Authorship attribution for small texts: Literary and forensic experiments

Ol'ga Feiguina
Cherches and Associates
Montreal, Canada
olga82@gmail.com

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, Canada M5S 3G4
gh@cs.toronto.edu

ABSTRACT

To capture syntactic structure as a feature for the classification of short texts by their authorship, we use the frequencies of bigrams in the stream of syntactic labels produced by a partial parser. We experimented on literary data (from the Brontë sisters) and simulated forensic data. Syntactic label bigrams were found to be helpful with the former but not the latter.

Categories and Subject Descriptors

I.2.7 [**Artificial intelligence**]: Natural language processing—*Language models, text analysis*

Keywords

Authorship attribution, partial parsing, literary data, forensic data, text classification

1. INTRODUCTION

Methods of authorship attribution developed for literary analysis typically require the document to be long and the comparison corpus to be large (did Shakespeare or Marlowe write this play?). However, in many applications, these assumptions are not valid. In literary analysis, the texts might be relatively short poems or stories. In forensic situations, the documents are likely to be short and the corpus small. For example, the document might be an anonymous letter whose authorship is to be compared with samples from suspects. In the detection of plagiarism, short segments of a longer work may be compared with one another to see if they bear evidence of diverse authorship. Previous approaches to authorship attribution have been unsuccessful on short texts (Burrows 2002) or have succeeded only in narrow domains by using customized and highly domain-specific features (Zheng, Li, Chen, and Huang 2006).

In this paper, we present a method for authorship attribution that is suitable for texts as short as around 200 words. The method makes better use of the syntactic information in the texts than prior approaches.

Earlier research on the use of syntax in authorship attribution has shown both the strengths and the limitations of using either full parses or simple syntactic chunking. Baayen, van Halteren, and Tweedie (1996), working on long, hand-parsed literary texts, represented a text as the bag of rewrite rules used in the syntactic derivation of each of its sentences, and applied vocabulary-richness measures to this bag. The results were better than those obtained by applying the same measures directly to the vocabulary of the text, but Baayen et al despaired of using their method with automatic parsing because its accuracy would be insufficient. Stamatatos, Fakotakis, and Kokkinakis (2000), working with news texts averaging around 1100 words, used simple non-embedded chunking of the text to derive a number of quantitative features for authorship classification. While their results were quite good, the method was dependent on artifacts of their particular chunker, and the error rate was high for shorter texts.

2. BIGRAMS OF SYNTACTIC LABELS

To obtain the strengths of using syntax while minimizing the weaknesses, we use partial parsing (Abney 1996), which produces an embedded but not recursive syntactic structure for sentences. We then represent a sentence as a sequence of the syntactic labels of its bracketed substructures and words, ignoring the brackets and the words themselves; this can be thought of as an approximation to the syntactic structure of the sentence (Hirst and Feiguina 2007). As an example, Figure 1 shows a fragment of a sentence, the corresponding structure from partial parsing, and the stream of labels that we take as its representation. A document can then be represented by the frequencies of bigrams of these syntactic labels; Figure 1 shows the bigrams extracted for the example fragment. These frequencies are then used as features for text classification by authorship. In addition, following Baayen et al, we also use the rewrite rules from the partial parser as a feature — both by simple frequency of use and by vocabulary richness.

3. TESTS WITH LITERARY DATA

We first tested the method on text by the Brontë sisters, selecting them because Koppel, Schler, and Mughaz (2004) had found them to be very hard to discriminate even by sophisticated authorship attribution methods. We took 250,000 words each from novels by Anne Brontë and Char-

SIGIR '07 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection.

Sentence [fragment]:
Let it be theirs to conceive the delight of joy ...

Partial parse:

```
[vp [vx [vb Let]]]
[c [c0 [nx [prp it]] [vx [be be]]] [nx [prp theirs]]]
[infp [inf [to to] [vb conceive]]
      [ng [nx [dt the] [nn delight]] [of of] [nx [nn joy]]]] ...
```

Stream of syntactic labels:

```
vp vx vb c c0 nx prp vx be nx prp infp inf to vb ng nx dt nn of nx nn ...
```

Bigrams of syntactic labels:

```
vp-vx vx-vb vb-c c-c0 c0-nx nx-prp prp-vx vx-be be-nx nx-prp prp-infp infp-inf inf-to to-vb vb-ng ng-nx
nx-dt dt-nn nn-of of-nx nx-nn ...
```

