Intrinsic Detection of Plagiarism based on Writing Style Grouping

Maryam Elamine[1], SeifEddine Mechti[2], Lamia Hadrich Belguith[3]

[1]ANLP Group, FSEGS, University of Sfax
mary.elamine@gmail.com
[2]LARODEC Laboratory, ISG of Tunis, University of Tunis,
mechtiseif@gmail.com
[3]ANLP Group, MIRACL Laboratory, FSEGS, University of Sfax
l.belguith@fsegs.rnu.tn

Abstract. In this paper, we tackle the task of intrinsic plagiarism detection, also referred to as author diarization. This task deals with identifying segments within a document written by multiple authors [2]. The main goal is to discover deviations in the writing style, looking for parts of the document that could potentially be written by another person [4]. In this paper, we present our hybrid approach that constructs a style function from stylometric features and detects the outliers. The proposed approach has been evaluated on two publicly available corpora. The obtained results outperform the ones obtained by the best state-of-the-art methods.

Keywords: author diarization, plagiarism, intrinsic plagiarism detection, outliers detection.

1 Introduction

Plagiarism detection is the task of identifying text reuse in a document or collection of documents [2]. It can be clustered as two main tracks: 1) Intrinsic plagiarism detection and 2) Extrinsic plagiarism detection. The former is the process of verifying the unity of a document against itself using local analysis. It focuses on finding whether the document was written by the same author or if there exists some parts written by other ones. The latter is the process of evaluating a document and verifying if there exists some parts that have been copied from external sources [8], thus the suspected document is compared with a collection of source documents [3].

The term "author diarization" came from the domain of speaker diarization, which is concerned with clustering and identifying various speakers from a single audio speech signal. Then the frequency range of the speakers' voices are analyzed (e.g. a class discussion on a particular topic). Likewise, the task of author diarization deals with a written document instead of audio conversations [5]. Its objective is to identify and cluster different authors within a single document [2]. The task of author diarization is extended and generalized by the introduction of within-document author clustering problems [9]. Author diarization consists of the following three sub tasks [1]:

- *Traditional intrinsic plagiarism detection:* There exists one main author who wrote at least 70% of the considered document
- *Diarization with a given number of authors:* The document is written by a known number of authors
- *Unrestricted diarization:* The number of collaborating authors is unknown.

In this paper, we present our proposed approach for intrinsic plagiarism detection by automatic grouping of writing style and for that purpose, we will explore two corpora; PAN 16[1] and PAN 17[2]. We first group authors based on their writing style and for that, we explore a writing style function. Then, from the generated clusters there is a high probability that an author has plagiarized from another, so following this hypothesis, we segment the documents into segments of 500 characters, attribute a style function to each fragment, if there exists some parts with a different style in the document then it is plagiarized. Our approach uses a hybridization of stylistic features that include lexical features (Mean sentence length, Type-token ratio, punctuation marks and letter count) and syntactic features (POS tags and function words ratio). The obtained results outperform the ones obtained by the best state-of-the-art methods that have also exploited the same corpora (i.e. the PAN16 corpus and the PAN17 corpus).

2 Related Work

Plagiarism is rising as a serious problem in the academic and educational domains [3]. With the explosive growth of content found throughout the Web, people can find nearly everything they need for their written work. Thus, detection of such cases can become a monotonous task [4].

In their study, M. Kuznetsov et al. [1] used stylometric features such as character n-grams, word n-grams, punctuation marks and pronouns count. The authors used the PAN 11 corpus for experimenting and the PAN 16 corpus for evaluating. They formulated the task of intrinsic plagiarism as text segments classification. They exploited a per-sentence approach [12]. This approach constructs disjoint segments to different length and detects plagiarism on sentence level. In fact, sentences are labeled following this rule: if more than half the characters in a sentence "s" are plagiarized, then it is labeled as plagiarized, otherwise it is labeled as non-plagiarized. For the features used, first they exploited the word frequencies trait, which is based on analyzing occurrences of text words; the lowercased sequences of characters with the exception of stopwords. Second, they utilized the n-gram frequencies attribute in which they count the n-gram frequencies. In fact, experiments showed that for better use, it's best to exploit 1-grams, 3-grams and 4- grams jointly. The resulting n-gram feature returns for each of the considered n-grams three statistics. Finally, for the final feature, the authors count for each sentence the number of occurrences of the most common punctuation marks (!,.?- ;) and the universal POS tags (VERB, NOUN, ADJ, ADV, PRON, CONJ, ADP, DET,

