Intrinsic Plagiarism Detection using N-gram Classes

Imene Bensalem
MISC Lab
Constantine 2 University, Algeria
bens.imene@gmail.com

Paolo Rosso
NLE Lab
PRHLT Research Center
Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Salim Chikhi
MISC Lab
Constantine 2 University, Algeria
slchikhi@yahoo.com

Abstract

When it is not possible to compare the suspicious document to the source document(s) plagiarism has been committed from, the evidence of plagiarism has to be looked for intrinsically in the document itself. In this paper, we introduce a novel language-independent intrinsic plagiarism detection method which is based on a new text representation that we called n-gram classes. The proposed method was evaluated on three publicly available standard corpora. The obtained results are comparable to the ones obtained by the best state-of-the-art methods.

1 Introduction and Related Works

Intrinsic plagiarism detection is an essential alternative in situations where the plagiarism source does not have a digital version, e.g. an old book, or the plagiarized text was directly written by another author without copying from any source, e.g. the case of a student who asked someone else to write for him parts of his essay or thesis. Hence, the task of detecting plagiarism intrinsically is to identify, in the given suspicious document, the fragments that are not consistent with the rest of the text in terms of writing style.

The automatic analysis of the writing style is an important component of many NLP applications. For some of them, when analyzing the style, a document is considered as a whole, which is the case of the authorship identification (Stamatatos, 2009a) and the authorship verification (Koppel and Seidman, 2013). For other applications, a document is perceived as a set of fragments, for each of them the writing style needs to be analyzed individually. Examples of such applications include: paragraph authorship clustering (Brooke and Hirst, 2012), authorial segmentation of multi-author documents (Akiva and Koppel, 2013), detection of stylistic inconsistencies between consecutive paragraphs (Graham et al., 2005) and plagiarism direction identification (Grozea and Popescu, 2010).

For intrinsic plagiarism detection, it is crucial to analyze the writing style at fragments level. However, the majority of methods tend to analyze the whole document writing style as well. Indeed, intrinsic plagiarism detection puts together, in one research problem, many difficulties that are not present, or present separately, in the aforementioned related problems. Its main difficulties are listed below.

In contrast to multi-author documents related problems, the number of authors in the suspicious documents is unknown, i.e., it might be one author if the document is plagiarism-free or many unknown authors if it contains plagiarism.

Unlike the authorship attribution and verification, where the examined text and the potential author text are separate (and hence their writing styles could be readily characterized and compared), these two parts are both merged in the same document with unknown boundaries. Furthermore, the plagiarized fragments in a suspicious document might stem from different authors, which renders the computational characterization of plagiarism difficult.

As opposed to the problem of authorship clustering, where the task is merely to attribute already defined fragments of a given document to different authors, the segmentation is a crucial and inevitable task in a real scenario of intrinsic plagiarism detection. Indeed, a granular segmentation may lead to an undependable style analysis, and a coarse segmentation may prevent the identification of the short plagiarized texts.

Due to the aforementioned difficulties, intrinsic plagiarism detection is still a challenging

problem. This is evidenced by the still low performance scores of the majority of methods[1]. To the best of our knowledge, just two methods, namely Stamatatos (2009b) and Oberreuter et al. (2011), reached an f-measure greater than 0.30 on a standardized corpus. Other methods, for instance (Stein et al., 2011) and (Tschuggnall and Specht, 2013), obtained better performance scores. Nonetheless, they have been evaluated on only selected documents from the whole standardized evaluation corpus which makes their results not comparable to the others.

Although the writing style analysis is an old research area and has been applied successfully to solve many problems, notably authorship attribution, it is obvious that its application to identify the plagiarized fragments still needs to be investigated further. In this paper, we address this research problem by proposing a novel way of quantifying the writing style that we called n-gram classes. We show that our method, which is supervised classification-based, is able to discriminate between the plagiarized and the original text fragments with a performance comparable to the best state-of-the-art methods despite it uses a small number of features when building the classification model.

