Authorship Attribution with Thousands of Candidate Authors

Moshe Koppel
Dept. of Computer Science
Bar Ilan University
Ramat Gan, Israel
moishk@gmail.com

Jonathan Schler
Dept. of Computer Science
Bar Ilan University
Ramat Gan, Israel
schlerj@cs.biu.ac.il

Shlomo Argamon
Dept. of Computer Science
Illinois Institute of Technology
Chicago,IL 60616
argamon@iit.edu

Eran Messeri
Dept. of CS
Bar-Ilan Univ.
Ramat Gan, Israel
masrie@cs.biu.ac.il

ABSTRACT
In this paper, we use a blog corpus to demonstrate that we can often identify the author of an anonymous text even where there are many thousands of candidate authors. Our approach combines standard information retrieval methods with a text categorization meta-learning scheme that determines when to even venture a guess.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning; J.4 [**Social and Behavioral Sciences**] *Sociology*

Keywords
Authorship attribution, blog analysis

1. INTRODUCTION
In the most straightforward version of the authorship attribution problem, we are asked to determine which of a small set of candidates is the actual author of a given text. This version of the problem can be thought of as a rather standard text categorization problem (DeVel et al 2002, Diederich et al 2003, Koppel & Schler 2003, Peng et al 2004, Hill & Provost 2003). It is distinguished from ordinary topic-based categorization problems only by the set of features likely to be relevant to its solution. In real life, however, it is rarely the case that we are given a small closed set of candidates from which to choose. Typically, we either have no closed set at all with which to work or we have a set of candidates so enormous that the usual text categorization approach is inapplicable.

In this paper, we will show that authorship attribution can be solved to a reasonable extent even when the number of candidate authors numbers in the many thousands.

2. PROBLEM DEFINITION
A convenient testbed for such a problem is the blogosphere, the emergence of which as a popular medium provides us with an essentially unlimited number of authors. Our corpus consists of over 18,000 blogs (each blog is the full set of posts by a given author), each of which includes at least 200 occurrences of common English words. The (self-reported) age and gender of each author is known and for each age interval the corpus includes an equal number of male and female authors. All but 10,000 blogs are set aside for training and validation, as will be explained. For each of the 10,000 blogs in our test set, we snip off at least 500 words at the (chronological) end of the blog, snipping off as much as necessary so as not to break up an individual post. We call these "snippets".

Our object is to determine which of the 10,000 bloggers is the author of a given snippet. We will attempt to solve the problem independently for each of the 10,000 snippets.

3. FIRST ATTEMPT: IR APPROACH
It is quite clear that it is impractical to construct models of each of the 10,000 candidate authors. Rather, as a first step we use an information retrieval approach. For a given snippet, we wish to determine whether the snippet includes linguistic features that tie it uniquely to one of the candidate authors. Since such features might reflect either a distinct writing style or a preoccupation with a particular topic, we consider three representations of texts:

1. *content tfidf*: tfidf restricted to content words
2. *content idf*: binary idf restricted to content words
3. *style*: tfidf on stylistic features (function words and strings of non-alphabetic, non-numeric characters)

For each of these representation methods, we use the usual cosine measure (Salton & Buckley 1988) to quantify the similarity of a given author's known work with a given snippet. The hope is that for a given snippet, some distinctive feature or features will render the snippet uniquely similar to exactly one of the candidate authors.

In fact, this crude approach does work to some limited extent. For each of the three document representation methods, and each of

the 10,000 snippets, we rank the authors by similarity to the snippet. In Figure 1, we show the percentage (*y*-axis) of snippets for which the actual author is assigned rank k or better (*x*-axis). It can be seen that, for each representation methods, over 20% of the snippets are most similar to the actual author. In fact, 42% of the snippets are most similar to the actual author for at least one of the three representation methods.

