Authorship and Plagiarism Detection Using Binary BOW Features

Notebook for PAN at CLEF 2012

Navot Akiva

The Computer Science Department
Bar Ilan University, Israel Affiliation
navot.akiva@gmail.com

Abstract. Identifying writing style shifts and variations are fundamental capabilities when addressing authorship related tasks. In this work we examine a simplified approach for unsupervised authorship and plagiarism detection which is based on binary bag of words representation. We evaluate our approach using PAN-2012 Authorship Attribution challenge data, which includes both open/closed class authorship identification and intrinsic plagiarism tasks. Our approach proved to be successful achieving overall average accuracy of 84% over and a 2nd place rank in the competition.

Keywords: Intrinsic plagiarism, authorship attribution, outlier detection.

1 Introduction

Authorship and stylistic variation identification over documents has a broad range of applications from identifying specific author's writing style, author verification, plagiarism detection etc. Vast research efforts invested in approaching these areas has been conducted over recent years, by applying various feature representation and algorithmic approaches [1][2]. In this work we evaluate the extent to which a straightforward feature representation method could be successfully utilized for authorship and plagiarism identification.

We consider the problems of authorship identification either open/closed class and intrinsic plagiarism detection, both included in PAN-2012 Authorship Identification competition. For each problem type, we first represent each document as a binary vector that encodes the presence or absence of common words in the text.

2 Our Approach

For all the tasks in this competition we use a single vector representation that captures the presence/absence of common words in a text. We have previously demonstrated the power of this representation elsewhere [3][4] .

In both tasks we are given the authorial chunks' boundaries either over documents (authorship) or paragraphs (plagiarism). The challenge resides in the number of known authors, document length and open/closed class (authorship) as well as short text, ordered/unordered sequences and varying author's number (plagiarism)

2.1 PAN 2012 Authorship Identification Competition

The competition includes 6 tasks for authorship attribution and 2 tasks for intrinsic plagiarism. The tasks description and notation are listed in Table 1.

Table 1. PAN-2012 Authorship Identification competition tasks.

Task Name	Type	Number of Authors	Description
A	Authorship Attribution	3	Short texts, closed class
B	Authorship Attribution	3	Short texts, open class (of task A)
C	Authorship Attribution	8	Short texts, closed class
D	Authorship Attribution	8	Short texts, open class (of task C)
I	Authorship Attribution	12	Novel length texts, closed class
J	Authorship Attribution	12	Novel length texts, open class (of task I)
E	Intrinsic Plagiarism	2-4	Mixed set of paragraphs by individual authors. Paragraphs by any individual author are not guaranteed to be in original order.
F	Intrinsic Plagiarism	2	Consecutive intrusive paragraphs by a single (intrusive) author.

2.2 Supervised Authorship Attribution

For traditional authorship problems (tasks A-D, I-J) we use a supervised learning approach over known authors' documents, using support vector machines as our learning method. We create two separate vectors collections, each containing a single file per known authors, and then in turn we used each as train/test data over the other one respectively, using linear SVM-Light (default setting) modeling.

For the open-class task J where the training examples are novel-length we use the unmasking method of [5], in which different feature representations are applied for evaluating candidate authors' separation robustness. Unmasking distinguish between 2 authors by iteratively training a model and then deliberately impairing it by removing the most discriminative features between the two texts. [5] identified "degradation curves" patterns representing accuracy drop thorough a model's iterations, which drops faster for same-author cases than for different authors.

For the open-class tasks B and D where the texts are short we use the impostors method described in [6], in which similarity of a snippet to random impostor texts is used as a baseline. The method produces artificial impostors for known authors and iteratively tests a suspected text's chunks similarity with all impostors, using different feature set each time. The author of the suspected text is determined by the majority similar author-labels to its chunks.

2.3 Plagiarism Detection

For clustering/plagiarism problems (tasks E and F), we treat each paragraph as a separate document and apply the n-cut clustering algorithm described in [7].

In Task E the number of authors varies between 2 to 4 and there is no assumption of sequenced plagiarized paragraphs. Therefore we cluster the data using number of clusters K=2,3,4 and select the optimal cluster count (K) by applying an inner-cluster similarity and centroids' dissimilarity measures as a convergence criteria.

In task F there one plagiarized author and additional assumption of continuous plagiarized paragraphs sequence is provided. We cluster the data into K=2,3 clusters and look for the maximal original sequences grouped by a single cluster in the 2-clusters result that remain grouped over the 3-clusters sequence. The maximal consistent sequence (which belongs to the minority chunks at K=2) is presumed to be the plagiarized one.

3 Training and Evaluation Results

3.1 Authorship Attribution

Closed Class

Table 2. Closed Class authorship results.

Task Name	Average Training Accuracy	Evaluation Results
A	100%	100%
C	65%	75%
I	97%	86%

In task C there are more candidate authors to "confuse" than task A, where the training is based on a single example only and the train/test documents of the same author are different in length. Therefore one of the train experiments resulted in 87% accuracy while the other achieved only 50%. Nonetheless, when measuring the gap

between the SVM scores of the top class and the 2nd top one (in %) - all misclassified test examples were below average of all the gaps.

Open Class

Table 3. Open Class authorship results.