The remainder of the paper is organized as follows. Section 2 presents our motivation. Sections 3 and 4 present the new features and the proposed method. Section 5 provides the evaluation results. Finally, Section 6 draws our conclusions.

2 Motivation

The idea of our method is inspired by the work of Grozea and Popescu (2010), in the context of plagiarism direction identification. They reported that the character 8-grams of a plagiarized text fragment are more frequent in the source document (because the author is the same) than in the plagiarized document. Thus, we believe that, it is possible to distinguish the plagiarized fragments from the original ones on the basis of the frequency of their character n-grams in the suspicious document. That is, if many of the character n-grams of a fragment are infrequent in the document, it would be probably a plagiarized fragment. However, if many of them are frequent, then the fragment is likely to be original.

On the other hand, according to the authorship attribution researches, character n-grams are a powerful tool for characterizing the writing style (Stamatatos, 2009a). Moreover, they have been used in one of the best intrinsic plagiarism detection methods (Stamatatos, 2009b).

Generally, in n-gram based methods the text is represented by a vector of n-grams with their frequencies. The shortcoming of this text representation is the increase of its size with the increase of the text or the n-gram length.

Our method proposes a novel way of using character n-grams[2] for text representation. The idea is to represent the fragments of the suspicious document in a reduced vector where each feature value is the frequency of a class of n-grams instead of a particular n-gram. Therefore, the dimension of any fragment vector is always equal to the number of classes rather than the number of n-grams. The class of an n-gram is determined according to its frequency level in the given document as we will show in the next section.

3 N-gram Classes

Formally, we define an n-gram class as a number from 0 to $m-1$ such that the class labeled 0 involves the *least frequent* n-grams and the class labeled $m-1$ contains the *most frequent* n-grams in a document. If $m > 2$, classes between 0 and $m-1$ will contain n-grams with *intermediate frequency levels*.

Concretely, to assign the n-grams of a given document to m classes, first, the document is represented by a $2 \times l$ matrix (l is the total number of n-grams), where the first row contains the n-grams ng_i (i =1..l) and the second one contains their number of occurrences $freq_i$ (raw frequency).

Let *max_freq* denotes the maximum frequency, so:

$$max\_freq = \text{argmax} \quad freq_i ; \quad i=1..l \qquad (1)$$

Then, the class of a n-gram ng_i is computed as follows:

$$\text{Class } ng_i = \text{Log}_{\text{base}} (freq_i); \qquad (2)$$

Given that:

$$\text{base} = \sqrt[m-1]{max\_freq} \quad . \qquad (3)$$

By computing the base of the logarithm as shown in the equation (3), the most frequent n-grams (i.e. the n-grams with the maximum number of occurrences) will be in the class $m-1$, and

[1] See for instance PAN workshop (http://pan.webis.de) series, from 2007 to 2012, where several papers on intrinsic plagiarism detection have been published.

[2] In the rest of the paper, when not said differently, the term n-gram is always used to denote character n-gram.

the least frequent n-grams (e.g. the ones that appear only once) will be in the class 0, and the n-grams with intermediate levels of frequency will be in the classes between 0 and $m-1$. Figure 1 illustrates an example of computing the n-gram classes of a document. The chosen number of classes m in this example is 3.

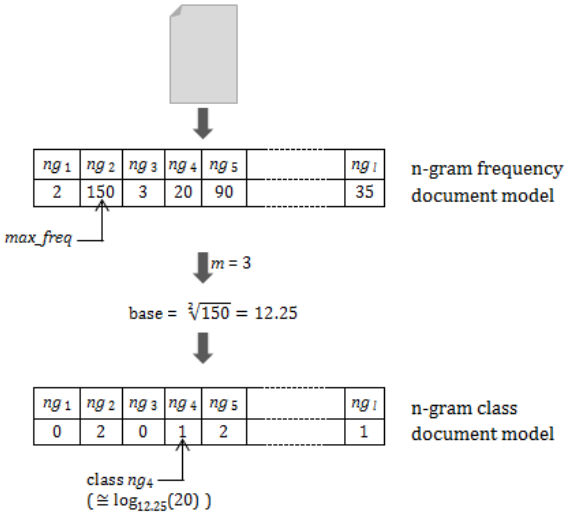

Figure 1. Steps for computing the n-gram classes of a document. The number of classes in this example is 3 (class labels are from 0 to 2).

