Proving and Improving Authorship Attribution Technologies

Patrick Juola* and **John Sofko**
Duquesne University
Pittsburgh, PA 15282
UNITED STATES OF AMERICA
juola@mathcs.duq.edu and sofko936@hotmail.com

Abstract

Who wrote *Primary Colors*? Can a computer help us make that call? Despite a century of research, statistical and computational methods for authorship attribution are neither reliable, well-regarded, widely-used, or well-understood. This paper presents a survey of the current state-of-the-art as well as a framework for uniform and unified development of a tool to apply the state-of-the-art, despite the wide variety of methods and techniques used.

1 Introduction

Determining the author of a particular piece of text has been a methodological issue for centuries. Questions of authorship can be of interest not only to humanities scholars, but in a much more practical sense to politicians, journalists, and lawyers. In recent years, the development of improved statistical techniques (Holmes, 1994) in conjunction with the wider availability of computer-accessible corpora (Nerbonne, 2004) has made the automatic inference of authorship (variously called "authorship attribution" or more generally "stylometry") at least a theoretical possibility, and research in this area has expanded tremendously. From a practical standpoint, acceptance of this technology is dogged by many issues — epistemological, technological, and political — that limit and in some cases prevent its wide acceptance.

This paper presents a framework for development and analysis to address these issues. In particular, we discuss two major usability concerns, accuracy and user-friendliness. In broad terms, these concerns can only be addressed by expansion of the number of clients (users) for authorship attribution technology. We then present a theoretical framework for description of authorship attribution to make it easier and more practical for the development and improvement of genuine off-the-shelf attribution solutions.

2 Background

With a history stretching to 1887 (Mendenhall, 1887), and 3,520 hits on Google[1], it is apparent that statistical/quantitative authorship attribution is an active and vibrant research area. With nearly 120 years of research, it is surprising that it has not been accepted by relevant scholars : "Stylometrics is a field whose results await acceptance by the world of literary study by and large."[2] This can be attributed at least partially to a limited view of the range of applicability, to a history of inaccuracy, and to the mathematical complexity (and corresponding difficulty of use) of the techniques deployed.

For example, and taking a broad view of "stylometry" to include the inference of group characteristics of a speaker, the story from Judges 12:5–6 describes how tribal identity can be inferred from the pronunciation of a specific word (to be elicited). Specifically, the Ephramites did not have the /sh/ sound in their dialect, and thus pronounced words with such sounds differently than the Gileadites. A more modern version of such *shibboleths* could involve specific lexical or phonological items; a person who write of sitting on a "Chesterfield" is presumptively Canadian, and an older Canadian at that (Easson, 2002). (Wellman, 1936)[p. 114] describes how an individual spelling error — an idiosyncratic spelling of "toutch" was elicited and used in court to validate a document for evidence.

At the same time, such tests cannot be relied upon. Idiosyncratic spelling or not, the word "touch" is rather rare [86 tokens in the million-word Brown corpus (Kučera and Francis, 1967)], and although one may be able to elicit it in a writing produced on demand, it's less likely that one will be able to find it independently in two different samples.

[1] Phrasal search for "authorship attribution," July 30, 2004

[2] Anonymous, personal communication to Patrick Juola, 2004

People are also not consistent in their language, and may (mis)spell words differently at different times; often the tests must be able to handle distributions instead of mere presence/absence judgements. Most worryingly, the tests themselves may be inaccurate [see especially the discussion of CUSUM (Farringdon, 1996) in (Holmes, 1998)], rendering any technical judgement questionable, especially if the test involves subtle statistical properties such as "vocabulary size" or "distribution of function words," concepts that may not be immediately transparent to the lay mind.

Questions of accuracy are of particular importance in wider applications such as law. The relevance of a document (say, an anonymously libelous letter) to a court may depend not only upon who wrote it, but upon whether or not that authorship can be demonstrated. Absent eyewitnesses or confessions, only experts, defined by specialized knowledge, training, experience, or education, can offer "opinions" about the quality and interpretation of evidence. U.S. law, in particular, greatly restricts the admissibility of scientific evidence via a series of epistemological tests[3]. The *Frye* test states that scientific evidence is admissible only if "generally accepted" by the relevant scholarly community, explicitly defining science as a consensus endeavor. Under *Frye*, (widespread) ignorance of or unfamiliarity with the techniques of authorship attribution would be sufficient by itself to prevent use in court. The *Daubert* test is slightly more epistemologically sophisticated, and establishes several more objective tests, including but not limited to empirical validation of the science and techniques used, the existence of an established body of practices, known standards of accuracy (including so-called type I and type II error rates), a pattern of use in non-judicial contexts, and a history of peer review and publication describing the underlying science.

At present, authorship attribution cannot meet these criteria. Aside from the question of general acceptance (the quote presented in the first paragraph of this section, by itself, shows that stylometrics couldn't pass the *Frye* test), the lack of standard practices and known error rates eliminates stylometry from *Daubert* consideration as well.

