Short samples in authorship attribution: a new approach

Maciej Eder
maciejeder@gmail.com
Institute of Polish Language
Polish Academy of Sciences, Poland

Introduction

The question of minimal sample size is one of the most important issues in stylometry and non-traditional authorship attribution. In the last decade or so, a few studies concerning different aspects of scalability in stylometry have been published (Zhao and Zobel, 2005; Hirst and Feiguina, 2007; Stamatatos, 2008; Koppel et al., 2009; Mikros, 2009; Luyckx and Daelemans, 2011), but the question has not been answered comprehensively. In his recent study, Eder proposed a systematic approach to solve the problem in a series of experiments, claiming that a sample should have at least 5,000 running words to be attributable (Eder, 2015).

The above studies (and many other as well) tacitly assume that there exists a certain amount of linguistic data that allows for reliable authorial recognition, and the real problem at stake is to determine that very value. However, one can assume that the authorial fingerprint is not distributed evenly in a collection of texts. Just the contrary, many experiments seem to suggest that the authorial voice is sometimes overshadowed by other signals, such as genre, gender, chronology, or translation. Some authors, say Chandler, should be easily attributable, while some other authors, say Virginia Woolf, will probably have their fingerprint somewhat hidden. Moreover, authorship attribution is ultimately a matter of context: telling apart Hemingway and Dickens will always be easier than distinguishing the Bronte sisters. On theoretical grounds, then, the minimal sample size can not be determined once and forever for the entire corpus, but may be different for different texts in the corpus.

Method and Data

To scrutinize the above intuition, a controlled experiment has been designed, in which particular text samples were assessed independently (one by one) and compared against the corpus. A following procedure was applied: the entire corpus served as a training set, out of which one text at a time was excluded. This temporarily excluded text was further pre-processed: in many iterations, longer and longer samples of randomly chosen words were excerpted (100 independent samples in each iteration), and then tested against the training set. In each iteration, the total number of correctly "guessed" authorial classes – a single value between 0 and 100 – was recorded, resulting in a row of accuracy scores for a given text as a function of its sample size. The same procedure was repeated for each text in the corpus. The above setup does not need to be supplemented by any cross-validation, because the experiment itself is a variant of a leave-one-out cross-validation scenario. Moreover, each text is re-sampled several times, which can be perceived as an additional way of neutralizing potential model overfitting.

The experiments were repeated a few times. Firstly, three different classification methods have been tested: Support Vector Machines (SVM), Nearest Shrunken Centroids (NSC), and a distance-based learner that is routinely used in authorship attribution tests, namely Burrows's Delta (Burrows, 2002). However, Delta was used as a general classification framework combined with a few custom kernels that seem to outperform the original setup. These included Cosine Delta (Evert et al., 2016), min-max measure (Kestemont et al., 2016), Eder's Delta (Eder et al., 2016), and, obviously, the original measure as introduced by Burrows and mathematically justified by Argamon (2011). Secondly, all the tests have been repeated for different vectors of input features, or most frequent words: 100, 200, 300, 500, 750 and 1,000. While the choice of the vectors' lengths was arbitrary, it was aimed to follow usual stylometric scenarios in their various flavors, ranging from a considerably short list of mostly frequent words, to a longish vectors overwhelmed by content words.

The aforementioned method of testing was applied into two roughly similar corpora (one at a time): a corpus of 100 English novels by 33 authors (male and female), covering the years 1840–1940, and a similar corpus of 100 Polish novels. Both corpora, referred to as the Benchmark Corpus of English and the Benchmark Corpus of Polish, have been compiled by Jan Rybicki (pers. comm.). The corpora used in the experiment, as well as the complete code needed to replicate the study, will be available in a GitHub repository.

Results

A lion's share of tested samples revealed a very consistent and clear picture. According to intuition, the performance for short samples falls far beyond any acceptance rate, sometimes showing no correct "guesses" at all. This is followed, however, by a very steep increase of performance which immediately turns into a plateau of statistical saturation, despite the number of analyzed features (frequent words). An example of such a behavior is *The Ambassadors* by Henry James (Figure 1), as well as many other novels by Blackmore, Chesterton, Foster, Lytton, Meredith, Morris, Thackeray, and Trollope. As one can see, the amount of text needed for a reliable attribution is less than 2,000 words (!), an amount radically smaller than the previous study suggests (Eder, 2015). Sometimes the picture is somewhat blurry, nevertheless the same general shape reappears, as in the case of *Felix Holt* by George Elliot (Figure 2). As one can see, using shorter vectors of features requires longer samples to extract the authorial profile.

