Authorship Identification of Short Texts Using Only Syntactic Features

The treebanking tools developed in association with Alpheios and Perseids contribute directly to significant progress in an important problem in the field of computational authorship attribution, namely the classification of short text passages. While machine-based stylometry has a strong record of success for the identification of authors of large texts, how to bring a similar rate of accuracy to the analysis of shorter texts is an open question. Eder (2015) has systematically investigated small text size in multi-class categorization and has suggested that, speaking generally, 5000 words is a minimum for useful results. Later, considering the issue more closely, Eder has found that 2000 words may be sufficient when the texts in question "exhibit a clear authorial signal" (Eder 2017). Of course, even this limit is too large for many authorship problems. In classics, for example, evaluating the authorial signature of fragmentary texts is an important problem, but these passages are usually much smaller than the suggested 2000 words. This paper will demonstrate how data from the Ancient Greek and Latin Dependency Treebank (AGLDT) can be used for accurate identification of texts beneath the suggested minimum.

Most attempts at computational authorship attribution depend to a greater or lesser degree on measures of vocabulary richness. With this approach, the particular words chosen by a writer are taken to be constitutive of a stylometric "authorial signature" allowing texts to be distinguished. A principle drawback of measuring vocabulary is that the lexicon of a text is highly sensitive to factors besides authorship: genre, topic, addressee, etc. For this reason, researchers often prefer function words—prepositions, conjunctions, articles, and the like—rather than content words as input for their algorithms. The method adopted here will go farther along this line than usual: we avoid completely any consideration of vocabulary richness and pay

no notice the occurrence of individual words. Instead, syntactic features, representing both shallow and deep linguistic characteristics, make up the data to be analyzed.

The texts for our experiment, drawn from the AGLDT, are ancient Greek works in both prose and verse. The corpus selected for analysis here contains 28 texts by 13 different authors. These works consist of 582,487 tokens in total. The particular units selected for analysis are constructed by combining two levels of information: the morphological annotation and dependency relationship for each word and the same features for each word's parent. Paradoxically, the next step after the creation of plausible units to represent the syntactic characteristics of a text is the ruthless culling of these same units, which can quickly multiply beyond the capacity of even a fast computer. The details of these two crucial stages of analysis, "feature engineering" and "feature reduction," will be briefly outlined in the presentation.

The greatly reduced (899 items) set of features serves as input to standard classification algorithms, logistic regression (LR) and the support vector machine (SVM). Both are widely used and available. In the experiment presented here, we first divide the input texts into 2000 word segments (the minimum size suggested for analysis based on vocabulary). 90% of the segments are used to train the algorithm; the remaining 10% are kept out for testing. Multiple iterations of training/testing are performed. After each round, the size of the segments is reduced by 100.

The results of testing were very good. For input texts of 2000 words, both LR and SVM produced practically no identification errors (success rate, SVM 0.9998, LR 1.000). For input texts of 1000 words, the success rate had barely dropped: SVM 0.9991, LR 0.9989). At 500 words, the relevant numbers are SVM 0.9958, LR 0.996. Even with text as small as 100 words, the syntactic features derived from the AGLDT still yield the correct author identity more than 9

times out of 10: SVM 0.943, LR 0.9433. These success rates, high by comparison to previous experiments, should not be seen as exceptional for the method described here. To prepare the input data, the simplest and most naïve approach to feature reduction was chosen. There is good reason to suppose that a more sophisticated and considered approach would increase success significantly. Thus we can have some confidence that we are close to the practical application of classification techniques to texts smaller than 100 words, at least in ancient Greek.

Bibliography

Maciej Eder. 2015. “Does size matter” Authorship attribution, small samples, big problem.” *Digital Scholarship in the Humanities,* 30.2, pp. 167-182

--------. 2017. “Short samples in authorship attribution: a new approach.” https://pdfs.semanticscholar.org/8e55/de66c9c8060cd19ecac8bac25b311ad42184.pdf Accessed 8-14-2018.