Authorship identification in large email collections: Experiments using features that belong to different linguistic levels

Notebook for PAN at CLEF 2011

George K. Mikros[1] and Kostas Perifanos[2]

[1]Department of Italian Language and Literature, National and Kapodistrian University of Athens, Greece
gmikros@isll.uoa.gr
[2]Department of Linguistics, National and Kapodistrian University of Athens, Greece
kperifanos@phil.uoa.gr

Abstract The aim of this paper is to explore the usefulness of using features from different linguistic levels to email authorship identification. Using various email datasets provided by PAN'11 lab we tested several feature groups in both authorship attribution and authorship verification subtasks. The selected feature groups combined with Regularized Logistic Regression and One-Class SVM machine learning methods performed well above average in authorship attribution subtasks and below average in authorship verification subtasks.
Keywords: authorship attribution, authorship verification, stylometry, Regularized Logistic Regression, LIBLINEAR, LIBSVM, One-Class SVM, Support Vector Machines

1 Introduction

Authorship identification refers to the connection of a text of unknown authorship to a specific author using a set of quantifiable text features as indicators of the author's style. Since the late 1990s authorship identification has known a new impetus based on developments in a number of key research areas such as Information Retrieval, Machine Learning and Natural Language Processing [12]

The authorship identification dataset provided by the 5^{th} International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse PAN'11 provided a test bed for comparing different strategies in both feature selection and classification algorithms. Our approach to authorship identification is based mainly on the idea that an author's style is a complex multifaceted phenomenon affecting the whole spectrum of his/her linguistic production. Following the old theoretical notion of "double articulation" of the Prague School of Linguistics we accept that stylistic information is constructed in blocks of segments of increasing semantic load, from character n-grams, to word n-grams. In order to capture the multilevel manifestation of stylistic traits we should detect these features, which belong to many different linguistic levels, and utterly combine them for achieving the most accurate representation of an author's style.

2 Features and classification algorithms

Authorship identification research has used an impressive array of stylometric features ranging from characters to syntactic and semantic units. We selected our features taking into consideration the best practices established in authorship identification research published from the 1990's till today [12,4,6]

As mentioned above we decided to focus on features that cover a wide range of linguistic levels and at the same time are easy to implement and are language independent. We used five single feature groups and in a later stage we combined them in a feature group labeled "All", a methodology that gave us the best results in the validation set and is generally accepted as better strategy [6,13]. In all our features we normalized their frequency in relation to the text length. In order to construct the "All" feature group we used the 1000 most frequent features from each single feature group in the training corpus resulting in a total vector of 5000 features. The single feature groups we combined are described below:

- Character Bigrams (cbg): Character bigrams provide a robust indicator of authorship and many studies have confirmed their superiority in large datasets e.g. [7].

- Character Trigrams (ctg): Character trigrams capture significant amount of stylistic information and have the additional merit that they also represent common email acronyms like FYI, FAQ, BTW, etc.

- Word Unigrams (ung): Word frequency is considered among the oldest and most reliable indicators of authorship outperforming sometimes even the n-gram features [1,3].

- Word Bigrams (wbg): Word bigrams have long been used in authorship attribution with success e.g. [3].

- Word Trigrams (wtg): Word trigrams have also been found to convey useful stylistic information [5,10] since they approach more closely the syntactic structure of the document.

Character n-grams approach phonology and morphology capturing quantitative information regarding syllable structure, phonotactics, consonant clusters, prefix and suffix structure. Word n-grams on the other hand approach syntax organization including different lexical bundles, phrases, collocation structures among others.

The most frequent unigrams were detected using a custom PERL script which identified tokens as a sequence of alphanumeric characters using the regular expression \w+. Later a custom PERL script took as input a list of the most frequent tokens in the training corpus and produced a vector containing text length normalized frequency of occurrence of each token in all the texts contained in the datasets.

The most frequent n-grams were detected using the Ngram Statistics Package (NSP) [2], a PERL module designed word and character n-gram identification. Tokenization in n-gram identification followed the following rules:

– Token was identified any sequence of alphanumeric characters using the following regular expression: \w+

– As tokens were identified also the punctuation marks defined in the following regular expression: [\.,;:\?!]. Punctuation usage often reflects author-related stylistic habits [8] and n-grams with punctuation can capture better possible these stylistic idiosyncrasies.

– All tokens were converted to lowercase.