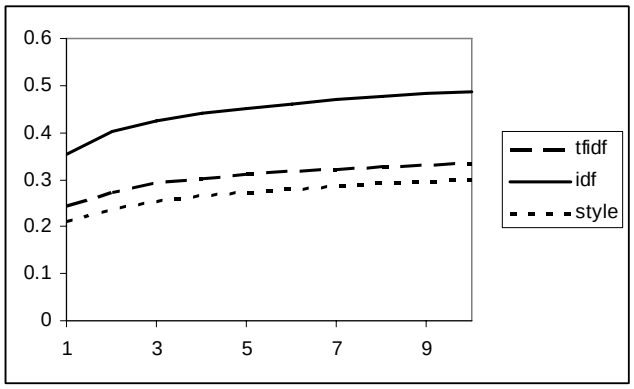

Figure 1: Percentage (*y*-axis) of snippets for which the actual author is assigned rank k or better (*x*-axis)

4. REFINEMENT: META-LEARNING

While this may be a surprisingly high number (given the large number of candidates), it is quite useless in the sense that we could never confidently assert that a given snippet was written by a given author; after all, we are still wrong most of the time. Therefore, we wish to isolate those cases for which we have high confidence that a particular representation method produces the right author; in all other cases, we will report that the results are inconclusive. The object is to return some answer in as many cases as possible and to achieve a high level of accuracy over those cases.

In order to accomplish this, we use a meta-learning scheme that exploits the 8,000+ blogs that we set aside for this purpose (the holdout set). For each representation method, we consider each pair consisting of a snippet and the author ranked most similar to that snippet. We call the pair a *successful* pair if the top-ranked author is in fact the actual author. We use the holdout set to learn to distinguish successful pairs from unsuccessful pairs. Each pair is represented in terms of a set of 18 meta-features, including the absolute similarity of the snippet to the top-ranked author, the gap in degree of similarity between the top-ranked author and the k-ranked author, the rank of the top-ranked author using the other two representation methods and so forth. For each representation method, a linear SVM is used to classify a pair as successful or not.

Now we return to our original experiment on the 10,000 snippets. The style-based representation method is hypothesized to be reliable by its meta-learner for 21.5% of the snippets and correctly classifies 84.0% of these snippets. For content tfidf and content idf, the numbers are 25% hypothesized reliable with 81.1% correct and 34% hypothesized reliable with 79.7% correct, respectively.

We combine the methods in the following way. For a given snippet, if exactly one of the representation methods yields a top-ranked author such that the snippet/author pair is hypothesized by the respective meta-learned classifier to be successful, we output that author as the right answer. If none of the three representation methods yields a successful pair, we return *Don't Know*. Likewise, if two representation methods yield different top-ranked authors and both are hypothesized to be reliable, we return *Don't Know*. As can be seen in Figure 2, overall we venture an answer in 31.3% of all cases and this answer is correct for 88.2% of such cases.

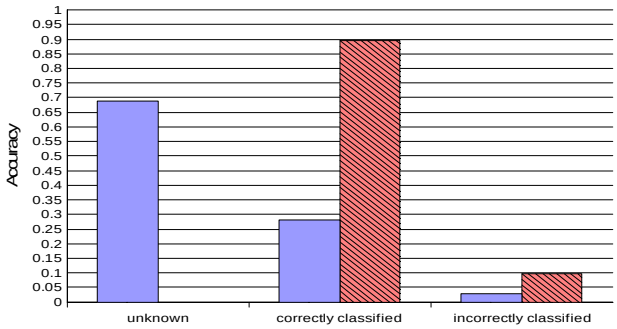

Figure 2: Percentage of snippets not classified, correctly classified and incorrectly classified. Diagonal hatching refers to percentage correctly classified among those classified.

Thus, we conclude that, provided we are willing to live with the response *Don't Know* in many cases, we can achieve reasonably reliable authorship attribution even where the number of candidate authors numbers in the many thousands.

5. REFERENCES

De Vel, O., M. Corney, A. Anderson and G. Mohay (2002), E-mail Authorship Attribution for Computer Forensics, in Applications of Data Mining in Computer Security, Barbara, D. and Jajodia, S. (eds.), Kluwer.

Diederich, J., J. Kindermann, E. Leopold and G. Paass (2003), Authorship Attribution with Support Vector Machines, Applied Intelligence 19(1), pp. 109-123

Hill, S. and F. Provost (2003), The Myth of the Double Blind Review? Author Identification using only Citations, SIGKDD Explorations Vol. 5 No. 2, 179-184.

Koppel, M. and J. Schler (2003), Exploiting Stylistic Idiosyncrasies for Authorship Attribution, in Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, pp. 69-72.

Peng, F. Schuurmans, D. and Wang, S. (2004). Augumenting Naive Bayes Text Classifier with Statistical Language Models , Information Retrieval, 7 (3-4), pp. 317 - 345

Salton, G. and C. Buckley (1988), Term Weighting Approaches in Automatic Text Retrieval, Information Processing Management Vol. 24 No. 5, 513-523.