[1] http://pan.webis.de/clef16/pan16-web/author-identification.html
[2] http://pan.webis.de/clef17/pan17-web/author-identification.html

PRT, NUM). For each sentence, they count its length in characters and the mean length of the sentence words. After constructing their features, the authors move on to detect the outliers. For the author diarization task, they adapted the intrinsic plagiarism approach to solve the next problem. The algorithm functions the same way as previously described, but instead of the outlier detection phase, this approach provides segmentations of series using the Hidden Markov Model (HMM) with Gaussian emissions [13]. For the task of author diarization with unknown number of authors, the authors estimated the number by computing an averaged *t*-statistic for all pairs of author segments. Afterwards, they iterated through a probable number n ($n \in [2..20]$), then they computed the time series segmentation for each n. For each segmentation, the measure of clusters discrepancy is computed [1]:

$$Q_n = \sum_{i,j=1}^{n} \frac{|m(c_i) - m(c_j)|}{\sqrt{\frac{\sigma(c_i)^2}{l(c_i)} + \frac{\sigma(c_j)^2}{l(c_j)}}} \quad (1)$$

Where m (c_i) is the mean of elements in cluster c_i, $\sigma(c_i)$ is the mean deviation, and l (c_i) is the cluster size.

The final estimation $\hat{n}$ maximizes Q (n). After obtaining the estimation, the algorithm performs a diarization with a known number of authors $\hat{n}$. The model obtained F-score 0.2 for intrinsic plagiarism detection, BCubed F-score[3] 0.54 for author diarization with a known number of authors and a BCubed F-score 0.5 for unrestricted diarization.

A. Sittar et al. [2] conducted their experiments on the same corpus (i.e. PAN 16) as [1]. They exploited stylistic features which include lexical attributes to uniquely identify an author's writing style in a given document. In fact, the authors segmented each text document into sentences, and for each sentence, they exploited a total of 15 lexical features (Character n-grams, digits count, spaces count, words count, etc.). Actually, their approach consists of 6 steps. First, the "Read Raw Input Text" in which the authors read all the documents as they are. Second, the "Break Down Text into Sentences" step, in which the authors segmented the documents into sentences. Third is the "Lexical Features Computation" step, in which they counted the ratio of each feature in each sentence. The forth step is the "Distance Calculation", the authors computed the distance for each sentence. The fifth step is "ClustDist[4] Computation [15]" which is also calculated for each sentence. And the final step is "Generating Clusters" in which, on the basis of the scores obtained in the fifth step, the authors applied K-Means algorithm for clustering their data.

In their experiments, they created a matrix V of order n x p. Each matrix row shows a vector of features for each sentence. In the training phase, their approach performed

[3] The BCubed F-measure is a measure defined for non-overlapping clustering. It is like the regular F-Score; but the BCubed algorithm calculates the precision and recall numbers for each entity in the document.

[4] ClustDist is a straightforward technique to compute the average distance from one portion (i.e. sentences) of text to all other pieces of text.

well with sentences of length 7. In fact, sentences of length 5 demonstrated better results for the author diarization with known number of authors subtask and in the unrestricted diarization subtask. Following the results obtained in the training phase, the authors used only sentences with lengths that demonstrated the highest results for each subtask, which are as follows: 7 for the first subtask and 5 for the second and third subtask.

