Baseline Approaches for the Authorship Identification Task

Notebook for PAN at CLEF 2011

Darnes Vilariño, Esteban Castillo, David Pinto, Saul León, and Mireya Tovar

Faculty of Computer Science
B. Universidad Autónoma de Puebla
{darnes, dpinto}@cs.buap.mx, ecjbuap@gmail.com, saul.ls@live.com

Abstract In this paper we present the evaluation of three different classifiers (Rocchio, Naïve Bayes and Greedy) with the aim of obtaining a baseline in the task of authorship identification. We decided to employ as features the original words contained in each document of the test set, with a minimum of preprocessing which included elimination of stopwords, punctuation symbols and XML tags. As may be seen in this paper, the obtained results are adequate, reflecting the aim of the experiments. In average, Rocchio slightly outperformed the Naïve Bayes and the Greedy classifier. However, we recommend using both, Rocchio and Naïve Bayes in future evaluations of the PAN competition as baselines from which other teams may compare their own approach.

1 Introduction

Authorship identification is the task of determining the real author of a given text. Nowadays, there exist many texts that have been written anonymously or under false names which lead to confusion on the identification of their authorship. The main challenge of authorship identification is to automatically assign a text to one of a set of known candidate authors. The experiments reported in this paper were carried out in the framework of the 5th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN'11).

Given texts of uncertain authorship and texts from a set of candidate authors, the task of "Authorship identification" of PAN'11 consists of mapping the uncertain texts onto their true authors among the candidates. For this purpose, the organizers of the aforementioned task have developed a new authorship evaluation corpus which is better described in [2].

We have tackled this problem through the use of supervised learning methods. The aim of this paper is to show the performance of three different methods in order to determine whether or not they may be used as baselines in future evaluations.

The rest of this document is structured as follows. Section 2 present the pre-processing techniques applied to the given corpus and a description of the classifiers used. Section 3 presents a discussion of the obtained results. Finally, in Section 4 the conclusions are given.

2 Authorship features and classification

The aim of our first participation in PAN'11 was to obtain baselines for the task of authorship identification. Therefore, we decided to employ as features the original words contained in each document of the test set, with a minimum of preprocessing which included elimination of stopwords, punctuation symbols and XML tags. Future evaluations will consider appropriate features which may improve the classification task. The description of the datasets used and the classifiers evaluated follows.

2.1 Datasets

Five training collections consisting of real-world texts (i.e., often short and messy) were given to the task participants. The first one with 26 different authors, the second one with 72 different authors, and the remaining three each with a single author (for author verification). In the first two cases, there were two testing scenarios, represented by a different test set: one with only authors from the training (not necessarily all), and one with authors from the training and from outside the training. For the verification task, each test set included documents from the author to be verified and some documents from other authors as well. A complete description of the rationale on the construction of these corpora is given in [2].

2.2 Classification methods

Supervised learning methods are used to calculate the most likely class given a set of features. We are given a training set D of labeled documents $< d, a >$, where $d \in D$ is set of documents, and $a \in A$ is the set of authors, in order to obtain the classification model.

In this paper we evaluated three supervised learning methods (Rocchio, Naïve Bayes and Greedy) which are described as follows.

Rocchio Rocchio classification is a form of Rocchio relevance feedback. The average of the relevant documents, corresponding to the most important component of the Rocchio vector in relevance feedback, is the centroid of the "class" of relevant documents (the authorship centroid in our case). We omit the query component of the Rocchio formula in Rocchio classification since there is no query in text classification. Rocchio classification can be applied to classes whereas Rocchio relevance feedback is designed to distinguish only two classes, relevant and nonrelevant.

The centroid calculation formulae is shown in Eq.(1) and the model training is depicted in algorithm 1.

$$\overrightarrow{\mu(a)} = \frac{1}{|D_a|} \sum_{d \in D_a} \overrightarrow{v}(d), \tag{1}$$

where D_a is the set of documents in D whose author is $a : D_a = \{d :< d, a >\in D\}$, and $\overrightarrow{v}(d)$ is the normalized vector of d.

In this approach we have used a vectorial document representation with a TF-IDF weighting schema[3]. The classification criterion presented in algorithm 2 uses the Euclidean distance.

