning in automated text categorization. ACM Comp. Surv., 34,1, 2002.
10. Witten, I., H., Frank, E.: Data Mining Practical Machine Learning Tools and Techniques, Elsevier, 2005.
11. Zechner, M., Muhr, M., Kern, R., Granitzer, M.: External and Intrinsic Plagiarism Detection using Vector Space Models. In: Proceedings of the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09), pp. 1-9, Donostia-San Sebastian, Spain, September 2009.