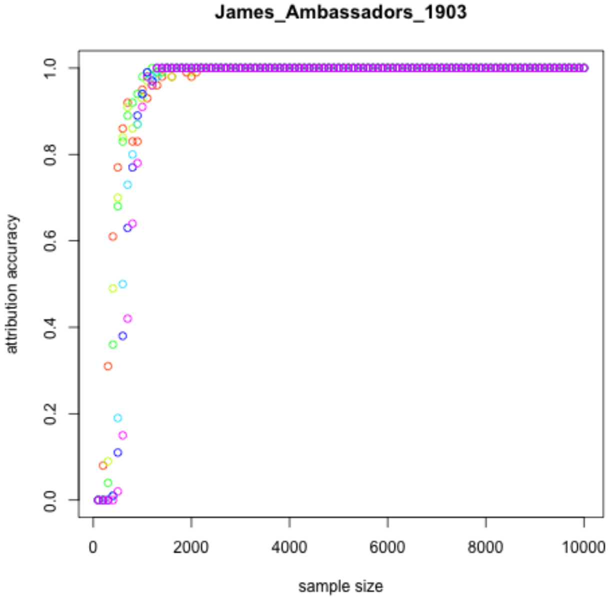

Figure 1: *The Ambassadors* by Henry James contrasted against a corpus of 100 English novels: the attribution accuracy as a function of sample size (in words). Colors represent the results for different vectors of MFWs: 100 (red), 200 (yellow), 300 (green), 500 (cyan), 750 (blue), and 1,000 (violet).

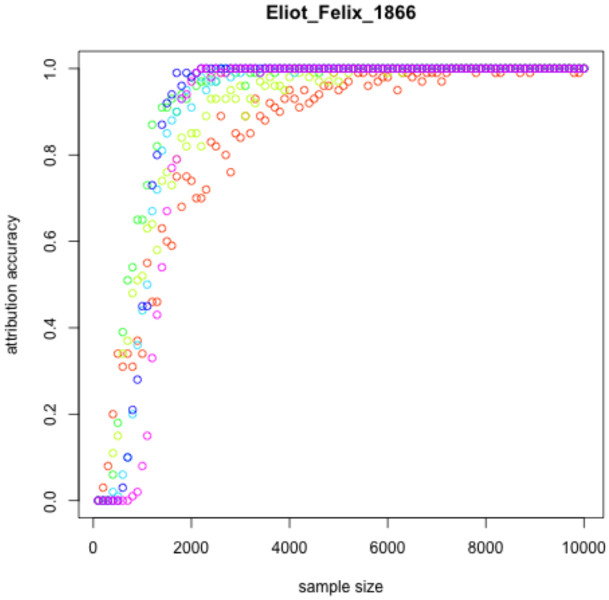

Figure 2: *Felix Holt* by George Eliot: the dependence of authorship recognition and sample size.

Optimistic as they are, however, the results might differ significantly. E.g., in some cases, the statistical saturation does not really take place, even if very long samples are used (Figure 3: scores for *Saints Progress* by John Galsworthy). What is more important, however, the final results additionally depend on the number of analyzed features. In Figure 4, a representative example of this behavior has been shown, namely *Bleak House* by Dickens.

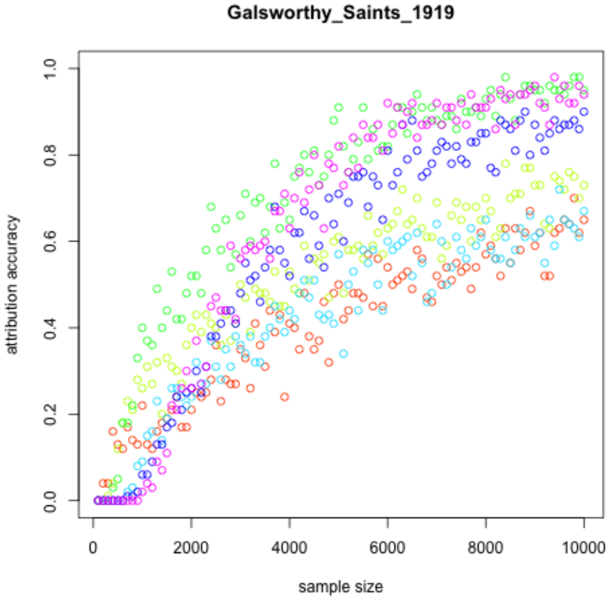

Figure 3: *Saints Progress* by John Galsworthy: the dependence of authorship recognition and sample size.

