Frequency-free authorship attribution: Testing the n-gram tracing method

Dr Andrea Nini
University of Manchester

Prof Jack Grieve
University of Birmingham

Introduction

The task of analysing an anonymous text to determine who wrote it is an important application of corpus linguistics, both for forensic and literary problems. In the past years, several methods have been tested with generally good degrees of success, especially when sophisticated computational techniques have been applied to frequency of words or characters (Stamatatos, 2009). The degree of success of these frequency methods, however, tends to be connected to the amount of data available (Luyckx and Daelemans, 2008; Eder, 2015). This is not so much a concern for certain literary questions but forensic cases often tend to be concerned with very small disputed texts. In addition, another important problem for both forensic and literary linguistics is that the results of several techniques that use frequencies are not easy to interpret. For example, although one of the best features for authorship attribution is character 2-grams, it is almost impossible to interpret and understand the underlying linguistic pattern that has led to a particular attribution if this feature is used. This is problematic for forensic purposes because the forensic linguist should be able describe in court the linguistic patterns that contribute to the attribution, as opposed to just the statistical/computational method used. It is also problematic for the literary scholar, as their ultimate interest is often to understand the style of authors. This paper presents a new method of analysis for authorship that tackles these two problems. This method, *n-gram tracing*, introduced in Grieve *et al.* (forthcoming) to solve a specific case, has here been tested more generally using several parameters to establish its usefulness in different contexts.

N-gram tracing and the *shared co-selection coefficient*

The method of *n-gram tracing* is based on measuring how much co-selection of linguistic material is in common between the disputed text and all the texts available from a candidate author. Given a disputed text and a set of candidate authors, each author with *n* number of texts, we can calculate the *shared co-selection coefficient* for author *K*, C_K, as follows

$$C_K = \frac{|\bigcup_{i=1}^{n} \kappa_i \cap Q|}{|Q|}$$

where Q is the set of n-grams of the disputed text and $|Q|$ is its cardinality, or the number of different n-grams in the text; κ_i is the set of n-grams of text *i* of candidate author *K*, with $|\kappa_i|$ being its cardinality and $\bigcup_{i=1}^{n} \kappa_i$ thus being the union of the sets of n-grams for all the texts written by candidate author *K.* A condition for the coefficient to be calculated is that $|\bigcup_{i=1}^{n} \kappa_i| \geq |Q|$, that is, that the data for the candidate author should always be greater than or equal to the questioned data, so that the maximum similarity is 1. The type (character, words, POS, etc) and size of n-grams are parameters that can be chosen depending on the case.

After a coefficient is calculated for each candidate author, the disputed text is attributed to the candidate author with the largest coefficient, which is therefore the candidate author with the highest amount of shared co-selection of n-grams with the author of the disputed text. However, this process is carried out while controlling for the number of word tokens sampled for each author. Since the number of types in a text is correlated with its number of tokens, equal number of tokens must be sampled for each author in order for the coefficients to be comparable.

An additional step in the method of n-gram tracing is to visualise the increase in shared encoding as more data is observed by plotting the coefficients as they are measured at different sample sizes, as done in Grieve *et al.* (forthcoming).

Methodology

A corpus was compiled using data from *Project Gutenberg.* Ten authors were chosen: Austen, Conrad, Dickens, Doyle, Eliot, Hardy, Lawrence, Stevenson, Thackeray, and Wells, and five novels per author were selected so that for each author more than 200,000 word tokens were available.

The parameters for the testing were type of n-grams, size of n-grams, number of word tokens of the disputed text, and number of word tokens of comparison data. Three types of n-grams were tested: word n-grams (1 to 8), character n-grams (1 to 20), and POS n-grams (1 to 15), for a total of 43 feature types. Seven sizes of disputed text were chosen: 25, 50, 100, 200, 300, 500, 1,000, while nine sizes of comparison data were chosen: 100, 1000, 2,000, 5,000, 10,000, 20,000, 50,000, 100,000, 200,000.

