Author attribution on paragraph level using simulated annealing

Marijn Schraagen M.P.Schraagen@uu.nl
Digital Humanities Lab, Utrecht University

Dirk van Miert D.K.W.vanMiert@uu.nl
Department of Philosophy and Religious Studies, Utrecht University

Keywords: author attribution, Enlightenment Philosophy, simulated annealing

Abstract

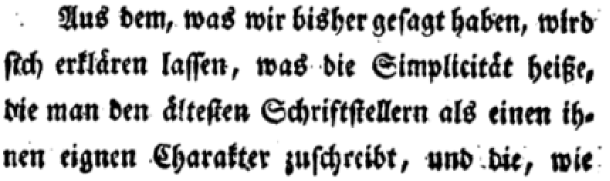
Authorship attribution, i.e., determining the author of a document automatically based on a reference corpus, is an established topic in computational linguistics and Digital Humanities. However, state of the art techniques are generally applied to documents of 5000 words or more. This research explores a method to attribute authorship on paragraph level, using simulated annealing to incrementally increase the accuracy of classification.

1. Introduction

In 1783, the German philosophers Christian Garve and Johann Feder published a co-authored review of Kant's "Critique of Pure Reason". Each author took a different view on the subject, which has prompted modern-day philosophers to investigate authorship of this text on a paragraph level. This task in turn provided the motivation to develop a stylometric algorithm to automatically determine authorship on this level of granularity. This abstract describes the iterative application of an established authorship algorithm to gradually improve classification of text fragments using simulated annealing, applied to works of Garve and Feder.

2. Related work

Stylometry and authorship attribution have received considerable attention in various research communities (Stamatatos, 2009). Besides traditional machine learning approaches such as k-nearest neighbors and Support Vector Machines, in stylometry *Burrows* Δ (Burrows, 2002) is widely used, which computes the difference between probabilities of frequent words in a test sample compared to reference data. The performance of Δ is comparable to other methods (Jockers & Witten, 2010) with the advantage of simplicity in parameters. However, for stylometry in general, minimum text length constraints do not allow accurate paragraph-level classification (Eder, 2015).



Aus dem, was wir bisher geſagt haben, wird ſich erklären laſſen, was die Simplicität heiße, die man den älteſten Schriftſtellern als einen ihnen eignen Charakter zuſchreibt, und die, wie

Figure 1. Fragment from Garve's *Abhandlungen.* Translation: *From what we have said so far, we can explain what is called the Simplicity* [. . .]

3. Data

The current experiments use as source data the books *Sammlung einiger Abhandlungen (G)* (Christian Garve, 1779, Figure 1) and *Lehrbuch der praktischen Philosophie (F)* (Johann Feder, 1770), available from Google Books as scans and OCR results. The narrow layout of the German Fraktur (Gothic) script causes a high occurrence of hyphenation, lowering accuracy of word frequency counts. To address this issue, a word at the start of a line is appended to the word at the end of the previous line in case the concatenation is found anywhere else in the document in the non-peripheral part of a line. OCR errors have limited impact on the analysis, since the author attribution algorithm is based on the most frequent words in a document (see Section 4), for which the OCR procedure has a relatively high accuracy.

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

4. Method

The goal of the current experiments is to classify fragments of 100 words each into two author classes. The method uses Burrows Δ as implemented in the `stylo` library for the R programming language (Eder et al., in press). To address the text length constraints of the Δ algorithm the problem is formulated as follows: given a set T of text fragments, find the best possible separation of this set into the two author classes.

