JDHCS 2009
Volume 1 Number 1

JGAAP: A System for Comparative Evaluation of Authorship Attribution

Patrick Juola, Department of Mathematics and Computer Science, Duquesne University

Introduction

In 2004, Potomac Books published *Imperial Hubris : Why the West is Losing the War on Terror*. Drawing on the author's extensive personal experience, the book described the current situation of the American-led war on terror and argued that much United States foreign policy was misguided.

Unfortunately, the author was anonymous. Finding out who this author was would go a long way to assessing his or her credibility. The task of determining from a document's text who wrote it has a long history but no clearly-defined best solution. We present a software system (JGAAP) to help the search for this best solution.

Problem Statement

"Authorship Attribution" is a long-standing problem in text analysis, and indeed in the humanities generally. Questions such as "Did William Shakespeare write the plays attributed to him?" or "Did the *Iliad* and the *Odyssey* have the same (single) author?" have generated literally centuries of discussion. These questions are traditionally answered—or at least analyzed—via close reading by scholars with expertise in the authors, languages, and fields under discussion. With the development of modern statistics and later modern computers, some scholars, among them de Morgan and Mendenhall have suggested that an analysis of statistical properties of writings — attributes like word or sentence length, vocabulary richness, or distribution of words — might settle these questions more objectively and accurately. The most famous example of this is the work of Mosteller and Wallace on the authorship of the *Federalist Papers*, where they used Bayesian statistics on a set of common function words to determine that the "disputed essays" had been written by Madison.

Since Mosteller and Wallace, there has been a virtual explosion of proposed techniques for doing this type of "nontraditional" or "statistical" authorship attribution; Rudman (1998) has identified more than one thousand proposed techniques in the literature. In general, the usual study goes as follows: a researcher identifies a potential "fingerprint" that is characteristic of an author he or she is interested in, collects an *ad hoc* group of texts including genuine authorial texts and distractors known to be by different authors, and demonstrates via experiment that the fingerprint can in fact distinguish the two groups with accuracy significantly better than chance as measured by the usual t-tests and p-values. Juola (2006) provides a recent survey both of studies and techniques.

Applications

Applications of this problem extend far beyond the literary salon, however. In the 2008 presidential election, one of the accusations leveled against candidate Barack Obama was that he had been closely associated with the former terrorist William Ayers; journalist Jack Cashhill offered in support a traditional stylistic analysis purporting to prove that Ayers had, in fact, ghost-written President Barack Obama's first book. But is this true? Cashhill himself had little of the high-level expertise we typically expect of, say, a Shakespearean scholar, and his arguments were colored through with his

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

personal biases. It may or may not be significant that he was unable to offer any statistical or computation evidence in support of his view, with the notable exception of a single study using a long-discredited method.

Law enforcement and forensic scientists are similarly interested in these methods; Chaski reports on a court case where a body was discovered near a computer with an apparent suicide note typed into it. In the case of a hand-written note, of course, handwriting specialists could be called in to verify that the note was in the deceased's handwriting, but one flat-ASCII "A" looks identical to any other. What was needed instead was an analysis based on writing style, and Chaski was able to prove that the deceased did not write the note, and that murder had been committed.

Key to such legal uses is that the analysis must be accurate enough to be relied upon, and in fact, the Federal Rules of Evidence more or less demand that any such "scientific" evidence be independently proven accurate before it is admissible in court. One of the problems with Cashhill's analysis of the Obama/Ayers question is that the method used ("cusum") is known to be highly unreliable and has in fact had some quite serious and well-publicized failures.[1] The unfortunate situation is that a scholar with a question of authorship is now faced with a bewildering array of possible methods to use, most of which have been proven to be "better than chance" (as though that was meaningful) but with little guidance as to how much better and under what circumstances maximum accuracy can be achieved.

Testing For Accuracy

With this situation (and more than 1000 techniques to choose from), what is necessary is a common framework and comparative evaluation to give guidance among candidate techniques. The JGAAP program wraps a user-friendly GUI around a simple three-phase model of authorship attribution, permitting the user to select from a variety of preprocessor, units of analysis, and specific analytic methods.

These three phases include:

- Canonization: Juola uses this term both for data type preprocessors (such as converting HTML or PDF documents into plaintext) as well as preprocessors that "canonize" the document by neutralizing uninformative or distracting variations such as variations in spacing, capitalization, and punctuation.

- Event Set Generation: The "canonical" document is then broken down into a sequence (Vector) of "Events." This may involve simple tokenization (as in breaking the document into words or characters), tokenization and recombination (as in the generation of word/character bigrams or trigrams), or substitution (stemming, POS tagging, or replacing each word with its length or frequency in a large neutral corpus).