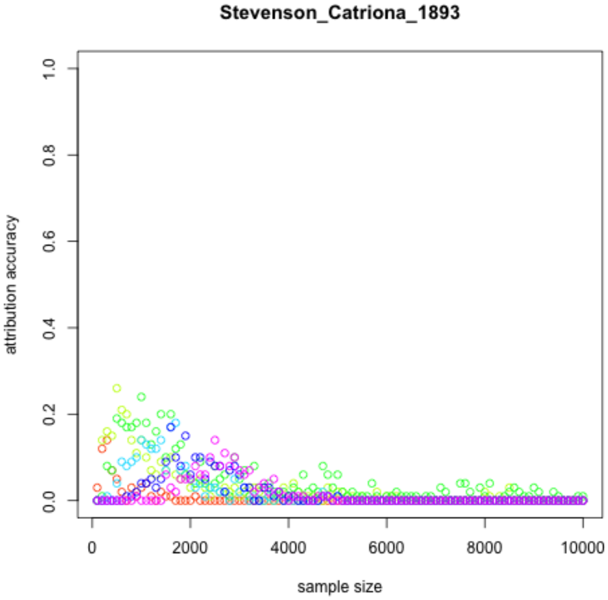

Figure 5: *Catriona* by Robert Louis Stevenson: the dependence of authorship recognition and sample size.

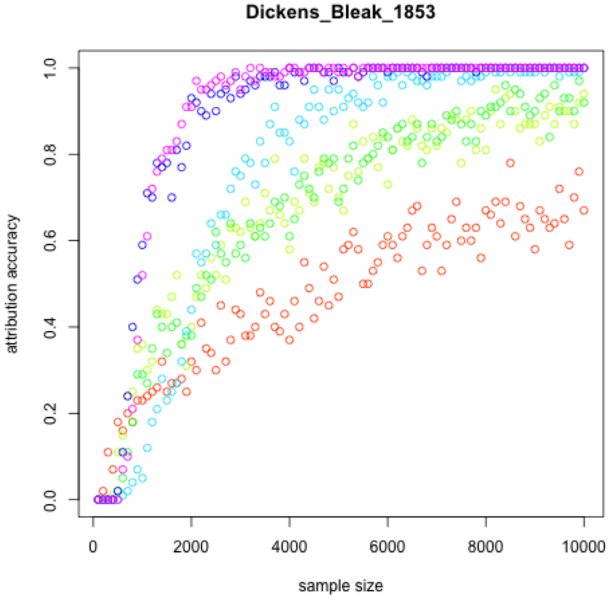

Figure 4: *Bleak House* by Charles Dickens: the dependence of authorship recognition and sample size.

Last but definitely not least, there are a few texts that are never correctly attributed, no matter how long the extracted samples are (Figure 5). The question why some novels were misclassified will be addressed in a separate study. Here, it should be emphasized that such a behavior is unpredictable. Certainly, it can be easily detected, as long as one tests novels of known authorship; it becomes an obstacle, however, when one tries to scrutinize an anonymous text.

Detecting Outliers

The outcome of the above experiment shows that the minimal sample size can be lowered substantially, from ca. 5,000 running words as suggested previously (Eder, 2015), to less than 2,000 words. However, this is true only for those texts that exhibit a clear authorial signal; otherwise the risk of severe misclassification appears. To take advantage of the above results, then, one has to be sure which category an analyzed text belongs to. In a controlled experiment, the task is simple, in a real-case attribution study, however, one has no chance to fine-tune the model by testing the disputed sample against the corpus. What if an anonymous text does not reveal a clear accuracy curve, as the one in Figure 1?