Algorithm 1: Rocchio algorithm for the training step

Input: A : Set of authors; D : Set of documents
Output: Centroids of each author document set

foreach $a_j \in A$ **do**
  $D_j \leftarrow \{d :< d, a_j >\in D\}$;
  $\overrightarrow{\mu}_j \leftarrow \frac{1}{|D_a|} \sum_{d \in D_a} \overrightarrow{v}(d)$;
end
return $\{\overrightarrow{\mu}_1, \cdots, \overrightarrow{\mu}_J\}$

Algorithm 2: Rocchio algorithm for the testing step

Input: $\{\overrightarrow{\mu}_1, \cdots, \overrightarrow{\mu}_J\}$: Centroids; d : Document of being classifed
Output: The best class (author) for the input document

return *arg min*$_j |\overrightarrow{\mu}_j - \overrightarrow{v}(d)|$

Naïve Bayes We have used a probabilistic supervised learning method named multinomial Naïve Bayes in order to determine the authorship attribution (as described in [1]). The probability of a document d being written by author a is computed as shown in Eq.(2).

$$P(a|d) \approx P(a) \prod_{1 \leq k \leq n_d} P(t_k|a) \tag{2}$$

where $P(t_k|a)$ is the conditional probability of the k-th term (t_k) occurring in a document written by author a. Actually, $P(t_k|a)$ measures the contribution of term t_k so that the document d belongs to class a. n_d is the number of terms in document d. $P(a)$ is the prior probability of a document written by author a. Since we are really interested in finding the best class (author) for the document, we may calculate the maximum a posteriori (MAP) as shown in Eq.(3).

$$a_{map} = \arg\max_{a \in A} P^*(a|d) = \arg\max_{a \in A} P^*(a) \prod_{1 \leq k \leq n_d} P^*(t_k|a) \tag{3}$$

$P^*(t_k|a)$ is estimated by using Laplace smoothing, which simply adds one to each count (See Eq. (4)).

$$P^*(t_k|a) = \frac{T_{at_k} + 1}{\sum_{t' \in V}(T_{at'} + 1)} \tag{4}$$

where T_{at_k} is the number of occurrences of t_k in training documents from class a, including multiple occurrences of a term in a document and V is the corpus vocabulary.

2.3 Greedy

A greedy approach was also employed in the task of authorship attribution. We calculated a similarity matrix among all the documents in the training set. Thereafter, we selected the 100 most similar documents with respect to each document of the test set. The final class (author) is obtained by counting the most frequent class (author) from those 100 documents.

3 Evaluation

In Tables 1, 2, 3 and 4 we may see the obtained results for the three different approaches. Although we have used the tags Rocchio, Naïve Bayes and Greedy for these approaches, in the task description paper [2] they may be reported as *authorship-vilarino-2011-05-31-1456/*, *authorship-vilarino-2011-05-31-1455/* and *authorship-vilarino-2011-05-31-1454/*, respectively.

We have also included include the best run, the worst run and the arithmetic mean of all the runs submitted at the competition, so that the reader may compare the three approaches reported in this paper with respect to the rest runs.

In general, Rocchio obtained a more stable behavior than Naïve Bayes and the Greedy approach. However, it is worth noticed that Naïve Bayes outperformed the Rocchio classifier in terms of macro average precision but lousy macro average recall, which is likely because it will tend to classify most documents into the largest document classes, and so will not do well for the less frequent authors.

By comparing the arithmetic mean (obtained with all the runs submitted to the competition) with respect to the presented approaches, we may suggest that both, Rocchio and Naïve Bayes would be used as baselines in future competitions of authorship attribution. Further analysis of variance is needed in order to confirm this claim. It is important to notice that these classifiers were executed with raw data, i.e., none feature selection was performed and, therefore, we expect that any features analysis would improve significantly the performance of these classifiers.

Table 1. PAN authorship evaluation results with the *LargeTest* corpus

Run	MacroAvg Prec	MacroAvg Recall	MacroAvg F1	MicroAvg Prec	MicroAvg Recall	MicroAvg F1
Best run	0.549	0.532	0.520	0.658	0.658	0.658
Rocchio	0.364	0.337	0.364	0.428	0.428	0.428
Naïve Bayes	0.534	0.095	0.103	0.238	0.238	0.238
Greedy	0.232	0.139	0.147	0.219	0.219	0.219
Worst run	0.021	0.017	0.013	0.056	0.055	0.055
All runs arithmetic mean	0.418	0.266	0.273	0.402	0.389	0.395

In Tables 5, 6 and 7 it is presented the results obtained when the classifiers were executed using corpora attempting to detect only one author. In other words, the task