Figure 1: Example of partial parse, with corresponding label stream and bigrams.

lotte Brontë, and divided them at sentence boundaries into fragments of approximately 100, 500, or 200 words. (That is, we pretended that instead of writing novels they had written many independent short texts.) We then tried to classify these short texts by author, using frequencies of syntactic label bigrams as features. As a baseline for comparison, we also tried the task with a number of standard lexical features commonly used in authorship attribution (Graham, Hirst, and Marthi 2005), including frequencies of function words, part-of-speech tags, and word lengths, and vocabulary richness measures. Lastly, we tried combinations of these features with the syntactic label bigram frequencies. For classification, we used a support-vector machine, with ten-fold cross-validation for testing.

Our results, which we present in greater detail in Hirst and Feiguina 2007, are shown in Table 1. These results are based on the complete 500,000-word dataset, but accuracies started to level off when the size of the training set reached about 40,000 words; we discuss training-set size in Hirst and Feiguina 2007. The random baseline (the accuracy that would be achieved by guessing randomly) is 50%.

We observe immediately that, contrary to the results of Baayen et al, vocabulary-richness measures on rules give poor results by themselves, and we exclude them from further discussion, although they do in all cases give a boost to the other syntactic features. For the 1000-word texts, there is a ceiling effect: all conditions do quite well, though syntactic label bigram frequencies alone achieve a 99% accuracy, effectively the same as all lexical features combined. For smaller text sizes, not surprisingly, accuracy drops. However, for both 500-word and 200-word texts, the accuracy achieved by the combination of all feature sets exceeds that of any single set; that is, our label bigram frequencies increase accuracy compared to the use of standard lexical features alone. An examination of the nine label bigrams that were most discriminating found that seven of them involved non-terminal labels, indicating that a sensitivity to syntactic structure is indeed making a difference to the classification.

	Text size		
Features	1000	500	200
Syntactic features			
Label bigram freqs	99.0	93.4	84.9
Rule freqs	93.2	93.4	83.8
Vocab richness of rules	76.6	76.7	70.3
Bigram and rule freqs	98.4	95.8	87.4
All syntactic features	**99.5**	94.2	87.5
Lexical features			
PoS freqs	93.8	93.4	82.7
Other lexical features	97.5	90.5	85.6
All lexical features	98.9	95.0	89.5
All features	99.2	**96.8**	**92.4**

Table 1: Average accuracy (in percent) in 10-fold cross-validation on pairwise classification of Brontë texts, by text size and features used. Boldface indicates best results for each text size.

4. TESTS WITH FORENSIC DATA

Given this success on short literary texts, we then turned to using the method on forensic data — for example, anonymous threatening letters, tip-off notes, etc. We assume that attested writing samples from suspects are available for comparison. Because there is no readily available corpus of such data for use in experiments, we used simulated forensic data: Chaski's (2005a,b) model forensic dataset of short texts. Chaski asked 11 different authors to each write approximately 2000 words in total, choosing from topics such as a threatening letter, an apology, or a complaint. There are a total of 73 texts, ranging from 4 to 10 texts per author and varying widely in length (average length, 265 words). Depending on the method of data analysis, Chaski's own syntactically-aware method achieves 95% accuracy (Chaski 2005a) or 81.5% accuracy (Chaski 2005b) in pairwise author identification on this dataset; we take the latter, the more-recent publication, as definitive.

We applied our method to this dataset, classifying both complete texts, regardless of length, and fragments of approximately 200 words. We tried both pairwise authorship classification (with a dataset size of about 4000 words) and multiclass (1 in 11) classification (with a dataset size of about 22,000 words). For pairwise classification, the random baselines are 50% for uniformly guessing an author and 50 to 70% (depending on the pair) for always guessing the author with the greater number of texts; for multiclass classification they are 9% and 14% respectively. The results are shown in Tables 2 and 3. (We have dropped vocabulary richness of rules as a feature and added some new combination feature sets.)

	Text size	
Features	Whole	200
Syntactic features		
Label bigram freqs	86.1	78.8
Rule freqs	87.3	72.4
Label bigram and rule freqs	88.3	75.4
Lexical features		
PoS freqs	89.2	84.1
Other lexical features	84.4	83.2
All lexical features	**91.2**	**85.6**
Combinations		
Label bigrams and other lexical	88.3	83.3
Label bigrams and all lexical	89.3	80.0
All features	88.7	75.6

Table 2: Average accuracy (in percent) in 10-fold cross-validation on pairwise classification of simulated forensic texts, by text size and features used. Boldface indicates best results for each text size.