As mentioned in [10], these are the obtained result by Kuznetsov et al. [1] and Sittar et al. [2] in the PAN 16 competition for intrinsic plagiarism detection:

Table 1. Intrinsic Plagiarism Detection Results [10]

		Micro			Macro		
Rank	Team	Recall	Precision	F	Recall	Precision	F
1	Kuznetsov et al.	**0.19**	**0.29**	**0.22**	**0.15**	**0.28**	**0.17**
2	Sittar et al.	0.07	0.14	0.08	0.10	0.14	0.10

N. Akiva [7] treated the problem of intrinsic plagiarism detection, which was the center of interest in the competition PAN 2011. The author's approach consisted of two phases: *chunks clustering* and *cluster properties detection*. For the first step, for a given document, the author divided the text into chunks consisting of 1000 characters. Then, he identified the 100 rarest words that appear in at least 5% of the fragments. Afterwards, the author created a numerical vector representing each chunk, its length is 100 and it corresponds to the presence or absence of the rare words in the fragment. The similarity between pairs is then measured using the Cosine metric. For the clustering, the author used a spectral clustering method called n-cut [14] for clustering the chunks. Later on, the author clustered the document to two parts only (true text and plagiarized text). For the second step, which purpose is to identify clusters that comprise plagiarized parts, the author ran the clustering algorithm on the training corpus and measured a variety of properties which include the relative and absolute size of each cluster, the similarity of each chunk to its own cluster, to the other clusters and to the whole document. Afterwards, the author represented each chunk in the training set as a numerical vector. Then, he used a supervised learning algorithm to learn decision trees[5] to distinguish plagiarized segments from non-plagiarized segments. The author utilized ten-fold cross-validation in order to optimize parameter settings and to estimate accuracy results. Actually, the author didn't exploit the full training set for efficiency reasons. The author ignored all documents with a percentage of plagiarism greater than 40%, and then randomly selected fragments from the remaining documents. On the PAN 11 evaluation set, the author achieved a precision of 12.7% and a recall of 6.6%.

S. Rao et al. [6] conducted their experiments on the same corpus (i.e. PAN 11) as [7]. They focused on features that model the author style (character n-grams, word fre-

[5] A decision tree consists of nodes and branches to partition a set of samples into a set decisions. The starting node is also known as the root of the tree. In each node, a single test or decision is made to obtain a partition. In the terminal nodes or leaves, a decision is made

quencies, means sentence length, stem suffixes frequency, closed class words frequency and frequency of discourse markers[6]. In fact, the authors combined their features to obtain better results in identifying the author style. Their approach first calculates the distance between two normalized feature vectors: the first one is composed of the whole document whereas the second one represents the partially overlapping sections of the documents of 2000 characters window with 200-step size. All the sections for which the style change value comes out to be greater than 2.0 are marked as plagiarized. Consecutive plagiarized sections that are 500 characters apart are merged to form a single plagiarized case to maintain a proper granularity value. Then, the authors measured the style change function distance between normalized stylometric feature vectors by exploiting a style change function. The corpus used has a total of 4753 documents for intrinsic setting. Their obtained result were mediocre and this was due to the low recall values and large number of false positive detection. Nevertheless, discourse markers based features along with the other traits exploited by the authors were successful in detecting intrinsic plagiarism.

G. Oberreuter and J. D. Velásquez [4] also treated the problem of plagiarism detection by detecting deviations in the writing style. They first preprocessed the document, for that they removed all characters leaving only those that belong to the a-z group, all the characters are considered in lowercase. Then, they explored word unigrams considering all the words including the stopwords. Afterwards, they applied a word-frequency-based algorithm to test the self-similarity of a given document. Next, they built for all the words in the document a frequency vector (which is not normalized) and then they clustered the document into groups. At first, the authors created these segments with the use of a sliding window, over the whole document, of length "*m*". For each segment, a new frequency vector is computed, this new vector is explored in further steps, and it's utilized to determine if a segment deviates from the complete document. All segments are classified according to their distance with the document's style. The authors evaluated their approach using the PAN corpora, which is publicly available. For the evaluation of their approach, they used the standard metrics for information retrieval (precision, recall and f-score). The obtained results show the unreliable nature of their approach because the precision is very low (0.3). Actually, their experiments were conducted on documents written in English; however, their approach is not language-dependent.