3 Recent developments

To meet these challenges, we present some new methodological and practical developments in the field of authorship attribution. In June 2004, ALLC/ACH hosted an "Ad-hoc Authorship Attribution Competition"(Juola, 2004a) as a partial response to these concerns. Specifically, by providing a standardized test corpus for authorship attribution, not only could the mere ability of statistical methods to determine authors be demonstrated, but methods could further be distinguished between the merely "successful" and "very successful." (From a forensic standpoint, this would validate the science while simultaneously, establishing the standards of practice and creating information about error rates.) Contest materials included thirteen problems, in a variety of lengths, styles, genres, and languages, mostly gathered from the Web but including some materials specifically gathered to this purpose. Two dozen research groups participated by downloading the (anonymized) materials and returning their attributions to be graded and evaluated against the known correct answers.

The specific problems presented included the following:

- *Problem A* (English) Fixed-topic essays written by thirteen Duquesne students during fall 2003.
- *Problem B* (English) Free-topic essays written by thirteen Duquesne students during fall 2003.
- *Problem C* (English) Novels by 19th century American authors (Cooper, Crane, Hawthorne, Irving, Twain, and 'none-of-the-above'), truncated to 100,000 characters.
- *Problem D* (English) First act of plays by Elizabethan/Jacobean playrights (Johnson, Marlowe, Shakespeare, and 'none-of-the-above').
- *Problem E* (English) Plays in their entirety by Elizabethan/Jacobean playrights (Johnson, Marlowe, Shakespeare, and 'none-of-the-above').
- *Problem F* ([Middle] English) Letters, specifically extracts from the Paston letters (by Margaret Paston, John Paston II, and John Paston III, and 'none-of-the-above' [Agnes Paston]).
- *Problem G* (English) Novels, by Edgar Rice Burrows, divided into "early" (pre-1914) novels, and "late" (post-1920).
- *Problem H* (English) Transcripts of unrestricted speech gathered during committee meetings, taken from the *Corpus of Spoken Professional American-English.*
- *Problem I* (French) Novels by Hugo and Dumas (pere).

[3] *Frye vs. United States*, 1923; *Daubert vs. Merrill Dow*, 1993.

- *Problem J* (French) Training set identical to previous problem. Testing set is one *play* by each, thus testing ability to deal with cross-genre data.

- *Problem K* (Serbian-Slavonic) Short excerpts from *The Lives of Kings and Archbishops*, attributed to Archbishop Danilo and two unnamed authors (A and B). Data was originally recived from Alexsandar Kostic.

- *Problem L* (Latin) Elegaic poems from classical Latin authors (Catullus, Ovid, Propertius, and Tibullus).

- *Problem M* (Dutch)
  Fixed-topic essays written by Dutch college students, received from Hans van Halteren.

The contest (and results) were surprising at many levels; some researchers initially refused to participate given the admittedly difficult tasks included among the corpora. For example, Problem F consisted of a set of letters extracted from the Paston letters. Aside from the very real issue of applying methods designed/tested for the most part for modern English on documents in Middle English, the size of these documents (very few letters, today or in centuries past, exceed 1000 words) makes statistical inference difficult. Similarly, problem A was a realistic exercise in the analysis of student essays (gathered in a freshman writing class during the fall of 2003) – as is typical, no essay exceeded 1200 words. From a standpoint of literary analysis, this may be regarded as an unreasonably short sample, but from a standpoint both of a realistic test of *forensic* attribution, as well as a legitimately difficult problem for testing the sensitivity of techniques, these are legitimate.

Results from this competition were heartening. (“Unbelievable,” in the words of one contest participant.) The highest scoring participant was the research group of Vlado Keselj, with an average success rate of approximately 69%. (Juola’s solutions, in the interests of fairness, averaged 65% correct.) In particular, Keselj’s methods achieved 85% accuracy on problem A and 90% accuracy on problem F, both acknowledged to be difficult and considered by many to be unsolvably so.

More generally, all participants scored significantly above chance. Perhaps as should be expected, performance on English problems tended to be higher than on other languages. Perhaps more surprisingly, the availability of large documents was not as important to accuracy as the availability of a large number of smaller documents, perhaps because they can give a more representative sample of the range of an author’s writing. Finally, methods based on simple lexical statistics tended to perform substantially worse than methods based on N-grams or similar measures of syntax in conjunction with lexical statistics. We continue to examine the detailed results in an effort to identify other characteristics of good solutions.

4 New technologies

The variation in these techniques can make authorship attribution appear to be an unorganized mess, but it has been claimed that under an appropriate theoretical framework (Juola, 2004b), many of these techniques can be unified, combined, and deployed. The initial observation is that, broadly speaking, all known human languages can be described as an unbounded sequence chosen from a finite space of possible events. For example, the IPA phonetic alphabet (Ladefoged, 1993) describes an inventory of approximately 100 different phonemes; a typewriter shows approximately 100 different Latin-1 letters; a large dictionary will present an English vocabulary of 50–100,000 different words. An (English) utterance is “simply” a sequence of phonemes (or words).