Note that, what we explained above is solely how to compute the class of each n-gram of a document. However, our purpose is to represent the document fragments using these classes. To this end, for each fragment, first, its n-grams are extracted. Then, each n-gram is replaced by its class obtained from the document model built previously. Finally, the proportion of each class in the fragment is computed. So, the fragment can be represented by a vector of m values, where the first value is the proportion of the class 0, the second value is the proportion of the class 1 and so on. Figure 2 illustrates these steps. For the sake of simplicity, we suppose that the fragment contains only 5 n-grams.

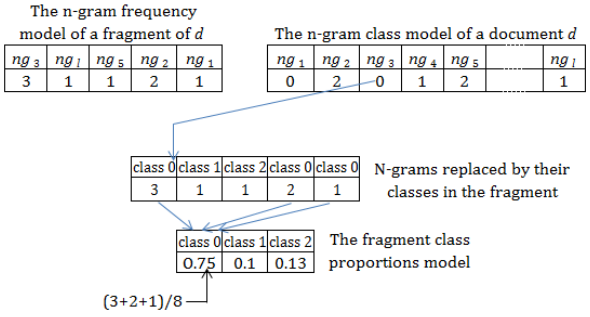

Figure 2. Steps for representing a document fragment by the proportion of 3 n-gram classes.

4 The Proposed Method

Once the suspicious document has been segmented to fragments and these latter have been represented by a set of features, an important phase in the process of the intrinsic plagiarism detection is to decide whether a fragment is plagiarized or original. This phase has been implemented in the literature methods using different techniques, notably clustering (Akiva, 2011), supervised classification (Meyer zu Eissen et al., 2007), distance functions with thresholds (Stamatatos, 2009b; Oberreuter et al., 2011) and density-based methods (Stein et al., 2011).

In our supervised method, the classification model is trained with a small number of features which are the proportions of the n-gram classes described in the previous section.

In detail, our method is composed of the following steps:

1. Segment each document d into fragments s_i by using the sliding window technique. Let S denotes the set of these fragments.
2. Build the n-gram class document model (see Figure 1) without considering numerals. We choose to consider the frequency of a n-gram ng_i as the number of its occurrence in d such that it is counted once per fragment. Therefore, the minimum value that could take a frequency is 1 if ng_i appears only in one fragment, and its maximum value is $|S|$ (the number of fragments in d) if ng_i occurs in each fragment $s_i \in S$.
3. Represent each fragment s_i by a vector of m features f_j , $j \in \{0,\ldots, m-1\}$. So that, each f_j is the proportion of the n-grams that belong to the class labeled j to the total number of n-grams in s_i.
4. Combine into one dataset the fragment vectors obtained from all the training corpus documents. Then, label each vector with its authenticity state, i.e. plagiarized, if the fragment plagiarism percentage exceeds 50% and original otherwise.
5. Build a classifier using the training set produced in the previous step. For this purpose, we trained and tested several classification algorithms implemented on WEKA software (Hall et al., 2009). The best results were obtained with the Naïve Bayes algorithm[3].

The aforementioned steps represent the training phase of our method, which aims to construct the classifier. In practice, in order to detect the plagiarism in a given document, this classifier is

[3] Consult the arff file from the archive file associated to this paper which contains the fragments class proportion model and the plagiarism prediction for each fragment.

directly applied to the document fragments after the step 3.

5 Evaluation

5.1 Datasets

We evaluated our method on 3 corpora: PAN-PC-09[4] and PAN-PC-11[5] which are the corpora used in the international competition of plagiarism detection in 2009 and 2011 respectively[6], as well as InAra corpus[7], which is a publicly available collection of artificial suspicious documents in Arabic (Bensalem et al., 2013). The three document collections include XML annotations indicating the plagiarized segments positions.