Output files from NSP were converted to vectors using custom PERL script which aggregated n-gram counts from each text file and normalized their frequency to the text length. Given the rather huge dimensionality of the extracted features, the task of training models is time and memory consuming, even for moderate number of training instances. Therefore a method for solving efficiently large scale classification problem was required.

For the purposes of the authorship attribution tasks, we used Regularized Logistic Regression (RLR) as implemented in LIBLINEAR [9], a relatively new and highly efficient package for classification tasks. For the case of verification tasks we used One-Class Support Vector Machines proposed by Schölkopf et al. [11] which is provided by the LibSVM package.

3 Evaluation

3.1 Authorship identification task

In the Authorship identification task we trained our classifier (RLR) using the default values ($Costparameter = 1$, $Epsilon = 0.01$). In order to evaluate the performance of our classifier in the training set we used both accuracy and F_1 averaged in 10-fold cross-validation of the testing sample. The trained model we obtained from this procedure was used for prediction in the LargeValid dataset. Its performance was also measured using accuracy by comparing the predictions made by the classifier with the labels provided by the GroundTruthLarge and Small Valid files. We compared different feature groups both single and in different combinations. The best results obtained from the "All" feature group which contained the 1000 most frequent features from each single feature group. The results from our experiments in LargeTrain, LargeValid and SmallTrain, SmallValid datasets are shown in the following table. We report only the "All" feature group and the single group results since the number of combinations we examined was large (25 feature group combinations for each dataset):

Table 1. Cross-Validated Accuracy and micro-averaged F_1 of our classifier in Large and Small Train and Valid data

Features	LargeTrain		LargeValid		SmallTrain		SmallValid	
	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
All	**0.481**	**0.465**	**0.51**	**0.498**	**0.683**	**0.662**	**0.674**	**0.64**
wbg	0.32	0.303	0.352	0.331	0.576	0.551	0.56	0.535
wtg	0.281	0.256	0.294	0.263	0.502	0.472	0.504	0.466
cbg	0.26	0.246	0.269	0.248	0.423	0.407	0.43	0.406
ctg	0.312	0.293	0.338	0.321	0.519	0.49	0.512	0.476
ung	0.322	0.311	0.347	0.334	0.59	0.568	0.57	0.541

From the above table it is obvious that combining the feature groups we get the best classification accuracy over all the datasets. In order to ensure further the superiority of the combined feature group we conducted a series of pairwise t-tests comparing the "All" feature group with each of the single feature groups. Since we had multiple comparisons a Bonferroni correction was applied to the p level of significance ($p = 0.01$) of all the t-test conducted. In all the comparisons employed, the "All" feature group obtained a statistical significant better classification accuracy and F_1 over each one of the single feature groups providing support to our claim that a combined feature group consisting of features from multiple and different linguistic levels capture more efficiently an author's style.

3.2 Authorship verification tasks

In authorship verification tasks (+ datasets) two subtasks were defined:

– Combined authorship attribution and verification: In this subtask the aim was to find which of the given texts in the LargeValid+ and SmallValid+ were written from authors within the corresponding training set and which from external authors. In the first case we trained a One-Class SVM model with RBF kernel using the LargeTrain and the Small Train datasets as one class. The trained model was applied to the LargeValid+ and SmallValid+ datasets in order to identify the texts that were written from the authors of the training sets. All "unknown" cases were assigned the label "unknown" and they were removed from the Valid+ datasets. Then we applied the previously trained models from the authorship identification subtask to the reduced Valid+ datasets and performed authorship attribution. The final results obtained from this procedure are shown in the Table 2 below:

Table 2. Accuracy and F_1 of our classifier in the SmallValid+ and LargeValid+ datasets

Features	LargeValid+		SmallValid+	
	Acc	F_1	Acc	F_1
	0.352	0.339	0.676	0.659

– Authorship verification task: The aim of the second subtask was to find if 3 specific authors (Verify1, 2 and 3 datasets) had written any and what texts from the provided validation datasets (Verify1+, 2+ and 3+). The procedure followed in this case was

the training of One-Class SVM model using RBF kernel and its subsequent application to the respective validation datasets. Furthermore, since our training set was small we used only the 2000 most frequent character bigrams of the training set in order to train our classifier. The final results from this procedure are shown in Table 3:

Table 3. Accuracy of our classifier in the Verify datasets

Dataset	Verify1+Valid	Verify2+Valid	Verify3+Valid
	1	1	1

The accuracy of our classifier in the validation datasets reached 1 in all datasets.