To overcome the sample size issue of unknown texts, an additional measure can be involved to supplement the accuracy scores. (Due to limited space in this abstract, a compact outline of the proposed solution will be presented, rather than a complete algorithm). In the case of misclassification, one would like to know if the wrong response is consistent, or if different classes were assigned chaotically. To address this question, an indicator of consistency would be useful. The Simpson index is a very simple measure of concentration when observations are classified into a certain number of types (Simpson, 1949):

$$\lambda = \Sigma pi2$$

where pi is the proportion of observations belonging to the ith type. The index can be easily adopted to indicate imbalance between assigned classes in supervised classification. To this end, the obtained classification scores (for a given sample size) have to be divided by the total number of trials (in this case, 100). The value 1 reflects purely consistent results, lower values mean that the assigned classes were fuzzy.

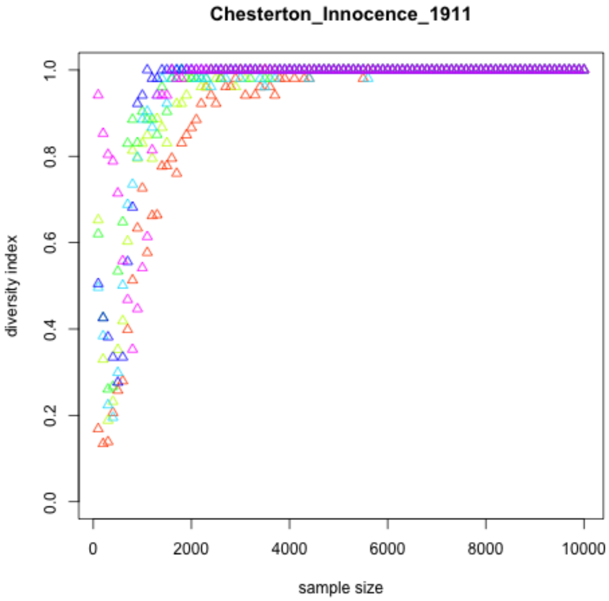

Figure 6: Diversity scores (Simpson index) as a function of sample size.

To make a long story short: the texts that distribute their accuracy curves as in Figure 1 will also exhibit the same shape of the diversity index (see Figure 6). However, when the accuracy scores are low and/or ambiguous, the diversity index might provide a priceless hint. It is especially important when the accuracy scores are consistent (Figure 5), and the Simpson index is not (Figure 7). Instead of being mislead ("Stevenson did not write Catriona", which is not true), we are warned that the classification is inconsistent. Thus, to reliably test a minimal size of a disputed text, one has to take into account two values (accuracy and diversity). The bigger the dispersion between the indices, the smaller the probability that the text is attributable – perhaps a longer sample has to be involved, or a different set of features?

Conclusion

The study was aimed at re-considering the minimum sample size for reliable authorship attribution. The results of the experiments suggest that a sufficient amount of textual data may be as little as 2,000 words in many cases. However, sometimes the authorial fingerprint is so vague, that one needs to use substantially longer samples to make the attribution feasible. A question of some importance is to which category an unknown (disputed) text belongs.

Bibliography

Argamon, S. (2011). Interpreting Burrows's delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2): 131–47.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267–87.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities,* 30(2): 167–82.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal,* 8(1): 107–21.

Evert, S., Jannidis, F., Proisl, T., Thorsten, V., Schöch, C., Pielström, S. and Reger, I. (2016). Outliers or key profiles? Understanding distance measures for authorship attribution. *Digital Humanities 2016: Conference Abstracts.* Kraków: Jagiellonian University & Pedagogical University, pp. 188–91.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4): 405–17.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F. and Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems With Applications*, 63: 86–96.

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.

Luyckx, K. and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1): 35–55.

Mikros, G. K. (2009). Content words in authorship attribution: an evaluation of stylometric features in a literary corpus. In Köhler, R. (ed), *Studies in Quantitative Linguistics,* vol. 5. Lüdenscheid: RAM, pp. 61–75.

Rybicki, J. and Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words?. *Literary and Linguistic Computing,* 26(3): 315–21.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163: 688.

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, 44(2): 790–99.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Proceedings of the Second Asia Conference on Asia Information Retrieval Technology.* (AIRS'05). Berlin, Heidelberg: Springer-Verlag, pp. 174–89.