Table 2. PAN authorship evaluation results with the *LargeTest+* corpus

Run	MacroAvg Prec	MacroAvg Recall	MacroAvg F1	MicroAvg Prec	MicroAvg Recall	MicroAvg F1
Best run	0.688	0.267	0.321	0.779	0.471	0.587
Rocchio	0.347	0.245	0.263	0.368	0.368	0.368
Naïve Bayes	0.488	0.084	0.088	0.222	0.222	0.222
Greedy	0.153	0.092	0.089	0.175	0.175	0.175
Worst run	0.001	0.011	0.000	0.001	0.001	0.001
All runs arithmetic mean	0.430	0.147	0.162	0.407	0.270	0.314

Table 3. PAN authorship evaluation results with the *SmallTest* corpus

Run	MacroAvg Prec	MacroAvg Recall	MacroAvg F1	MicroAvg Prec	MicroAvg Recall	MicroAvg F1
Best run	0.662	0.451	0.475	0.717	0.717	0.717
Rocchio	0.236	0.284	0.358	0.432	0.432	0.432
Naïve Bayes	0.359	0.141	0.157	0.374	0.374	0.374
Greedy	0.150	0.061	0.098	0.091	0.091	0.091
Worst run	0.150	0.061	0.098	0.091	0.091	0.091
All runs arithmetic mean	0.459	0.297	0.303	0.519	0.515	0.517

consisted on detecting whether or not a particular text were written by the given author. All the classifiers obtained a very poor performance in this particular task. We consider that the problem is that we do not have enough term frequencies for determining the discrimination degree of each term. A similar behavior is observed in the *SmallTest* and *SmallTest+* corpus. Besides considering the term frequency, it is important to reduce the noise in the documents as done, for instance, in [4].

4 Conclusion

In this paper we have presented the performance of three supervised learning methods for the task of authorship attribution. All these classifiers were feeded with the original training data, i.e., without considering feature selection techniques of any kind. The purpose was to determine baselines for future comparisons in this task. In general, the Rocchio classifier performed better than the other two evaluated.

A simple manner of improving the obtained results would be by lemmatizing the corpus, so that we would increase the term frequencies and, therefore, the classifiers will be able to determine the discrimination degree of each term.

Table 4. PAN authorship evaluation results with the *SmallTest+* corpus

Run	MacroAvg Prec	MacroAvg Recall	MacroAvg F1	MicroAvg Prec	MicroAvg Recall	MicroAvg F1
Best run	0.737	0.161	0.193	0.824	0.457	0.588
Rocchio	0.200	0.157	0.195	0.349	0.349	0.349
Naïve Bayes	0.371	0.077	0.084	0.301	0.301	0.301
Greedy	0.140	0.030	0.049	0.065	0.065	0.065
Worst run	0.140	0.030	0.049	0.065	0.065	0.065
All runs arithmetic mean	0.527	0.109	0.119	0.506	0.293	0.336

Table 5. PAN authorship evaluation results with the *Verify1* corpus

Run	MacroAvg Prec	MacroAvg Recall	MacroAvg F1
Best run	1.000	0.333	0.500
Naïve Bayes	0.100	0.333	0.500
Rocchio	0.043	0.667	0.900
Greedy	0.033	0.333	0.500
Worst run	0.045	0.333	0.080
All runs arithmetic mean	0.263	0.400	0.363

Table 6. PAN authorship evaluation results with the *Verify2* corpus

Run	MacroAvg Prec	MacroAvg Recall	MacroAvg F1
Best run	0.400	0.800	0.533
Naïve Bayes	0.071	0.400	0.571
Greedy	0.031	0.400	0.571
Rocchio	0.026	0.400	0.571
Worst run	0.035	0.600	0.067
All runs arithmetic mean	0.185	0.400	0.346

Table 7. PAN authorship evaluation results with the *Verify3* corpus

Run	MacroAvg Prec	MacroAvg Recall	MacroAvg F1
Best run	0.211	1.000	0.348
Naïve Bayes	0.091	0.333	0.500
Rocchio	0.037	0.583	0.833
Greedy	0.034	0.333	0.500
Worst run	0.036	0.500	0.067
All runs arithmetic mean	0.078	0.437	0.323

Having seen that Rocchio performs well, we are in conditions of analyzing and exploring different feature selection techniques for improving the authorship identification task.

References

1. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
2. Potthast, M., et al.: Authorship attribution: Task description. In: Same working notes (2011)
3. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
4. Yang, Y.: Noise reduction in a statistical approach to text categorization. In: Proc. of the 18th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR-ACM. pp. 256–263 (1995)