3 Our Proposed Approach

In this study, we address the intrinsic plagiarism detection problem. In order to identify plagiarism in textual documents, we focused on stylometric features that best describe the writing style, plus we introduced the hybrid aspect (hybridization of lexical and syntactic features). Our proposed approach, as shown in figure 1, comprises five steps.

[6] Discourse markers are words that do not change the meaning of the text. They are either used as filler elements in the text or out of author's habit. They are used frequently and most likely twice every 2 or 3 sentences. Examples of discourse markers are:"well","actually","then", etc.

First, we have the clustering step, in which we grouped documents by writing style. Second, from the obtained clusters we tokenized each document into clusters of 500 characters so that it would be easy to use our features in the following phase. Next, and after creating a vector of features, we constructed a style function by which we determine the designated style for each cluster. Finally, we have the phase of detecting outliers. In fact, each cluster having a deviant style function than the rest of clusters in one single document is detected as an outlier.

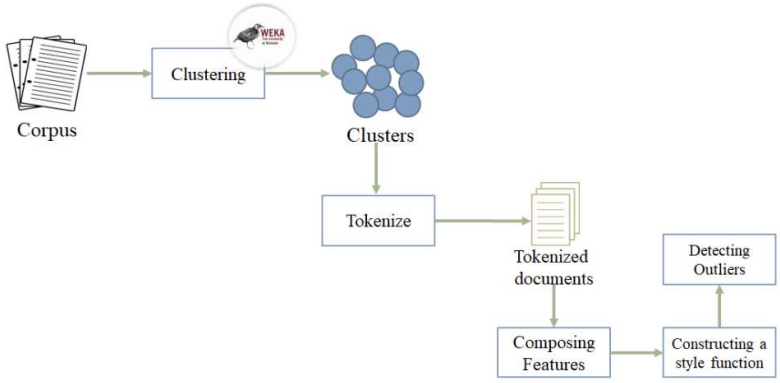

Fig. 1. Main Steps of our Proposed Approach

3.1 Clustering

In this step, we grouped together documents, which are with high probability written by the same author, based on their writing style. Actually, we explored a variety of features in this phase, such as POS tags, means sentence length, function words, type-token ratio, punctuation marks, etc. We also used various classification algorithms such as KNN[7], SVM[8] and decision trees using Weka[9].

3.2 Tokenize

In this step, we parsed the documents and tokenized them by clusters of 500 characters each. Since our approach is inspired by works proposed in the competition PAN@CLEF, we followed the same format demanded in the PAN competition.

---

[7] In k-nearest neighbor (KNN), the nearest neighbor is calculated on the basis of the value of k, that specifies how many nearest neighbors are to be considered to define a class of a sample data point.

[8] SVM is a learning machine for two-group classification problems. It defines a linear decision as an optimal hyperplane with maximal margin between the vectors of two classes.

[9] http://www.cs.waikato.ac.nz/ml/weka/

3.3 Composing Features

In this step, we vectorized text sentences and constructed feature description. We first used our features separately then we explored the hybrid aspect. From the obtained results, we constructed a vector of features.

3.4 Constructing a Style Function

In this phase, and after creating our vector, we constructed the style function. Actually, an author style function is generated as an output of a classifier trained on basic features [1]. In this step, we attribute to each cluster generated in the second step a style function.

3.5 Detecting Outliers

In the final phase, we tried to detect outliers. In each document, we inspected each cluster created in phase 2; each style appearing to be different from the other styles within the same document will be marked as plagiarized. In fact, we explored the KNN algorithm in this step; given a document, the KNN algorithm will segment the document into fragments based on the writing style.

