MEASURING THE USEFULNESS OF FUNCTION WORDS FOR AUTHORSHIP ATTRIBUTION

Shlomo Argamon *(argamon@iit.edu)*
Illinois Institute of Technology
Shlomo Levitan *(levishl@iit.edu)*
Illinois Institute of Technology

INTRODUCTION

Some forty years ago, Mosteller and Wallace suggested in their influential work on the Federalist Papers that a small number of the most frequent words in a language ('function words') could usefully serve as indicators of authorial style. The decades since have seen this work taken up in many ways including both the use of new analysis techniques (discriminant analysis, PCA, neural networks, and more), as well as the search for more sophisticated features by which to capture stylistic properties of texts. Interestingly, while use of more sophisticated models and algorithms have often led to more reliable and generally applicable results, it has proven quite difficult to improve on the general usefulness of function words for stylistic attribution. Indeed, John F. Burrows in his seminal work on Jane Austen has demonstrated that function words can be quite effectively used for attributing text passages to different authors, novels, or individual characters.

The intuition behind the utility of function words for stylistic attribution is as follows. Due to their high frequency in the language and highly grammaticalized roles, function words are very unlikely to be subject to conscious control by the author. At the same time, the frequencies of different function words vary greatly across different authors and genres of text - hence the expectation that modeling the interdependence of different function word frequencies with style will result in effective attribution. However, the highly reductionistic nature of such features seems unsatisfying, as they rarely give good insight into underlying stylistic issues, thus the various efforts at developing more complex textual features while respecting constraints on computational feasibility.

One especially promising line of work in this regard has been the examination of frequent word sequences and collocations for stylistic attribution, particularly Hoover's recent (2004) systematic work on clustering analysis of several text collections using frequent word collocations. A "word collocation" is defined as a certain pair of words occurring within a given threshold distance of each other (such as "is" and "certain" appearing within 5 words of each other in this sentence). Given such a threshold, the most frequent such collocations are determined over the entire corpus, and their frequencies in each text constitute its features for analysis. Hoover's analyses show the superiority, for his data set, of using frequent word collocations (for certain window sizes) over using frequent words or pairs of adjacent words.

We contend, however, that by using such a small data set (twenty samples of 10,000 words each, in one case), the discriminating power of a model based on function words will be much reduced, and so the comparison may not be fair. As has been shown for other computational linguistic tasks (see, e.g., Banko & Brill), even simple language modeling techniques can greatly improve in effectiveness when larger quantities of data are applied. We have therefore explored the relative effectiveness of frequent words compared to frequent pairs and collocations, for attribution of both author identity and national origin, increasing the number of text passages considered over earlier work.

We performed classification experiments on the twenty novels considered by Hoover, treating each separate chapter of each book as a separate text (rather than using just the first 10,000 words of each novel as a single text). Table 1 gives the full list with numbers of chapters and average number of words per chapter. We used a standard state-of-the-art machine learning technique to derive linear discrimination models between pairs of authors. This procedure gave results that clearly show a superiority of function words over collocations as stylistic features. Qualitatively similar results were obtained for the two-class problem of attributing the national origin (American or British) of a text's author. We conclude from this that larger and more detailed studies need to be done to effectively validate the use of a given feature type for authorship attribution.

Author	Book	# Chapters	Avg. Words
Cather	*My Antonia*	45	1826
	Song of the Lark	60	2581
	The Professor's House	28	2172
Conrad	*Lord Jim*	45	2913
	The Nigger of the Narcissus	5	10592
Hardy	*Jude the Obscure*	53	2765
	The Mayor of Casterbridge	45	2615
	Tess of the d'Urbervilles	58	2605
James	*The Europeans*	12	5003
	The Ambassadors	36	4584
Kipling	*The Jungle Book*	13	3980
	Kim	15	7167
Lewis	*Babbit*	34	3693
	Main Street	34	4994
	Our Mr. Wrenn	19	4126
London	*The Call of The Wild*	7	4589
	The Sea Wolf	39	2739
	White Fang	25	2917
Wells	*The Invisible Man*	28	1756
	The War Of The Worlds	27	2241

Table 1. Corpus composition.

METHODOLOGY

Given each particular feature set (frequent words, pairs, or collocations), the method was to represent each document as a numerical vector, each of whose elements is the frequency of a particular feature of the text. We then applied the SMO learning algorithm (Platt) with default parameters, which gives a model linearly weighting the various text features. SMO is a support vector machine (SVM) algorithm; SVMs have been applied successfully to a wide variety of text categorization problems (Joachims).