4 Conclusion

The Author Identification competition organized by the PAN 2011 Lab was an interesting and challenging task in which we had the opportunity to test the usefulness of both features and machine learning methods in a variety of authorship attribution and verification scenarios. Our features covered a wide range of linguistic levels, from sub-word entities (character bigrams, trigrams) to word and hyper-word formations (word unigrams, bigrams and trigrams). In the authorship attribution subtask we used the above mentioned features combined with the Regularized Logistic Regression. This approach scored well in all the performance indices except the macro-averaged precision probably due to the large dimensionality of our solution. In total our system ranked in the 5^{th} (out of 13 groups) and 3^{rd} (out of 12 groups) position in the LargeTest and SmallTest dataset correspondingly.

On the subtasks of authorship verification our approach scored below the average performance of the participating research groups. In the combined scenario of authorship attribution and verification we trained a One-Class SVM in the training datasets with the same features used in authorship attribution subtask in order to identify and exclude the texts that weren't part of the training set. Then we performed an authorship attribution to the remaining texts. This approach was ranked in the 6^{th} (out of 9 groups) and 5^{th} (out of 9 groups) position in the LargeTest+ and SmallTest+ dataset correspondingly. In the plain authorship verification task we trained a One-Class SVM using the 2000 most frequent character bigrams with little success in the respective datasets. Our approach obviously suffered from overtraining since we obtained high recall but low precision values in Verify2+Test and Verify3+Test datasets. More specifically our system ranked in the 4^{th} (out of 7 groups), the 7th (out of 7 groups) and 5^{th} (out of 7 groups) position in the Verify1+Test, Verify2+Test and Verify3+Test dataset correspondingly. We believe that some of the factors affecting the performance of our system in the verification tasks were the usage of the LIBSVM One-Class algorithm combined with the usage of the character bigrams in the plain verification subtask. There is a constant decline of our system's performance when we employ LIBSMV One-Class in the combined subtask (Large+, Small+ datasets) and a further decrease when we shrink our features to character bigrams.

The results obtained from the authorship attribution subtask are encouraging and support our claim that authorship is based on textual features that are scattered in a wide spectrum of linguistic levels. Future research will be directed to detect features from other linguistic levels and use them in attribution tasks taking into consideration not only their frequency but also their discriminative power especially in small classes in order to improve our macro-average performance indices. Furthermore, we will continue our experimentation in the verification subtask with different one-class learning algorithms and varying feature groups.

References

1. Allison, B., Guthrie, L.: Authorship attribution of E-Mail: Comparing classifiers over a new corpus for evaluation. European Language Resources Association (ELRA), Marrakech, Morocco (2008)
2. Banerjee, S., Pedersen, T.: The design, implementation, and use of the ngram statistic package pp. 370–381 (2003)
3. Coyotl-Morales, R., Villaseñor-Pineda, L., Montes-y Gómez, M., Rosso, P.: Authorship Attribution Using Word Sequences, Lecture Notes in Computer Science, vol. 4225, pp. 844–853. Springer Berlin / Heidelberg (2006)
4. Grieve, J.W.: Quantitative authorship attribution: an evaluation of techniques. Literary and Linguistic Computing 22(3), 251–270 (2007)
5. Guzmàn-Cabrera, R., Montes-y Gómez, M., Rosso, P., Villaseñor-Pineda, L.: A Web-Based Self-training Approach for Authorship Attribution, pp. 160–168. Springer-Verlag, Berlin, Heidelberg (2008)
6. Juola, P.: Authorship attribution. Foundations and Trends® in Information Retrieval 1(3), 233–334 (2008)
7. Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. Literary and Linguistic Computing 26(1), 35–55 (2011)
8. Mikros, G.K.: Stylometric experiments in Modern Greek: Investigating authorship in homogeneous newswire texts, pp. 445–456. Mouton de Gruyter, Berlin / New York (2007)
9. R.-E. Fan, e.a.: Liblinear: A library for large linear classication. Journal of Machine Learning Research 9, 1871–1874 (2008)
10. S. Raghavan, e.a.: Authorship attribution using probabilistic context-free grammars. In: for Computational Linguistics, A. (ed.) Proceedings of the ACL 2010 Conference Short Papers. pp. 158–164 (2010)
11. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Tech. Rep. MSR-TR-99-87, Microsoft Research (1999)
12. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
13. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 57(3), 378–393 (2006)