4 Experiments

We tested our approach on the PAN 16 and the PAN 17 corpora for the task of author diarization and style breach. We conducted several individual experiments testing our features separately and combined.

4.1 Dataset Description

The original problem of intrinsic plagiarism detection is related to the question, whether an author has misused parts of a text from others without proper references, and if yes, which parts are plagiarized. Thus, in a given document, the writing style needs to be analyzed to identify the authors [10].

The PAN 16 Corpus:

The task at PAN 16 focuses on identifying authorships within a single document. Thereby, the task is not only focused on searching for plagiarism, but also to identifying contributions of different authors in a given document. The former is the case, where it can be assumed that the main text is written by one author and only some fragments are by other writers. The latter is the case, where in a single document, there exists multiple authors. Such documents may be the result of a collaborative work which is known as "author diarization" (e.g. a combined master thesis written by two students or scientific papers written by a known number of cooperating researchers.) Author diarization consists of three subtasks: Intrinsic plagiarism detection, diarization with a known number of authors and unrestricted diarization. For all three subtasks, distinct training and test

datasets have been provided, which are based on the Webis-TRC-12 dataset [11], with 150 topics from TREC Web Tracks from 2009-2011, whereby professional writers were hired to compose a single document on a given topic. In fact, from the written documents, the datasets for the three subtasks have been generated by varying several configurations such as the number of authors in a text and their respective contributions, the decision if the authors are uniformly distributed or if switches are permitted within a sentence, at the end of a sentence, or only between paragraphs, etc. Since the training set has been partly published, the test documents are created only from unpublished documents. Overall, the number of training/test documents for the respective subtasks are 71/29 for traditional intrinsic plagiarism detection, 55/31 for diarization with a given number of authors, and 54/29 for unrestricted diarization [10].

The PAN 17 Corpus:

The PAN 17 corpus provides documents written in English. The PAN 17 task focuses on detecting style breaches within documents, i.e. to locate borders where authorships change. Therefore, it deals with the task of text segmentation, however; it does not focus on detecting switches in the topic. Thus, given a document, the task is to identify whether the document is multi-authored, and if yes, the borders where authors switch should be determined. The documents provided in this corpus may contain zero up to arbitrarily many style breaches. Thereby, switches of authorships may only occur at the end of sentences and not within them[10].

4.2 Results

Our approach achieved good results with both corpora; Table 2 compares the results of our approach obtained with the PAN 16 and PAN 17 corpora. As evaluation measures, we used precision, recall and f-measure. Actually, we did several benchmarking tests using the PAN@CLEF 2016 and PAN@CLEF 2017 corpora. Figure 2 illustrates the performance of our features compared to those of the hybridization. It is clear that the hybrid aspect gives great result compared to exploring the traits separately.

Table 2. Dataset Results

Corpus	Precision	Recall	F-score
PAN16	0.748	0.635	0.686
PAN17	0.701	0.6	0.646

[10] For more information visit the site of PAN17 : http://pan.webis.de/clef17/pan17-web/author-identification.html

