Explaining Delta, or: How do distance measures for authorship attribution work?

Stefan Evert, Thomas Proisl
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
{stefan.evert,thomas.proisl}@fau.de

Christof Schöch, Fotis Jannidis, Steffen Pielström, Thorsten Vitt
University of Würzburg, Germany
{christof.schoech,fotis.jannidis,
thorsten.vitt}@uni-wuerzburg.de
pielstroem@biozentrum.uni-wuerzburg.de

1 Introduction

Authorship Attribution is a research area in quantitative text analysis concerned with attributing texts of unknown or disputed authorship to their actual author based on quantitatively measured linguistic evidence (see Juola 2006; Stamatatos 2009; Koppel et al. 2009). Authorship attribution has applications in literary studies, history, forensics and many other fields, e.g. corpus stylistics (Oakes 2009). The fundamental assumption in authorship attribution is that individuals have idiosyncratic habits of language use, leading to a stylistic similarity of texts written by the same person. Many of these stylistic habits can be measured by assessing the relative frequencies of function words or parts of speech, vocabulary richness, and many other linguistic features. Distance metrics between the resulting feature vectors indicate the overall similarity of texts to each other, and can be used for attributing a text of unknown authorship to the most similar of a (usually closed) set of candidate authors.

The aim of this paper is to present findings from a larger investigation of authorship attribution methods which centres around the following questions: (a) How and why exactly does authorship attribution based on distance measures work? (b) Why do different distance measures and normalization strategies perform differently? (c) Specifically, why do they perform differently for different languages and language families, and (d) How can such knowledge be used to improve authorship attribution methods?

First, we describe current issues in authorship attribution and contextualize our own work. Second, we report some of our earlier research into the question. Then, we present our most recent investigation, which pertains to the effects of normalization methods and distance measures in different languages, describing our aims, data and methods. We conclude with a summary of our results.

2 Current issues in authorship attribution

There are several key elements to any authorship attribution study: the nature and extent of textual material available, the richness of metadata about the texts, the number and types of linguistic features used, the strategy used to normalize the resulting feature vectors, an optional dimensionality reduction step (often by principal component analysis), the measure used to assess distances between feature vectors, and the method for classification or clustering of the texts based on feature vectors and inter-text distances. All of these aspects are currently topics of investigation and debate in the authorship attribution community (e.g. Argamon 2008; Eder and Rybicki 2013). This paper is mainly concerned with the role of standardization and normalization of feature vectors, the choice of suitable features, and the impact of different distance metrics.

The current state of the art is to consider normalization and metric as one joint step in the process of authorship attribution. One ground-breaking measure, Burrows's Delta (Burrows 2002), can in fact be understood as a combination of standardization (i.e. z-transformation) of frequency counts combined with the well-known "Manhattan" (or "city block") metric. Many other measures proposed in the literature also amalgamate the two steps (e.g. Hoover 2004a, 2004b; Smith and Aldridge 2011). In this paper, we follow Argamon's (2008) lead and consider normalization strategy and distance measure separately from each other. This allows us to investigate the influence of each parameter on authorship attribution results as well as the interaction of these two parameters.

3 Previous work

In recent previous work, we describe an empirical investigation of the performance of 15 different text distance measures available for authorship attribution. For evaluating their performance, we compiled three collections of novels (English, French, German), each consisting of 75 complete texts of known authorship (three novels each by 25 authors), and ranging from the early nineteenth century to the first half of the twentieth century. The texts come from *Project Gutenberg*, the *TextGrid* collection and *Ebooks libres et gratuits*.

We compared the performance of the different text distance measures for feature vectors of 100–5000 most frequent words (mfw) and for all three corpora. We used two quantitative measures to evaluate performance: (a) the accuracy of the

clustering results relative to the gold standard if each cluster is labelled with the appropriate author; (b) a comparison of the average distance between works of the same author with the average distance between works by different authors.

