Exploiting Stylistic Idiosyncrasies for Authorship Attribution

Moshe Koppel Jonathan Schler
Dept. of Computer Science
Bar-Ilan University
Ramat-Gan, Israel

Introduction

Early researchers in authorship attribution used a variety of statistical methods to identify stylistic discriminators – characteristics which remain approximately invariant within the works of a given author but which tend to vary from author to author (Holmes 1998, McEnery & Oakes 2000). In recent years machine learning methods have been applied to authorship attribution. A few examples include (Matthews & Merriam 1993, Holmes & Forsyth 1995, Stamatatos et al 2001, de Vel et al 2001).

Both the earlier "stylometric" work and the more recent machine-learning work have tended to focus on initial sets of candidate discriminators which are fairly ubiquitous. For example, the classical work of Mosteller and Wallace (1964) on the Federalist Papers used a set of several hundred function words, that is, words that are context-independent and hence unlikely to be biased towards specific topics. Other features used in even earlier work (Yule 1938) are complexity-based: average sentence length, average word length, type/token ratio and so forth. Recent technical advances in automated parsing and part-of-speech (POS) tagging have facilitated the use of syntactic and quasi-syntactic features such as POS n-grams (Baayen et al 1996, Argamon-Engelson et al 1998, Stamatatos et al 2001, Koppel et al 2003).

However, human experts working on real-life authorship attribution problems do not work this way. They typically seek idiosyncratic usage by a given author which serves as a unique fingerprint of that author. For example, Foster (2000) describes his techniques for identifying a variety of notorious anonymous authors including the author of the novel, *Primary Colors*, and the Unabomber. These techniques include repeated use of particular neologisms or unusual word usage. In the case of unedited texts, spelling and

grammatical errors, which are typically eliminated in the editing process, can be exploited as well.

The purpose of this paper is to attempt to simulate the idiosyncrasy-based methods used by human experts. We construct classes of common idiosyncratic usage and assess the usefulness of such features, both in and of themselves and in conjunction with other types of features, for authorship attribution.

We use as our corpus an email discussion group since such unedited material allows us to take maximal advantage of the features we are interested in. The problem of authorship attribution on email has been studied by de Vel et al (2001). They use a combination of lexical, complexity-based and formatting features, as well as "structural" features (attachments, HTML tags, etc.) unique to email. In our work we do not use such structural features in order not to focus too narrowly on email. We do consider, in addition to the other types of features considered by de Vel et al, various types of systematic errors of usage and spelling.

The Corpus

We chose as our corpus an email discussion group concerning automatic information extraction. This corpus offers a number of important advantages for the kinds of tests we wish to run. First, it includes sufficient material from a sufficient number of authors. To be precise, it included 480 emails written by 11 different authors during a period of about one year. The average length of a post is just over 200 words. Second, as is customary in such discussion groups, the writing is not overly polished so that repeated errors of all sorts can be found. Third, the material is homogeneous with regard to topic and cohort type so that differences that do exist are largely attributable to writing style. Finally, the material is public domain. (Nevertheless, we thought it prudent to disguise the names of the authors.)

All material not in the body of the text, as well as quoted material, was not considered.

Feature Sets

For the purposes of our experiments, we considered three classes of features:

1. Lexical – We used a standard set of 480 function words. We filtered these by using the infogain ranking on each training corpus and choosing the top 200 words.

2. Part-of-Speech Tags – We applied the Brill (1992) tagger to the entire corpus to tag each word with one of 59 POS tags. We then used as features the frequencies of all POS bi-grams which appeared at least three times in the corpus. (In early experiments, bi-grams proved more useful than other n-grams so we adopted it as a standard.)

3. Idiosyncratic Usage – We considered various types of idiosyncratic usage: syntactic, formatting and spelling. For example, we checked for frequency of sentence fragments, run-on sentences, unbroken sequences of multiple question marks and other punctuation, words shouted in CAPS and so forth. In addition, we considered various categories of common spelling errors such as inverted letters, missing letters, and so forth. The full list of 99 stylistic idiosyncrasies that we considered is shown in Table 1.

In order that our entire process be automated, we used the following procedure for detecting errors: We ran all our texts through the MS-Word application and its embedded spell-checker. Each error found in the text by the spell checker was recorded along with the best suggestion (to correct the error) suggested by the spell-checker. Each pair <error, suggestion> was processed by another program, which assigned it an "error type" from among those in the list we constructed. For certain classes of errors, we found MSWord's spell and grammar checker to be inadequate, so we prepared scripts ourselves for capturing them.

Error Type	*# Features*
Sentence Fragment	*1*
Run-on Sentence	*1*
Repeated Word	*1*
Missing Word	*1*
Mismatched Singular/Plural	*1*
Mismatched Tense	*1*
Missing hyphen	*1*
'that' *following comma*	*1*
Single consonant instead of double	*16*
Double consonant instead of single	*13*
Confused Letters 'x' and 'y'	*6*
Wrong vowel	*6*
Repeated Letter	*19*
Only One of Doubled Letter	*17*
Letter Inversion	*1*
Inserted Letter	*1*
Abbreviated Word	*1*
ALL CAPS words	*1*
Repeated non-letter\non-numeric characters	*10*

Table 1: List of 99 error features used in classification experiments.