The set T is constructed by dividing both G and F into two equal parts, set aside one part of each document as a combined reference corpus $C = C^G \cup C^F$, dividing the remaining part of each document into fragments of 100 words, and combining the two sets of fragments (791 fragments in total) into T. Next, the set T is split proportionally to the source documents into two sets T^G and T^F. At this stage, both sets contain an equal proportion of fragments from each author. The next step of the method consists of a simulated annealing procedure, iteratively swapping a random selection of 10 fragments between the two sets. In each iteration, Δ is computed to measure the distance of T^G and T^F to C^G and C^F, respectively. Equation (1) shows the computation of $\Delta(T^{A \in \{G,F\}})$, with variables MFW_i as the set of i most frequent words in C, w_X for the relative frequency of word w in X, and $w_{\sigma,C}$ as standard deviation of the frequency of w in C. If $\Delta(T^G) + \Delta(T^F)$ is smaller than the best result so far, then the current swap is retained, otherwise the previous state is restored. The process is repeated until convergence, defined as the absence of improvement for a predefined number m of consecutive iterations.

$$\Delta(T^A) = \frac{1}{i} \cdot \sum_{w \in MFW_i} \left| \frac{w_{T^A} - w_C}{w_{\sigma,C}} - \frac{w_{C^A} - w_C}{w_{\sigma,C}} \right| \quad (1)$$

5. Preliminary results

To validate the use of Δ for the current data, in an initial experiment the proportion of the fragments has been gradually shifted from one author to the other. Figure 2 shows a correlation, i.e., increasing the proportion of fragments of an author improves the Δ score for this author (and reduces the score for the other author), although not fully monotonous. As main experiment, the above method has been applied using $m = 1000$ and variable i (Table 1).

6. Discussion and future work

The simulated annealing procedure is capable of approaching Δ scores for the corresponding original single-author text, for MFW_{50} even exceeding the reference score. However, the percentage of actually correctly classified fragments, although significantly improved over the random initialization, is relatively low (69% at best). This indicates that the same Δ score may be reached by documents with a different distribution of fragments over authors. In terms of stylometry this means that some fragments do not contribute to the distinctive 'style' of an author, and instead some fragments by another author are more typical of the first author than the non-contributing fragments.

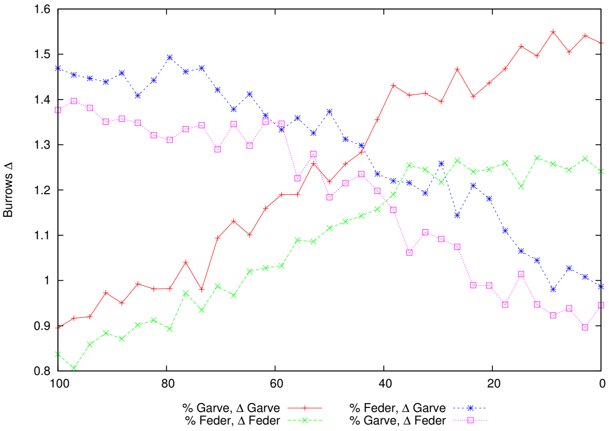


Figure 2. Proportion of fragments vs. Δ score, MFW_{100}

Interestingly, during iteration a higher percentage of correctly classified fragments is occasionally obtained. In these cases the Δ score continues to improve, while the classification percentage moves down and stabilizes below the intermediate maximum. Both observations challenge the correlation between Δ score and proportional distribution of fragments over authors.

Increasing the number of MFWs has a negative effect on the Δ score. However, the percentage correctly classified fragments does increase with more MFWs.

The experiments discussed here provide a some preliminary insight into the problem of paragraph-level author attribution. Systematic investigation of the impact of data quality, parameters and other classification algorithms is necessary to improve the results further, using the current method as a starting point.

Table 1. Authorship attribution results, best of 5 runs.

MFW_i	Single-author	Init	Final	Correct
i 50	Δ 1.75	2.39	1.64	% 59.4
100	1.73	2.42	1.76	62.2
250	1.72	2.44	1.86	66.7
500	1.76	2.45	1.91	69.2

References

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

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, *30*, 167–182.

Eder, M., Rybicki, J., & Kestemont, M. (in press). Stylometry with R: A package for computational text analysis. *The R Journal*.

Jockers, M., & Witten, D. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, *25*, 215–223.

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