Task Name	Evaluation Results
B	90%
D	65%
J without Unmasking	88%
J with Unmasking	100%

For tasks B and D we initially apply a learned classifier to the test examples, based on the assumption that actual known author's test examples would get the correct label. To eliminate examples not authored by any of the candidates, we then apply the impostor method introduced in [6].

Instead of producing an auxiliary artificial impostor data, we utilize all 11 known authors (of both tasks A and B) as impostors for identifying outliers only, while ignoring the non-outlier labels.

We chunk each author's texts into 1000 words for providing the test examples chunks the option to "confuse" their labelled class and suggest inconsistency. Taking into account the option that there is no guarantees that all 11 authors are unique ones, as well as the lower number of potential "impostors" (at least for task C) we decide the following:

1. Use this method for validating "outliers" existence only and ignore the resulting coherent labels of the other ones.

2. Repeat the impostor experiment per task using its original known authors and looks for labelling consistency when expanding the experiment with all 11 authors. In the case where all chunks of a test example were coherently labelled by both experiment as a single train author (of its original task) then the labelling is valid, otherwise considered as an outlier.

While applying the methodology above we discover that there were a couple of test examples (one per task) which got a coherent label on the tasks' known authors experiment and another coherent label at the expanded experiment, both from the other task's know authors. This fact suggested that potentially these "switching" authors are the same one and in order to validate that (and to not eliminate these examples) we re-ran both tasks experiments and eliminate one of these authors respectively. This time the labels were consistent which assessed our assumption and we did not considered these examples as outliers

For task J we apply the unmasking method for each classification results where the difference between the first 2 top SVM scores is below average. The unmasking algorithm revealed 2 cases where the predicted class was not the correct author and therefore these items were marked as 'None'. The overall accuracy is thus raised from 88% to 100%.

3.2 Intrinsic Plagiarism

Table 4. Intrinsic plagiarism results.

Task Name	Evaluation Results
E	76%
F	89%

For task F we create a train set of example documents with similar properties as the provided test files (~200 words per paragraph, 20 paragraphs per file) over 15 "Guttenberg project" books written by different authors. Each example is attached with a plagiarism rate of 0-40%. For each of these we measure its plagiarized sequence consistency between K=2 and K=3 clustering results.

Table 5. Consistent chunks mapping of the train set (above 60%).

	Plagiarism Rate			
	10%	**20%**	**30%**	**40%**
% of documents	57%	70%	69%	77%

Table 5 shows the percentage of documents, grouped by their plagiarism rate, in which the original plagiarized chunks are consistently mapped to a single K=3 cluster with purity level above 60%. In all the plagiarism levels above the purity of 50% exists for 95% of the documents. The test evaluation results presented in Table 4 show the robustness of our method for task F, achieving a high accuracy level of 89%.

In Task E the number of authors is larger and varies between 2 to 4. In addition there is no assumption of sequenced plagiarized paragraphs. Our approach of seeking optimal cluster number among K=2,3,4 by applying an inner-cluster similarity criteria obtained overall accuracy of 76%. The obtained accuracy level is lower than the one in task F due to the larger number of participating authors as well as the lack of sequential assumption. Nonetheless this result is surprisingly high considering the task's complexity and the straightforward representation and algorithmic approach we've applied for capturing individual paragraph's authorship source.

4 Conclusion and Future Work

In this paper we present an evaluation of the obtained accuracy level for authorship related tasks by using a simplified representation of binary bag of words approach. Our evaluation over the PAN-2012 Authorship Identification challenge data is split over authorship identification (open/closed class) and author clustering/intrinsic plagiarism. We show that binary BOW representation works quite well for capturing authorship for all the tasks. There are a couple of factors which affects the accuracy levels: number of participating authors and the length of the examined documents. Our method appears to be effective for both long and short texts but more sensitive towards high number of authors. For this reason the authorship tasks C and D and plagiarism task E has somewhat lower accuracy level than the other, though still solid. Therefore an immediate future work would be to improve achieved accuracy results for a high number of authors for both tasks.

References

1. Argamon, S., Juola, P., Overview of the International Authorship Identification Competition at PAN-2011. CLEF (Notebook Papers/Labs/Workshop). (2011).
2. Stamatatos, E., A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3):538–556. (2009).
3. Koppel, M., Akiva, N., Dershowitz, I. and Dershowitz, N., Unsupervised Decomposition of a Document Into Authorial Components, Proceedings of ACL, Portland OR, June 2011, pp. 1356-1364.
4. Koppel, M. and Akiva, N., Identifying Distinct Components of a Multi-Author Document, Proceedings of EISIC 2012, to appear.(2012).
5. Koppel, M., Schler, J. and Bonchek-Dokow, E., Measuring Differentiability: Unmasking Pseudonymous Authors, JMLR 8, pp. 1261-1276. (2007).
6. Koppel, M., Schler, J.,, Argamon, S. and Winter, Y. : The “Fundamental Problem” of Authorship Attribution, English Studies, 93:3, 284-291 (2012).
7. Dhillon, I. S. , Guan, Y. and Kulis, B., "Weighted graph cuts without eigenvectors: a multilevel approach", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 29(11):1944-1957.(2007)