Psycholinguistically Plausible Events and Authorship Attribution

Juola, Patrick
juola@mathcs.duq.edu
Duquesne University

Authorship attribution (Juola, 2008) is an important emerging subdiscipline of digital scholarship, but it suffers from a lack of connection to other areas and disciplines, which in turn strongly limits both applicability and uptake. It is now unquestionable that computers can infer authorship attributes with high accuracy, but the accurate inference processes tend not to inform us about the actual authors (Craig, 1999). Among the best methods, for example, are the analysis of the most frequent function words such as prepositions (e.g., Binongo, 2003), but knowing that a particular person uses the word "above" a lot tells us little about that person. Argamon (2006) has provided a theoretical analysis of one particular method, but in the unfamiliar and "inhuman" language of statistics, which again sheds little light on authorial language and authorial thought. By contrast, studies of gender differences in language (e.g., Coates, 2004) offer not only lists of differences, but explanations in terms of the social environment.

This is in marked contrast to some of the early (pre-computer) work in authorship analysis, which attemped to infer authorship on the basis of personality traits or psychological attributes. For example, one of the oft-suggested measures is vocabulary size, which we can easily associate with both high intelligence (a personal trait) as well as high education (a background trait). This idea can be attributed both to Simpson (1949) and Yule (1944) as well as to Talentire (1976) [which admittedly is not pre-computer]. Similarly, average word length has been often proposed [going back to De Morgan (1851)] but never successful.

Why? Why the apparent disconnect between the useful measures (such as preposition count) and meaningful measures like vocabulary richness? And in particular, why does this disconnect persist when we can find both linguistic patterns that predict personality (Argamon et al, 2005; Nowson and Oberlander, 2007) and well as medically useful linguistic diagnostics (Brown et al, 2005). We suggest two possibilities; first, that the meaningful measures proposed may not be sufficiently fine-grained, and second, that the statistical measures performed lose too much information. As an example of the first, consider that very few words, even in high-level educated writing, exceed eight letters, meaning that "word length" is an extremely coarse-grained discretization of language. Similarly, the standard method of calculating "averages" (or even means and variances) reduces the entire data set for a given author to two numbers. Many authors have suggested (and recent findings tend to support) that multivariate analysis methods should work better for authorship attribution.

In this paper, we explore a set of multivariate analyses of well-established psycholinguistic variables. The English Lexicon Project (Balota et al, 2007) provides standardized behavioral data for a set of approximately 40,000 words, including average time for lexical decision tasks (seeing a string of characters on the screen and determining whether or not they form a word), and naming time (seeing a set of letters on the screen and naming the word they form). These are widely regarded as measures of the cognitive load involved in processing that particular word, i.e. a measure of the mental "difficulty" of that word. Following similar logic to De Morgan and Yule, we assume that some people (smarter people?) will be more comfortable using "difficult" words, and that difficulty is more appropriately measured via behavioral data than via either frequency or length.

However, rather than focusing purely on average difficulty, we apply more complex multivariate statistics to the data distribution, for example, by calculating the Kolmogorov-Smirnoff distances between the distributions, a distance that can be substantial even in instances where the means and variances of the data sets are identical. The JGAAP software package (Juola, 2009) provides many different combinations of analysis methods and preprocessing, allowing us to provide a fairly comprehensive discussion of the accuracy

and usefulness of these measurements in comparison with control techniques such as simple lexical statistics.