For the evaluation on English documents, the classifier has been trained on PAN-PC-11 test corpus and evaluated on this same corpus using 10-fold cross validation as well as PAN-PC-09 test corpus. For the evaluation on Arabic documents, the classifier has been trained and tested on InAra corpus using 10-fold cross validation.

5.2 Results

As evaluation measures we used macro-averaged precision, recall, f-measure, granularity and plagdet as they were defined in (Potthast et al., 2010).

In order to choose the parameters of our method, we trained the classifier using various training sets generated by using the different combinations of the n-gram length *n* (from 1 to 10) and the number of classes *m* (from 2 to 10). We adopted the parameters that yielded the higher f-measure, namely $n = 6$ and $m = 4$.

With regard the sliding window parameters, we used three different options for the window size, which are 100, 200 and 400 words, with a step equal to the quarter of the window size. Only one option is applied to a given document depending on its length.

We deliberately use similar sliding window parameters as the method of Oberreuter et al. (2011)[8] in order to compare the two methods without being much affected by the segmentation strategy.

Table 1 compares the results of our method to the one of Oberreuter et al. (2011) being the winner in PAN 2011 competition and considered one of the best intrinsic plagiarism detection methods.

		Our method	Oberreuter et al.[9]
PAN-PC-09	Precision	0.31	**0.39**
	Recall	**0.49**	0.31
	F-measure	**0.38**	0.35
	Granularity	1.21	**1.00**
PAN-PC-11	Precision	0.22	**0.34**
	Recall	**0.50**	0.31
	F-measure	0.30	**0.33**
	Granularity	1.13	**1.00**
InAra	Precision	0.24	**0.29**
	Recall	**0.69**	0.25
	F-measure	**0.35**	0.27
	Granularity	**1.27**	1.44

Table 1. Performance of the n-gram frequency class method on 3 corpora.

From Table 1 it can be appreciated that our method in terms of recall noticeably outperforms Oberreuter et al. (2011), although precision and granularity still needs to be further improved. Nonetheless, in comparison with other methods such as the one of Stamatatos (2009b), that obtained the best results in PAN 2009 competition on plagiarism detection, precision is still very much competitive: 0.31 vs. 0.23 (PAN-PC-09) and 0.22 vs. 0.14 (PAN-PC-11). In terms of f-measure, Oberreuter et al. (2011) method is significantly higher than our method on PAN-PC-11 corpus, but both methods have statistically similar results on InAra[10].

Considering plagdet, which is a score that represents the overall performance of a plagiarism detection method, our method could be ranked the 2nd, after Oberreuter et al. (2011) and

[4] http://www.uni-weimar.de/en/media/chairs/webis/research/corpora/corpus-pan-pc-09/

[5] http://www.uni-weimar.de/en/media/chairs/webis/research/corpora/corpus-pan-pc-11/

[6] We used only the corpora parts that are dedicated to the evaluation of the intrinsic approach.

[7] http://sourceforge.net/projects/inaracorpus/

[8] Oberreuter et al. (2011) used mainly 400 words as the window size that may change according to the document length.

[9] The results of Oberreuter et al. method (2011) on PAN-PC-09 and PAN-PC-11 are taken from his paper. However, we re-implemented this method in order to evaluate it on InAra. Note that our re-implementation maybe not perfectly similar to the original one since the authors did not provide details on the parameters tuning.

[10] The Kolomogorov Smirnov test with a significance level of 5% has been used to compare the two methods f-measures on PAN-PC-11 and InAra. Unfortunately, on the PAN-PC-09 corpora we were unable to carry out this test since we do not have the results of Oberreuter et al. per each document.

before Stamatatos (2009b) as it is shown in Table 2.

	Oberreuter et al.	Our method	Stamatatos
PAN-PC-09	0.35	0.33	0.25
PAN-PC-11	0.33	0.28	0.19

Table 2. Plagdet of our method in comparison with the two best methods on PAN competition corpora.

6 Conclusion

In this paper we have shown that representing the text fragments of a given suspicious document by the proportion of character n-gram classes (the most frequent, the least frequent and intermediate levels) is a promising way for detecting plagiarism intrinsically.