For pairwise classification (Table 2), the accuracy is lower overall than for the Brontë data, but is of course based on much smaller training data; in fact, the accuracy is much higher than for a Brontë dataset of the same size. But we observe quite a different pattern in the results compared to those for the Brontë texts. Here, the standard lexical features do better than the syntactic features and better than Chaski's method, especially for the smaller texts. Moreover, adding the syntactic features to the lexical features degrades performance. Examination of the most-discriminating label bigrams shows that, contrary to the Brontë case, almost all of them involved terminal symbols, and the method's sensitivity to syntactic structure was hardly used at all.

The results for multiclass classification also showed a superiority for standard lexical features, although the overall pattern was quite different yet again (Table 3), with the efficacy of a feature set or combination varying widely with text size. For whole texts, while label bigrams perform better than lexical features in general, frequencies of part-of-speech tags alone do best, and combinations do worse or no better. For the 200-word fragments, however, the performance with part-of-speech tags degrades severely, while that of other lexical features actually improves.

These results can be explained in part simply by the small and unbalanced dataset that was used. However, it is also clear that the writing styles in the simulated forensic data were more distinct from one another than the styles of the Brontë sisters are, and their differences were more at the lexical level than the syntactic level. That is, “ordinary writers” are an “easier problem” than the Brontë sisters. This suggests that the use of an intermediate feature such as PoS-tag bigrams might be more successful for this kind of data. (Spassova and Turell (2006) have carried out experiments with high-frequency PoS-tag trigrams for authorship attribution and reported promising results.)

5. CONCLUSION

Syntactic label bigrams were found to be a helpful feature in discriminating authorship of short texts by the Brontë sisters, but were not helpful on simulated forensic data in which syntactic distinctions seemed to be less necessary. This can be attributed in part to the imbalance and small size of the forensic dataset, but it is also a reminder that features for authorship attribution can be very genre- or situation-specific.

	Text size	
Features	Whole	200
Syntactic features		
Label bigram freqs	57.5	39.6
Rule freqs	56.2	25.5
Label bigram and rule freqs	56.2	34.9
Lexical features		
PoS freqs	**60.3**	34.0
Other lexical features	41.4	**50.9**
All lexical features	51.0	49.1
Combinations		
Label bigrams and other lexical	**60.3**	48.1
Label bigrams and all lexical	58.9	50.0
All features	**60.3**	37.7

Table 3: Average accuracy (in percent) in 10-fold cross-validation on multiclass (1 in 11) classification of simulated forensic texts, by text and features used. Boldface indicates best results for each text size.

6. ACKNOWLEDGEMENTS

Our research is funded by the Natural Sciences and Engineering Research Council of Canada. We are grateful to Carole Chaski for allowing us to use her dataset of simulated forensic texts, to Neil Graham for his assistance and for the use of his software for the lexical features that were used in his work, and to Nadia Talent for helpful comments on an earlier draft of this paper.

7. REFERENCES

[1] Abney, Steven (1996). Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4): 337–344.

[2] Baayen, R. Harald; van Halteren, Hans; and Tweedie, Fiona J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3): 121–131.

[3] Burrows, John (2002). ‘Delta’: A measure of stylistic difference and likely authorship. *Literary and Linguistic Computing*, 17(3): 267–287.

[4] Chaski, Carole E. (2005a). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1).

[5] Chaski, Carole E. (2005b). Computational stylistics in forensic author identification. *SIGIR Workshop on Stylistic Analysis of Text for Information Access.*

[6] Graham, Neil; Hirst, Graeme; and Marthi, Bhaskara (2005). Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4): 397–415.

[7] Hirst, Graeme and Feiguina, Ol'ga (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, to appear.

[8] Koppel, Moshe; Schler, Jonathan; Mughaz, Dror (2004). Text categorization for authorship verification. *Eighth International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, Florida.

[9] Spassova, Maria S. and Turell, M. Teresa (2006). The use of morpho-syntactically annotated tag sequences as markers of authorship. *Proceedings of the Second European IAFL Conference on Forensic Linguistics / Language and the Law*, Barcelona.

[10] Stamatatos, Efstathios; Fakotakis, Nikos; and Kokkinakis, George (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4): 471–495.

[11] Zheng, Rong; Li, Jiexun; Chen, Hsinchun; and Huang, Zan (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3): 378–393.