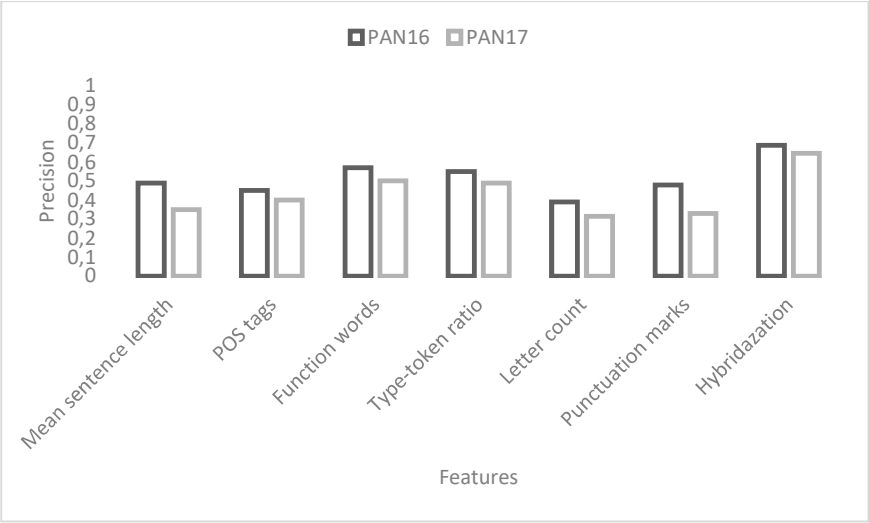

Fig. 2. Performance of Features

Figure 2 shows the performance of the features exploited in our experiments based on their precision. It is obvious that our traits performed well with the PAN 16 corpus than with the PAN 17 corpus and that the hybridization has a better performance than the features used separately.

5 Conclusion

In this paper, we proposed our approach for the task of intrinsic plagiarism detection. We explored a hybrid approach to optimize the performance; we combined various features (stylistic and syntactic attributes) for the construction of a style function. The experiments focused on this exploration were performed on two corpora comprising documents in English. We explored different aspects in our work such as the exploration of the KNN classifier to detect the outliers in a given document and feature hybridization. We obtained good results that outperform the ones obtained by the best state-of-the-art methods; the method achieved an f-score of 0.686 for the PAN 16 corpus and an f-score of 0.646 for the PAN 17 corpus.

As future works, we intend to experiment on other features such as n-grams. Moreover, in our approach we only considered texts written in English, therefore we would like to improve our approach so it would be language independent.

References

1. Kuznetsov, M., Motrenko, A., Kuznetsova, R. and Strijov, V.: Methods for Intrinsic Plagiarism Detection and Author Diarization (2016)
2. Sittar, A., Iqbal, H. R. and Nawab, R. M. A.: Author Diarization using Cluster-Distance Approach (2016)
3. K, V. and Gupta, D.: Detection of Idea Plagiarism using Syntax-Semantic Concept Extractions with Genetic Algorithm (2016)
4. Oberreuter, G. and Velásquez, J. D.: Text Mining Applied to Plagiarism Detection: The use of Words for Detecting Deviations in the Writing Style (2013)
5. Miro, X. A., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G. and Vinyals, O.: Speaker diarization: A review of recent research (2012)
6. Rao, S., Gupta, P., Singhal, K. and Majumder, P.: External & Intrinsic Plagiarism Detection: VSM & Discourse Markers based Approach, Notebook for PAN at CLEF 2011 (2011)
7. Akiva, N.: Using Clustering to Identify Outlier Chunks of Text, Notebook for PAN at CLEF 2011 (2011)
8. Magooda, A., Mahgoub, A. Y., Rashwan, M., Fayek, M. B. and Raafat, H.: RDI System for Extrinsic Plagiarism Detection (RDI_RED) Working Notes for PAN-AraPlagDet at FIRE 2015 (2015)
9. Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B. and Potthast, M.: Clustering by authorship within and across documents (2016)
10. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M. and Stein, B.: Overview of PAN'16: New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation (2016)
11. Potthast, M., Hagen, M., Völske, M. and Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: Proceedings of ACL 13. ACL. (2013)
12. Zechner, M., Muhr, M., Kern, R. and Granitzer, M.: External and Intrinsic Plagiarism Detection Using Vector Space Models (2009)
13. Keogh, E., Chu, S., Hart, D. and Pazzani, M.: Segmenting Time Series: A Survey and Novel Approach (2004)
14. Dhillon, I. S., Guan, Y. and Kulis, B.: Kernel k-means, Spectral Clustering and Normalized Cuts (2004)
15. Guthrie, D.: Unsupervised Detection of Anomalous Text; Ph.D. thesis (2008)