Authorship Identification Using a Reduced Set of Linguistic Features

Notebook for PAN at CLEF 2012

Stefan Ruseti, Traian Rebedea

Department of Computer Science and Engineering
University Politehnica of Bucharest, Romania
stefan_ruseti@yahoo.com, traian.rebedea@cs.pub.ro

Abstract. The proposed solution for authorship attribution combines a couple of the most important features identified in previous research in this domain with classification algorithms in order to detect the correct author. We consider that the most relevant aspect of our work is the small number of linguistic features and the use of the same framework to solve both the open and the closed class authorship problem, by only changing the classification algorithm. This approach obtained an overall 77% accuracy with regard to the total number of correctly classified documents.

1 Introduction

The problem of authorship identification or attribution of text documents has been widely studied in the last decades, especially in the last 20 years, but the solutions are not mature enough to consider the problem solved. Nowadays, the Web offers very large amounts of texts to be used as corpora for authorship identification, but it also provides many different types of discourse that may be analyzed: from narratives and e-mails to online conversations and social network updates. It is obvious that each type of discourse should be treated independently; nevertheless even the problem of identifying the author of large narrative texts is far from being closed.

Authorship attribution may be divided into two different subtasks: determining the most descriptive features of the texts under consideration, then applying a classification algorithm in order to detect the most probable author [1]. The methods for conducting the classification stage range from principal component analysis and cluster analysis to support vector machines (SVMs) and neural networks.

The proposed solution has started from examining the most important features and most powerful classification algorithms developed for PAN 2011 [2]. Of course, as the discourse type has changed from e-mails to short narratives, the feature set also had to be changed as described in the next section. An extensive set of features used for previous works in authorship attribution has also been presented in [3]. The most successful team in PAN 2011 has used a broad set of features, including several specific to e-mail conversations, and a maximum entropy classifier [4]. Another approach used a semi-supervised approach based just on character trigrams and SVM in order to determine documents in the test corpus that had the highest probability of being written by one of the authors. These documents were then added to the training set and the classifiers were retrained [5]. A last interesting approach that offered good results was a voting approach that used several classifiers that might elect of veto directly the author of a document. The results of all the classifiers were combined [6].

The following section describes the small feature set chosen for solving the proposed problem, while section 3 briefly presents the choice of the classifier. Then, we describe the results that were obtained and how the answers for the open problem were derived and wrap up with some conclusions.

2 The Reduced Set of Linguistic Features

In the multitude of approaches to author identification over the last years, a large list of features was used, ranging from character and lexical to the semantic layer. Of course, some of them need to be problem specific [3]. From all these features, we extracted a reduced set that proved to be suitable for the type of discourse in the PAN 2012 corpus. Thus, features related to the layout of the text or spelling errors were irrelevant because the corpus was composed of short narratives (or novels), which are usually written correctly and most times are also edited.

The remaining features used to describe the texts and to solve the classification problem are:

- *Character trigrams* – the most common 100 trigrams from the training corpus were selected and then the distribution of each text has been computed.
- *POS bigrams and trigrams* – the most common 50 bigrams and 100 trigrams were selected. The POS tagging was realized using the RITa POS Tagger[1]. However, as most other taggers, it returned very specific POS tags that did not offer enough generality needed to extract each author's style. For example, *nnps* was for proper noun plural, but only the first letter from these tags was considered sufficient and descriptive for the author's writing style.
- *Suffixes* – the most common 32 English suffixes were counted in each text. The percentage of suffixed words from all words was recorded as well. To check if a word has a certain suffix, we first checked to see if the word was composed using a suffix by using a stemmer. After this test, we only checked if the word ended with a suffix from the list. This approach is not 100% correct, but it had a very small error rate that did not influence the classification.
- *Word length* – word lengths from 1 to 15 were counted, any word longer than 15 characters was considered in the 15 category. These features should capture the author's vocabulary richness.
- *Syntactic complexity and structure* – we used the Stanford parser[2] to create the parsing tree for each sentence and to extract the syntactic dependencies. The average sentence length, the average and the maximum tree depth, the average and

[1] http://www.rednoise.org/rita/documentation/ripostagger_class_ripostagger.htm

[2] http://nlp.stanford.edu/software/lex-parser.shtml

the maximum distance between the elements of a dependency were recorded. Each dependency type was also counted to try to represent the author's predisposition for certain syntactic structures.

- *Percentage of direct speech* – some authors may tend to use more dialogue in their texts than others, so also took this under consideration. Sometimes, this feature can be irrelevant because the type of the text can also determine the percentage of dialogue. In the evaluation stage, this feature increased the overall accuracy, so we decided to use it.

Each feature was normalized so that the lengths of the texts do not interfere with the results. Because there were only 2 training texts for each author, we split each one into smaller pieces. The cross-validation for only 2 examples would have been very irrelevant, because only one text would remain as a training example, so no generalization could be made. For the sets A and C 5kB pieces were used, and for set I 50kB. This produced 100-200 training examples for each author, so a better generalization could be made by the classifier. The splitting took into account sentences, so it would not interfere with the syntactic features. Also, the last slice could not be smaller than half of the average slice size in the training set.

3 Classification Task

For classification we used a SVM implementation available in WEKA, the Sequential Minimal Optimization (SMO) algorithm. The test documents were also split into pieces of the same size as the training data, and the most common result was used as the output of the classifier each document.

For the open class problem, we used the same classifier, but with a logistic regression model for the output because we needed a more exact probability estimation for each author in the training set. If the classifier offered an expected value over 0.75 for an author for a text on average (the text was also split into pieces of different sizes), the classifier outputs that class; otherwise the answer is "*other*".

We have also tried a Naive Bayes classifier, but the results were not as good as for SVM when using cross-validation on the training set. However, it offered very close results, so it can also be a viable classifier.

4 Results

In order to determine the reduced feature set presented in section 2, different combinations of features have been evaluated and we have selected the ones that had the best results in cross-validation. Different split sizes were used for texts as well.

The experimental validation concentrated only on the closed attribution problem. In the 10-fold cross-validation, the results were very good:

- 100% - set A (using 5kB and 10kB slices)
- 96.6% - set C (using 5kB slices)

- 99.5% - set I (using 20kB and 50kB slices)

However, these results are not very relevant, because the training examples are from the same document, so one expects many linguistic similarities between them. It was clear that the results on the test corpus will be significantly under these levels. However, the described approach turned out to yield good results on the PAN 2012 test corpora as well, both for the closed and open problems:

- A – 4/6 (66.66%)
- B – 8/10 (80%)
- C – 6/8 (75%)
- D – 12/17 (70.58%)
- I – 12/14 (85.71%)
- J – 13/16 (81.25%)

As expected, the results are not as good as for cross-validation, but the depreciation was not very steep. Thus, our solution obtained an overall document accuracy of 77%, ranking 3rd in the author attribution competition and an average accuracy over all the 6 test sets of 76%, ranking 7th, at a very close distance from the previous 4 places.

5 Conclusions

Using only a reduced set of linguistic features has proven to offer good results for the author identification task. These results might have improved by adding more application specific features. Moreover, spiting the training texts proved to be a good solution for training, evaluation and scoring the test documents. The last conclusion is that using logistic regression over the solution designed for the closed class problem provided competitive results for the open class problem as well.

References

1. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. **1** (2006) 233-334
2. Argamon, S., Juola, P.: Overview of the International Authorship Identification Competition at PAN-2011. In: Petras, V., Forner, P., Clough, P.D. (eds.): CLEF 2011 (Notebook Papers/Labs/Workshop)
3. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. **60** (2009) 538-556
4. Tanguy, L., Urieli, A., Calderone, B., Hathout, N., Sajous, F.: A Multitude of Linguistically-rich Features for Authorship Attribution. In: Petras, V., Forner, P., Clough, P.D. (eds.): CLEF 2011 (Notebook Papers/Labs/Workshop)
5. Kourtis, I., Stamatatos, E.: Author Identification Using Semi-supervised Learning. In: Petras, V., Forner, P., Clough, P.D. (eds.): CLEF 2011 (Notebook Papers/Labs/Workshop)
6. Kern, R., Seifert, C., Zechner, M., Granitzer, M.: Vote/Veto Meta-Classifier for Authorship Identification. In: Petras, V., Forner, P., Clough, P.D. (eds.): CLEF 2011 (Notebook Papers/Labs/Workshop)