JGAAP : A System for Comparative Evaluation of Authorship Attribution

Patrick Juola // Duquesne University, Pittsburgh PA // juola@mathcs.duq.edu

"Nontraditional" authorship attribution, the process of inferring authors or authorial traits via statistical analysis of text, has seen something of a resurgence in recent years (see Juola, 2008 for a survey). Many methods have been proposed, and most work (after a fashion). However, the sheer number of methods proposed and the sheer number of test corpora developed makes it difficult to identify any clear "best practices" or particularly accurate techniques. The techniques that appear to score best, such as Support Vector Machines, can be difficult to understand and hard for non-technical specialists to apply.

We will describe and demonstrate a program to address some of these issues. JGAAP (Java Graphical Authorship Attribution Program, available for download from http://www.jgaap.com) is a freely available Java program to perform authorship attribution. Using Java provides two major advantages. First, it is a write-once, run-anywhere system that will work without modification on any system we have tested (including Windows, OS X, and several flavors of Linux). Second, using objects and inheritance has allowed us to create an easily extensible framework to allow researchers to apply their own method-of-choice. Specifically, the JGAAP framework takes "Documents," essentially strings generated from files of interest by any class instantiating the proper interface, converts them to "EventSets," again defined as any class implementing the proper interface, then analyzes them. Currently available EventSets include characters, words, syllables, reaction times, n-grams of any of the above, or "most common" versions of any of the above. Currently available analytic methods include nearest neighbor using a variety of distances such as Euclidean histogram distance, nominal Kolmogorov-Smirnov distance, and cross-entropy (Juola, 1997) as well as Support Vector Machines. We are currently working on a catalogue of effective and ineffective combinations based on the AAAC corpus (Juola, 2004).

More importantly, we have also demonstrated the ease of extensibility. The existing JGAAP framework focuses primarily on English documents (or at least documents written in languages using the Latin-1 character set and whitespace-separated words); Zhao and Juola (2008) have produced a simple extension to permit it to handle documents in Chinese. Modifying only the EventSets (in essence, adding new sets to handle word segmentation), we were able to apply JGAAP to authorship attribution in Chinese, using existing Event Sets (such as n-grams) and existing analysis methods. It was relatively easy to establish, for example, that nominal Kolmogorov-Smirnov appears to be the most accurate distance function for Chinese authorship attribution, and that single characters and words segmented via Forward Maximum Matching were the most accurate event sets.