- Analysis: These Events can then be analyzed using a variety of standard and not-so-standard classification methods, including nearest-neighbor methods (distances are calculated between each pair of documents and documents of unknown authorship are assigned to author of the closest document with a known author), Principle Component Analysis, Linear

[1] Juola 2006.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

> Discriminant Analysis, Support Vector Machines, and Naive Bayesian Analysis. Several variations are possible for many of these; for example, Support Vector Machine analysis can be done with a variety of kernels, many different prior probabilities can be used with NB, and of course, the definition of "distance" or "closest" can vary with a topologist's whim, as will be seen later.

Taking all possible combinations together, we estimate that JGAAP as distributed is currently capable of more than 20,000 different authorship analysis techniques. Furthermore, as an open-source project written in Java, it is easy enough for us (or for any other interested group) to add additional methods. JGAAP takes advantage of Java's object-oriented nature by defining, for example, a generic ("abstract") Preprocessor class, and any additional preprocessors that might be wanted [converting Word documents to plaintext, eliminating internal quotations as suggested by Rudman (2005), eliminating or neutralizing proper names, and so forth) can be implemented by user-defined classes that implement ("extend") the Preprocessor. The overall JGAAP framework simply creates objects of appropriate type and operates on them with the functions shared by all Preprocessors.

Putting it all together

With 20,000 possible combinations,we expect that more than 19,000 oft hem will not prove to be the most accurate or more generally, best practices. The simple question is which? We are in the process of testing many different techniques using the AAAC corpus[2] in search of the few methods that do work well, and in particular, looking for families of techniques that seem to perform well generally.

In order to do this efficiently, we have been forced to make substantial modifications of the JGAAP software. Juola's original program envisioned the primary user of JGAAP as a researcher with a specific (small) dataset to analyze, using a small set of well-defined and well-established methods who is more comfortable with a GUI than a command line. However, running large-scale experiments involving hundreds of documents and potentially thousands of methods makes selecting these through radio buttons impractical. We therefore modified the source code of JGAAP to allow options (such as canonicizers and event set generators as well as files to analyze) to be set from the command line and generated large-scale shell scripts to allow a user to perform these large-scale tests. Our modifications have been incorporated into the currently distributed version of JGAAP in case any other researchers are interested in similar experiments.

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 white space separated words); Zhao and Juola[3] have produced a simple extension to permit it to handle documents in Chinese. Modifying only the Event Sets (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

[2] Juola 2004.

[3] As noted in their co-written 2008 article.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

distance function for Chinese authorship attribution, and that single characters and words segmented via Forward Maximum Matching were the most accurate event sets.

Although JGAAP was designed and produced as a system for authorship attribution, it has much broader applications as well. The system as a whole is completely agnostic with regard to the type of differences it is asked to infer; it could as easily be used to classify documents by age, authorial gender[4], genre, or many other categories. As a simple example, we note that the newly available KRYS I Corpus for Genre Classification Research[5] provides more than 6000 documents labeled with their genres, and a similar large-scale experiment (indeed, much larger than the 98 documents in the AAAC) could establish the parameters of good genre identification methods, and whether or not those parameters are the same as authorship identification. We therefore submit that the JGAAP program and framework maybe a very useful tool for such analysis.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. OCI-0721667. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Argamon, S. and Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *Proceedings of ACH/ALLC 2005*, Victoria, BC. Association for Computing and the Humanities.

Chaski, C. 2005. Who's at the Keyboard: Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4, no. 1, http://www.ijde.org/ (accessed May 31, 2007).

de Morgan, A. 1851. Letter to Rev. Heald 18/08/1851. In *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters*, edited by Sophia Elizabeth de Morgan.

Juola, P. 2004. Ad-hoc authorship attribution competition. Paper presented at the *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities* (ALLC/ACH 2004) in Göteborg, Sweden.

Juola, P. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval* 1:3.

Mendenhall, T.C. 1887. The characteristic curves of composition. *Science* IX: 237–49.

Mosteller, F. and D. L. Wallace. 1964. *Inference and Disputed Authorship:* The Federalist. Reading: Addison-Wesley.

---

[4] Argamon and Levitan 2005.

[5] http://www.krys-corpus.eu

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Rudman, J. 1998. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities* 31:351–365.

Rudman, J. 2005. The non-traditional case for the authorship of the twelve disputed *Federalist Papers*: A monument built on sand. Paper presented at the *Association for Computing and the Humanities* in *Proceedings of ACH/ALLC 2005* in Victoria, BC.

Zhao, M. and Juola. 2008. A Chinese version of an authorship attribution analysis program. Paper presented in *Proceedings of Digital Humanities 2008* in Oulu, Finland.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago