

Authorship Attribution and Verification with Many Authors and Limited Data ¹

Kim Luyckx

Walter Daelemans

University of Antwerp, CNTS Language Technology Group

Abstract

Most studies in statistical or machine learning based authorship attribution focus on two or a few authors. This leads to an overestimation of the importance of the features extracted from the training data and found to be discriminating for these small sets of authors. Most studies also use sizes of training data that are unrealistic for most situations in which stylometry is applied (e.g., forensics), and thereby overestimate the accuracy of their approach in these situations. In this paper, we show, on the basis of a new corpus with 145 different authors, what the effect is of many authors on feature selection and learning, and show robustness of a memory-based learning approach in doing authorship attribution and verification with many authors and limited training data when compared to eager learners.

1 Introduction

In traditional studies on authorship attribution, the focus is on small sets of authors. Trying to classify an unseen text as being written by one of two or of a few authors is a relatively simple task, which in most cases can be solved with high reliability and accuracies over 95%. The field is however dominated by studies potentially overestimating the importance of the specific predictive features in experiments discriminating between only two or a few authors. A second problem in traditional studies are the unrealistic sizes of training data, which also makes the task considerably easier. Researchers tend to use over 10,000 words per author, which is regarded to be a reliable minimum for an authorial set. When only limited data is available for a specific author (e.g., a letter or e-mail), the authorship attribution task becomes much more difficult. Traditional approaches appear to be less reliable than expected from reported results when applied to more realistic applications like forensics.

In this paper, we present the first systematic study of the effect of many authors and limited data on feature selection and learning in the tasks of authorship attribution and verification. We show robustness of a memory-based learning approach in doing authorship attribution and verification with many authors and limited training data when compared to eager learners such as SVMs and maximum entropy learning.

2 Corpus and Approach

The 200,000-word Personae corpus in this study consists of 145 student (BA level) essays of about 1400 words about a documentary on Artificial Life, thereby keeping markers of genre, register, topic and age relatively constant. These essays contain a factual description of the documentary and the students opinion about it. The students also took an online Myers- Briggs Type Indicator (MBTI) test and submitted their profile, the text and some user information via a website. The corpus can therefore not only be used for authorship attribution experiments, but also for personality prediction.

We approach authorship attribution as an automatic text categorization task. As in most text categorization systems, we take a two-step approach in which our system (i) achieves automatic selection of features

¹The full version of this paper appeared in: *Proceedings of the Twenty-Second International Conference on Computational Linguistics (COLING '08)*, 2008, pp. 513-520.

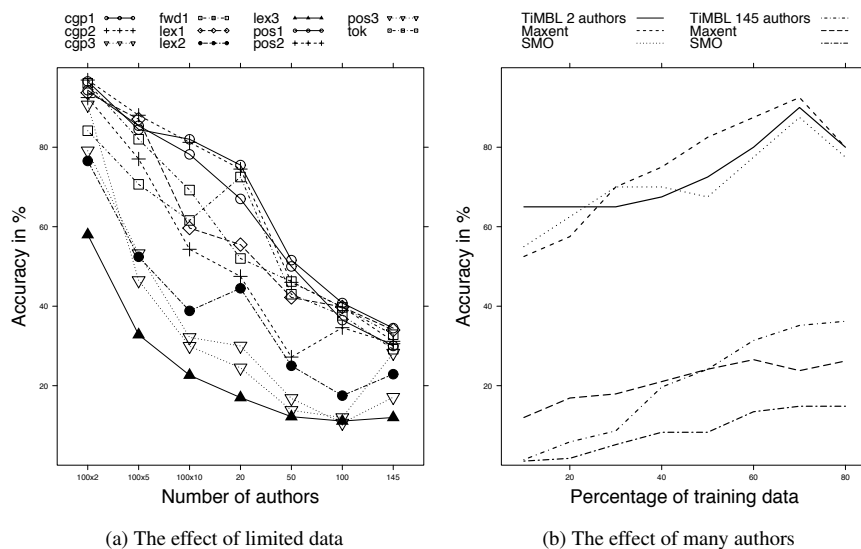
that have high predictive value for the categories to be learned, and (ii) uses machine learning algorithms to learn to categorize new documents by using the features selected in the first stage.

We use the Memory-Based Shallow Parser (MBSP) , for feature construction. MBSP tokenizes the input, performs a part-of-speech analysis, looks for noun phrase, verb phrase and other phrase chunks and detects subject and object of the sentence and a number of other grammatical relations. Features extracted include readability metrics, function word patters, n -grams of words and POS tags and vocabulary richness measures. We use the chi-square metric to select constructed features.

3 Experiments and Results (summary)

The paper focuses on three facets of authorship attribution, each with their own experimental set-up: (a) the effect of many authors on feature selection and learning; (b) the effect of limited data in authorship attribution; (c) the results of authorship attribution using many authors and limited data on learning.

For (a), we perform experiments in authorship attribution while gradually increasing the number of authors. We investigate (b) by performing authorship attribution on 2 and 145 authors while gradually increasing the amount of training data, keeping test constant at 20% of the entire corpus. The resulting learning curve will be used to compare performance of eager and lazy learners. The authorship attribution with many authors task (c) - which is closer to a realistic situation in e.g. forensics using limited data and many authors is approached as a one-class learning problem. For each of the 145 authors, we have 80% of the text in training and 20% in test. The negative class contains 80% of each of the other 144 authors training data in training and 20% in test.



We see, as expected, a marked effect of many authors and limited data in authorship attribution. When systematically increasing the number of authors in authorship attribution, we see that performance drops significantly (Figure a). On the positive side, similar types of features work well for different numbers of authors in our corpus, but generalizations about individual features are not useful. Memory-based learning shows robustness when dealing with limited data, which is essential in e.g. forensics (Figure b). Results from experiments in authorship attribution on 145 authors indicate that in almost 50% of the cases, a text from one of the 145 authors is classified correctly. Using combinations of good working lexical and syntactic features leads to significant improvements. The approximation of the authorship verification task is a much more difficult task, which only leads to a correct classification in 56% of the test cases. It is clear that studies reporting over 95% accuracy on a 2-author study are overestimating their performance and the importance of the features selected.

Further research with the 145-author corpus will involve a study of handling with imbalanced data and experimenting with other machine learning algorithms for authorship attribution and verification and a more systematic study of the behavior of different types of learning methods (including feature selection and other optimization issues) on this problem.