The first step of the testing was to create the disputed text by taking a random extract from a random text. This randomly sampled text was then excluded from further analysis and a comparison corpus was created using all the texts for the remaining nine authors plus all the texts minus the one sampled from the author of the disputed extract. By removing the text from which the disputed extract was sampled we can exclude the possibility that the attribution is done by identifying proper nouns belonging to the same novel.

For each test, the extract was attributed to the candidate author with the highest coefficient and the accuracy rate, or the percentage of correct attributions, was calculated. In those cases in which two candidate authors received the same coefficient, the attribution was considered unsuccessful. Because the candidate authors can be ranked by most likely to least likely, the percentage of successful attribution was also calculated for the correct author being within the top three candidates. Additionally, the results were reported for two scenarios: only two candidate authors are available or all 10 candidate authors are available, the former being the average accuracy of all the possible combinations of two authors in the corpus.

Results

Figure 1 shows the results of the tests performed. The vertical axis indicates the max percentage of accurate attributions achieved considering all n-gram types while the horizontal axis shows the increase in comparison data available, from 100 tokens to 200,000 tokens. Each facet shows the results for different sizes of the disputed text, from 25 tokens to 1,000. Three lines are plotted representing the three different tests performed: accuracy for only two candidates, accuracy for ten candidates, and the percentage of the correct author being in the top three most likely candidates.

The graph clearly indicates, as expected, that an increase in available comparison data leads to an increase in accuracy overall. Even a questioned text as short as 25 word can be attributed with 75% accuracy to the correct author among a set of two provided that each author has 200,000 tokens as comparison data. In this extreme scenario, if ten candidates are available, in 65% of the cases the correct author is one in the top three most likely candidates. In the more relatively common forensic scenario of a disputed text of 100 tokens, a comparison data set of 20,000 tokens can already lead to 75% accuracy provided that only two candidates are available. In this two candidate problem, over 90% accuracy is reached starting from a Q texts of 200 tokens with 200,000 tokens of comparison data. When the Q text becomes larger, thus 500 or 1,000 tokens, it is more likely than any test reaches more than 75%-80% accuracy, especially with more than 10,000 tokens of comparison material. Indeed, if more than 20,000 tokens of comparison data are available, then the correct author is within the top three candidates out of ten at least 75% of the times, and, in a two candidate problem, the correct author is identified with even greater accuracy. Near perfection in attribution is achieved, quite predictably, when both relatively long (1,000 tokens) disputed texts and comparison data (200,000 tokens) are available (99.2% that the correct author is in the top three).