The experiments described in this paper were performed on three corpora comprising documents in English and for the first time Arabic. We obtained comparable results to the best performing systems.

Our method best configuration is 6 as the n-grams length and only 4 as the number of classes (i.e. 4 features). As future work, it would be interesting to combine the most precise classes of different n-gram lengths in order to improve the precision. It would be important as well to try other segmentation strategies and post-processing techniques in order to improve the granularity. Another interesting experiment we plan to carry out in the future is to use the n-gram classes along with the traditional stylistic features such as the vocabulary richness, average sentence length, etc.

Acknowledgments

The first author would like to thank Parth Gupta for his helpful feedback and Gabriel Oberreuter for providing some implementation details of his method.

The work of the second author was carried out in the framework of DIANA APPLICATIONS-Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) and WIQ-EI IRSES (Grant No. 269180 within the EC FP 7 Marie Curie People) research projects, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

Navot Akiva. 2011. Using Clustering to Identify Outlier Chunks of Text - Notebook for PAN at CLEF 2011. In *Notebook Papers of CLEF 2011 LABs and Workshops, September 19-22, Amsterdam, The Netherlands*, pages 5–7.

Navot Akiva and Moshe Koppel. 2013. A Generic Unsupervised Method for Decomposing Multi-Author Documents. *Journal of the American Society for Information Science and Technology*, 64(11):2256–2264.

Imene Bensalem, Paolo Rosso, and Salim Chikhi. 2013. A New Corpus for the Evaluation of Arabic Intrinsic Plagiarism Detection. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *CLEF 2013, LNCS, vol. 8138*, pages 53–58, Heidelberg. Springer.

Julian Brooke and Graeme Hirst. 2012. Paragraph Clustering for Intrinsic Plagiarism Detection using a Stylistic Vector-Space Model with Extrinsic Features - Notebook for PAN at CLEF 2012. In *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy*.

Neil Graham, Graeme Hirst, and Bhaskara Marthi. 2005. Segmenting Documents by Stylistic Character. *Natural Language Engineering*, 11(04):397–415.

Cristian Grozea and Marius Popescu. 2010. Who ' s the Thief ? Automatic Detection of the Direction of Plagiarism. In *CICLing 2010, Iaşi, Romania, March 21-27, LNCS, vol. 6008*, pages 700–710. Springer.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Moshe Koppel and Shachar Seidman. 2013. Automatically Identifying Pseudepigraphic Texts. In *EMNLP 2013*, pages 1449–1454, Seattle,Washington, USA. Association for Computational Linguistics.

Sven Meyer zu Eissen, Benno Stein, and Marion Kulig. 2007. Plagiarism Detection without Reference Collections. In Reinhold Decker and Hans -J. Lenz, editors, *Advances in Data Analysis, Selected Papers from the 30th Annual Conference of the German Classification Society (GfKl), Berlin*, pages 359–366, Heidelberg. Springer.

Gabriel Oberreuter, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez. 2011. Approaches for Intrinsic and External Plagiarism Detection - Notebook for PAN at CLEF 2011. In *CLEF 2011 Evaluation Labs and Workshop – Working Notes*

Papers, September 19-22, Amsterdam, The Netherlands, pages 1–10.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In Chu-Ren Huang and Daniel Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 997–1005, Stroudsburg, USA. Association for Computational Linguistics.

Efstathios Stamatatos. 2009a. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science*, 60(3):538–556.

Efstathios Stamatatos. 2009b. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors, *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09)*, pages 38–46. CEUR-WS.org.

Benno Stein, Nedim Lipka, and Peter Prettenhofer. 2011. Intrinsic Plagiarism Analysis. *Language Resources and Evaluation*, 45(1):63–82.

Michael Tschuggnall and Günther Specht. 2013. Using Grammar-Profiles to Intrinsically Expose Plagiarism in Text Documents. In *NLDB 2013, LNCS, vol. 7934*, pages 297–302. Springer.