As a result, we were able to demonstrate that most modifications of Burrows's original Delta suggested in the recent literature do not yield better results, even though they have better mathematical justification. Our results indicate that Eder's Delta, a measure specifically designed for highly inflected languages, does perform slightly better on French texts. The best distance measure for authorship attribution is the cosine-based Delta measure recently suggested by Smith and Aldridge (2011). Also, most text distance measures work best if between 1000 and 2000 of the most frequent words are used (Jannidis et al. 2015).

4 Current research

This work has lead us to several further questions: First, how do the effects of normalization and distance measure interact with each other? Second, why does the performance of a given combination of normalization and distance measure vary across different languages? And can this variation be explained by looking at the frequency distributions of individual, highly frequent words across texts in different languages? Finally, how can we identify the words (or features) that contribute most to the overall distance between texts? Are there linguistic or distributional explanations why these words are particularly indicative of the authorship of a text?

We approach this set of problems from two perspectives. First, we look at some mathematical properties of the authorship classification problem, based on geometric and probabilistic interpretations of the text distance measures. Argamon (2008) suggests two versions of Delta that can be interpreted in terms of statistical significance tests. However, our previous empirical results show that they are inferior to other measures that lack a similarly well-founded mathematical motivation. We are currently investigating the reasons for this discrepancy, with a particular focus on the role of different normalization strategies and their interaction with various distance measures. The results will show which aspects of the word frequency profiles of text samples are exploited by successful authorship classification methods. They may also help to identify salient lexical features that distinguish the individual writing styles of different authors.

Second, we explore another strategy for obtaining the set of features. Instead of relying on a specified number of most frequent words (mfw), we systematically identify a set of discriminant words by using the method of recursive feature elimination. We repeatedly train a support vector classifier and prune the least important features until we obtain a minimal set of features that gives optimal performance. The resulting feature set is much smaller than the number of mfw typically required by Delta measures. It contains not only function words but also common and not so common content words. The features work well on unseen data from the same and from different authors, not only yielding superior classification results, but also outperforming the mfw approach for clustering texts. This preliminary finding stands in contrast to accepted stylometric lore that function words are the most useful feature for discriminating texts from different authors.

References

Argamon, S. 2008. "Interpreting Burrows' Delta: Geometric and probabilistic foundations." *Literary and Linguistic Computing* **23**(2), 131–147.

Burrows, J. 2002. "'Delta' – A measure of stylistic difference and a guide to likely authorship." *Literary and Linguistic Computing* **17**(3), 267–287.

Eder, M. and Rybicki, J. 2013. "Do birds of a feather really flock together, or how to choose training samples for authorship attribution." *Literary and Linguistic Computing* **28**(2), 229–236.

Juola, P. 2006. "Authorship Attribution." *Foundations and Trends in Information Retrieval* **1**(3), 233–334.

Jannidis, F; Pielström, S.; Schöch, C.; Vitt, Th. 2015 (to appear). "Improving Burrows' Delta. An empirical evaluation of text distance measures." In: *Digital Humanities Conference 2015*.

Hoover, D. 2004a. "Testing Burrows' Delta." *Literary and Linguistic Computing* **19**(4), 453–475.

Hoover, D. 2004b. "Delta Prime?" *Literary and Linguistic Computing* **19**(4), 477–495.

Koppel, M., Schler, J. and Argamon, S. 2009. "Computational methods in authorship attribution." *Journal of the American Society for Information Science and Technology* **60**(1), 9–26.

Oakes, M. P. 2009. "Corpus linguistics and stylometry." In A. Lüdeling & M. Kytö (eds.) *Corpus Linguistics: An International Handbook*, Berlin: Mouton de Gruyter, Berlin, pp. 1070–1090.

Smith, P. and Aldridge W. 2011. "Improving authorship attribution. Optimizing Burrows' Delta method." *Journal of Quantitative Linguistics* **18**(1), 63–88.

Stamatatos, E. 2009. "A survey of modern authorship attribution methods." *Journal of the American Society for Information Science and Technology* **60**(3), 538–556.