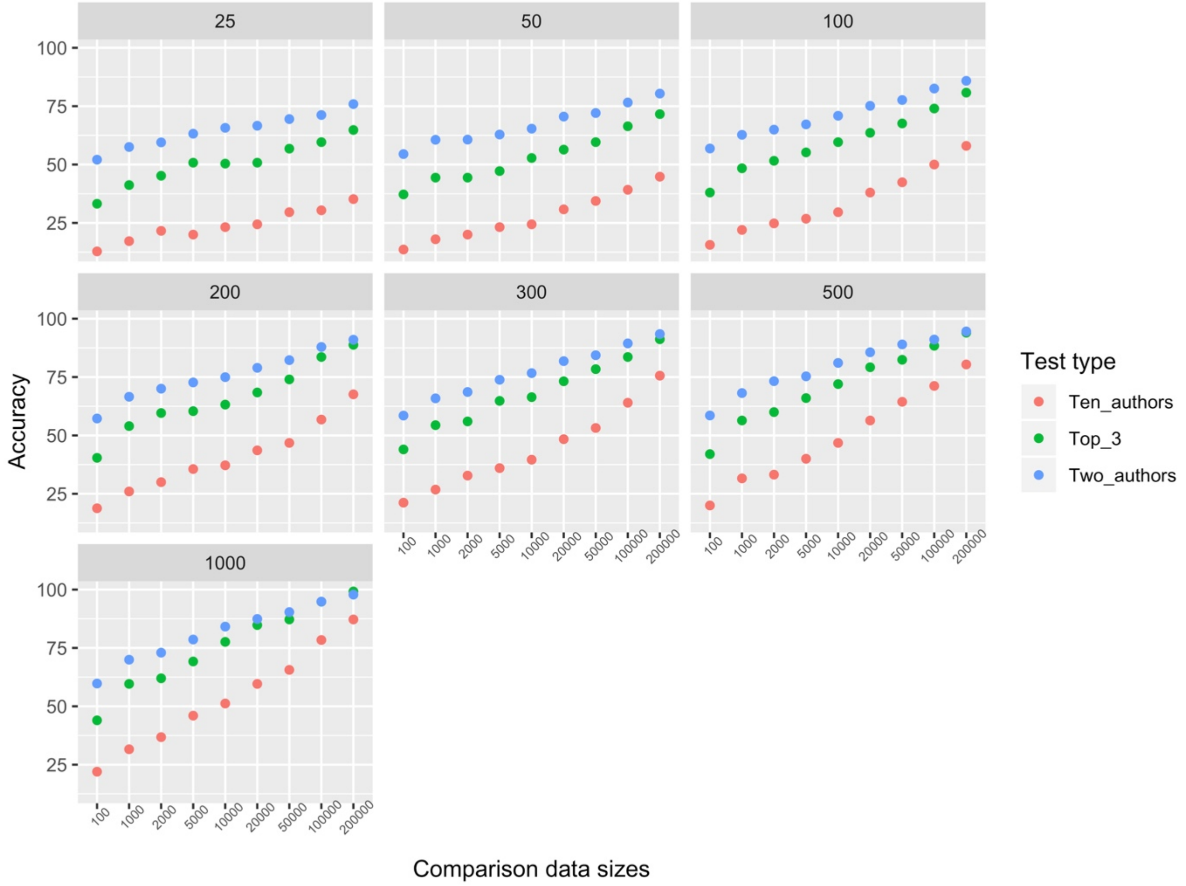

Figure 1 – Max accuracy for all the n-gram types for a series of tests (correct author among ten candidates, correct author between two candidates, correct author among top three most likely candidates) as a function of increasing comparison data available. Each facet represents a different size for the disputed text.

Figure 1 only shows the results for the most successful n-gram types, which were firstly long character n-grams (*n* = 9, 10, 11, 12, 13), followed by short word n-grams (*n* = 2, 3), followed by POS n-grams of medium length (*n* = 5, 6).

Conclusions

The method tested in this study shows preliminary evidence that in authorship attribution the size of comparison data is much more important than the size of questioned data and that information about frequency is useful but not essential. This fact is particularly important for both forensic and literary linguistics because this frequency-free method is much more amenable to interpretation, as it relies on large sets of combinations of rare or uncommon features as opposed to a small set of very frequent function words or characters. Theoretically, these findings also lend support to the theory of *idiolectal co-selection* (Coulthard, 2004), which says that individuality arises from the co-selection of different items.

References

Coulthard, M. (2004) 'Author identification, idiolect, and linguistic uniqueness', *Applied Linguistics*, 25, pp. 431–447.

Eder, M. (2015) 'Does size matter? Authorship attribution, small samples, big problem', *Digital Scholarship in the Humanities*, 30(2), pp. 167–182. doi: 10.1093/llc/fqt066.

Grieve, J., Chiang, E., Clarke, I., Gideon, H., Heini, A., Nini, A. and Waibel, E. (no date) 'Attributing the Bixby Letter using n-gram tracing', *Digital Scholarship in the Humanities*.

Luyckx, K. and Daelemans, W. (2008) 'Authorship attribution and verification with many authors and limited data', in *Proceedings of the Twenty-Second International Conference on Computational Linguistics (COLING 2008)*. Manchester, UK: ACL, pp. 513–520.

Stamatatos, E. (2009) 'A survey of modern authorship attribution methods', *Journal of the American Society for Information Science and Technology*, 60(3), pp. 538–556. doi: 10.1002/asi.21001.