Generalization accuracy was measured using 20-fold cross-validation, in which the 633 chapters were divided into 20 subsets of nearly equal size (3 or 4 texts per subset). Training was performed 20 times, each time leaving out one of the subsets, and then using the omitted subset for testing. The overall classification error rate was estimated as the average error rate over all 20 runs. This method gives a reasonable estimate of the expected error rate of the learning method for each given feature set and target task (Goutte).

RESULTS

Results of measuring generalization accuracy for different feature sets are summarized in Tables 2 and 3, which clearly shows that using the most frequent words in the corpus as features for stylistic text classification give the highest overall discrimination for both author and nationality attribution tasks.

Feature Set	Author	Nationality
Freq. Words	**99.00%**	**93.50%**
Freq. Pairs	91.60%	91.30%
Freq. Coll. (k=5)	88.94%	90.20%
Freq. Coll. (k=10)	84.00%	87.20%

Table 2. 20-fold cross-validation results for 200 most frequent words, pairs, and collocations (window size k = 5 or 10).

Feature Set	Author	Nationality
Freq. Words	93.20%	**93.50%**
Freq. Pairs	90.00%	88.60%
Freq. Coll. (k=5)	91.50%	92.10%
Freq. Coll. (k=10)	94.00%	92.10%

Table 3. 20-fold cross-validation results for 500 most frequent words, pairs, and collocations (window size 5 or 10).

DISCUSSION

Our study here reinforces many others over the years in showing the surprising resilience of frequently-occurring words as indicators of the stylistic character of a text. Our results show frequent words enabling more accurate text attribution than features such as word pairs or collocations, surprisingly contradicting recent results as well as the intuition that pairs or collocations should be more informative. The success of this study at showing the power of frequent words we mainly attribute to the use of more data, in the form of entire novels, broken down by chapters. The more fine-grained breakdown of text samples for each author enables more accurate determination of a good decision surface for the problem, thus better utilizing the power of all features in the feature set. Furthermore, using more training texts than features seriously reduces the likelihood of overfitting the model to the training data, improving the reliability of results.

It is indeed possible that collocations may be better than function words for different stylistic classification tasks; however such a claim remains to be proven. A more general interpretation of our results is that since a set of frequent collocations of a given size will contain fewer different words than a set of frequent

words of the same size, it may possess less discriminatory power. At the same time, though, such a feature set will be less subject to overfitting, and so may appear better when very small sets of texts are studied (as in previous studies). Our results thus lead us to believe that most of the discriminating power of collocations is due to the frequent words they contain (and not the collocations themselves), thus frequent words outperformed collocations given sufficient data.

CONCLUSIONS

Function words still prove surprisingly useful as features for stylistic text attribution, even after many decades of research on features and algorithms for stylometric analysis. We believe that significant progress is likely to come from fundamental advances in computational linguistics which allow automated extraction of more linguistically motivated features, such as recent work on extracting rhetorical relations in a text (Marcu).

More generally, our results argue for the importance of using larger data sets for evaluating the relative utility of different attribution feature sets or techniques. As in our case of comparing frequent words with frequent collocations, changing the scale of the data set may affect the relative power of different techniques, thus leading to different conclusions. We suggest that the authorship attribution community should now work towards developing a large suite of corpora and testbed tasks, to allow more rigorous and standardized comparisons of alternative approaches.

Bibliography

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

Banko, M., and E. Brill. "Scaling to very very large corpora for natural language disambiguation." *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 2001. 26-33.

Biber, D., S. Conrad, and R. Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

Burrows, J. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press, 1987.

Goutte, C. "Note on free lunches and cross-validation." *Neural Computation* 9.6 (1997): 1246-9.

Hoover, D. L. "Frequent collocations and authorial style." *Literary and Linguistic Computing* 18.3 (2004): 261-28.

Joachims, T. "Text categorization with Support Vector Machines: Learning with many relevant features." *Machine Learning: ECML-98, Tenth European Conference on Machine Learning*. 1998. 137-142.

Marcu, D. "The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach." *Comp. Ling.* 26.3 (2000): 395-448.

Matthews, R., and T. Merriam. "Neural computation in stylometry: An application to the works of Shakespeare and Fletcher." *Literary and Linguistic Computing* 8.4 (1993): 203-209.

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

Platt, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Microsoft Research Technical Report MSR-TR-98-14, 1998. Accessed 2005-03-08. <ftp://ftp.research.microsoft.com/pub/tr/tr-98-14.pdf>

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