It should be noted that we use the term "error" or "idiosyncrasy" to refer to non-standard usage or orthography in U.S. English, even though often such usage or orthography simply reflects different cultural traditions or deliberate author choice.

Experiments

We ran ten-fold cross-validation experiments on our corpus using various combinations of feature types and two different classification algorithms: linear SVM (Joachims 1999) and decision trees (C4.5). In all cases, our classifiers were allowed to assign a given test document to a single author only. Figure 1 shows results in terms of accuracy.

Figure 1: Accuracy (*y-axis*) on ten-fold cross-validation using various feature sets (*x-axis*) and classifying with linear SVM and C4.5, respectively.

Discussion

Several observations jump out of the data. First, for lexical features alone and POS features alone, SVM (47.9% and 46.2%, respectively) is more effective than C4.5 (38.0 and 40.4). This reflects the fact that SVM is designed to weigh contributions from a large number of features, while C4.5 selects out a relatively small number of thresholded features. For function word and POS bi-gram frequency, the relevant distinguishing information is typically spread around among many features. However, once errors are thrown into the mix the tables turn and C4.5 becomes more effective than SVM. The main point is that when classifying with C4.5, the difference between using errors or not using them is dramatic. In fact, errors completely set the tone when C4.5 is used with the other features hardly contributing. Errors alone achieve accuracy of 67.6 and in the best case, when all features are used, accuracy increases only to 72.0. For both classifiers, using lexical and POS features together without errors (C4.5: 61.7) under-performed using either one of them together with errors (C4.5/lexical: 68.8; C4.5/POS: 71.8).

Much insight can be gleaned by considering which features really do the work. Consider several interesting examples:

Author 1 uses British spelling. For example, he writes *organisation*, *summarisation*, and so forth. As a result the error type *confused 's' and 'z'* was extremely helpful for identifying Author 1.

Author 3 tended to double the letter 'n' at the end of names and words that more commonly (though not always) end in a single 'n', such as *Rosalynn*, *Bergmann*, and so forth. (Of course, such name spellings may not be errors at all but MSWord marks them as non-standard and certainly their repeated use in different names is significant.)

Author 7 tends to forget 'i's in the middle of words. For example, he writes *identifed*, *facilites*, *descripton* and so forth.

In each of these cases, these stylistic idiosyncrasies play the role of smoking guns. The problem is that such features are relatively rare and hence authors might make it through an entire short document without committing any of their habitual errors. Lexical and POS features, on the other hand, never quite disappear from view but they are rarely used with such outlandish frequency as to serve as smoking guns. Thus, as is evident in the results, the combination of these feature types is better than any one of them alone – but of the individual feature types stylistic idiosyncrasies constitute the most effective type.

Conclusions

We have found that the kinds of smoking guns that human experts exploit for authorship attribution can be identified and exploited in automated fashion. Moreover, the use of such features greatly enhances the accuracy of the results in comparison with methods which have generally been used in automated authorship attribution.

Certainly the list of stylistic idiosyncrasies we compiled for this study can be greatly enhanced. Neologisms of various types, non-standard use of legitimate words, awkward syntax and many other features a bit more difficult to detect using automated means would certainly help improve accuracy even more.

Although there is much anecdotal evidence that a small number of training documents is sufficient for authorship attribution, the sparseness of idiosyncratic features suggests that in this context even greater improvements might be expected when larger training corpora are available.

References

Argamon-Engelson, S., M. Koppel, G. Avneri (1998). Style-based text categorization: What newspaper am I reading?, in Proc. of AAAI Workshop on Learning for Text Categorization, 1998, pp. 1-4

Baayen, H., H. van Halteren, F. Tweedie (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, Literary and Linguistic Computing, 11, 1996.

Brill, E. (1992), A simple rule-based part-of-speech tagger, Proceedings of 3rd Conference on Applied Natural Language Processing, pp. 152—155

de Vel, O., A. Anderson, M. Corney and George M. Mohay (2001). Mining e-mail content for author identification forensics. SIGMOD Record 30(4), pp. 55-64

Foster, D. (2000). *Author Unknown: On the Trail of Anonymous*, New York: Henry Holt, 2000.

Holmes, D. (1998). The evolution of stylometry in humanities scholarship, Literary and Linguistic Computing, 13, 3, 1998, pp. 111-117.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features, Proceedings of 10th European Conference on Machine Learning, pp.137--142

Holmes, D. and R. Forsyth (1995). The Federalist revisited: New directions in authorship attribution, Literary and Linguistic Computing, pp. 111--127.

Koppel, M., S. Argamon, A. Shimony (2003). Automatically categorizing written texts by author gender, Literary and Linguistic computing, to appear

Matthews, R. and Merriam, T. (1993). Neural computation in stylometry : An application to the works of Shakespeare and Fletcher. Literary and Linguistic computing, 8(4):203-209.

McEnery, A., M. Oakes (2000). Authorship studies/textual statistics, in R. Dale, H. Moisl, H. Somers eds., Handbook of Natural Language Processing (Marcel Dekker, 2000).

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist.* Reading, Mass. : Addison Wesley, 1964.

Stamatatos, E., N. Fakotakis & G. Kokkinakis, (2001). Computer-based authorship attribution without lexical measures, Computers and the Humanities 35, pp. 193—214.

Yule, G.U. (1938). On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship, Biometrika, 30, 363-390, 1938.