The proposed framework postulates a three-phase division of the authorship attribution task, each of which can be independently performed. These phases are :

- Canonicization — No two physical realizations of events will ever be exactly identical. We choose to treat similar realizations as identical to restrict the event space to a finite set.

- Determination of the event set — The input stream is partitioned into individual non-overlapping “events.” At the same time, uninformative events can be eliminated from the event stream.

- Statistical inference — The remaining events can be subjected to a variety of inferential statistics, ranging from simple analysis of event distributions through complex pattern-based analysis. The results of this inference determine the results (and confidence) in the final report.

As an example of how this procedure works, we consider a method for identifying the language in which a document is written. The statistical distribution of letters in English text is well-known [see any decent cryptography handbook, including (Stinson, 2002)]. We first canonicize the document by identifying each letter (an italic *e*, a boldface **e**, or a capital E should be treated identically) and producing a transcription. We then identify each letter as a separate event, eliminating all non-letter characters such as numbers or punctuation. Finally, by compiling an event histogram and comparing it with the known distribution, we can determine a probability that the document was written in English. A similar process would treat each *word* as a separate event (eliminating words not found in a standard lexicon) and comparing event histograms with a standardized set such as the Brown histogram (Kučera and Francis, 1967). The question of the comparative accuracy of these methods can be judged empirically.

The Burrows methods (Burrows, 1989; Burrows, 2003) for authorship attribution can be described in similar terms. After the document is canonicized, the document is partitioned into words-events. Of the words, most words (except for a chosen few function words) are eliminated. The remainder are collected in a histogram, and compared statisically to similar histograms collected from anchor documents. (The difference between the 1989 and 2003 methods is simply in the nature of the statistics performed.)

Even the "toutch" method can be so described; after canonicization, the event set of words, specifically, the number of words spelled "toutch." If this set is non-empty, the document's author is determined.

This framework also allows researchers both to focus on the important differences between methods and to mix and match techniques to achieve the best practical results. For example, (Juola and Baayen, 2003) describes two techniques based on cross-entropy that differ only in their event models (words vs. letters). Presumably, the technique would also generalize to other event models (function words, morphemes, parts of speech), and and similarly other inference techniques would work on a variety of event models. It is to be hoped that from this separation, researchers can identify the best inference techniques and the best models in order to assemble a sufficiently powerful and accurate system.

5 Demonstration

The usefulness of this framework can be shown in a newly-developed user-level authorship attribution tool. This tool coordinates and combines (at this writing) four different technical approaches to authorship attribution (Burrows, 1989; Juola, 1997; Burrows, 2003; Kukushkina et al., 2000; Juola, 2003).

Written in Java, this program combines a simple GUI atop the three-phase approach defined above. Users are able to select a set of sample documents (with labels for known authors) and a set of testing documents by unknown authors. The user is also able to select from a menu of event selection/preprocessing options and of technical inference mechanisms. Currently supported, for example, are three different choices — a vector of all the letters appearing in the sample/testing documents, a vector of all *words* so appearing, or a vector of only the fifty most common words/letters as previously selected, representing a restriction of the event model. Similarly, a variety of processing classes can be [have been] written to infer a similarity between two different vectors. Authorship of the test document can be assigned to (the author of) the most similar document.

Parties interested in seeing or using this program should contact the corresponding author.

6 Discussion and Future Work

The structure of the program lends itself easily to extension and modification; for example, the result of event processing is simply a Vector (Java class) of events. Similarly, similarity judgement is a function of the Processor class, which can be instantiated in a variety of different ways. At present, the Processor class is defined with a number of different methods [for example, crossEntDistance() and LZWDistance()]. A planned improvement is to simply define a calculateDistance() function as part of the Processor class. The Processor class, in turn, can be subclassed into various types, each of which calculates distance in a slightly different way.

Similarly, preprocessing can be handled by separate instantiations and subclasses. Even data input and output can be modularized and separated. As written, the program only reads files from a local disk, but a relatively easy modification would allow files to be read from a local disk or from the network (for instance, Web pages from a site such as Project Gutenberg or *literature.org*).

From a broader perspective, this program provides a uniform framework under which competing theories of authorship attribution can both be compared and combined (to their hopefully mutual benefit). It also form the basis of a simple user-friendly tool to allow users without special training to apply technologies for authorship attribution and to take advantage of new developments and methods as they become available. From a standpoint of practical epistemology, the existence of this tool should provide a starting point for improving the quality of authorship attribution as a forensic examination – by allowing the widespread use of the technology, and at the same time providing an easy method for testing and evaluating different approaches to determine the necessary empirical valididation and limitations.

On the other hand, this tool is also clearly a "research-quality" prototype, and additional work will be needed to implement a wide variety of methods, to determine and implement additional features, to establish a sufficiently user-friendly interface. Even questions such as the preferred method of output — dendrograms? MDS subspace projections? Fixed attribution assignments as in the present system? — are in theory open to discussion and revision. It is hoped that the input of research and user groups such as the present meeting will help guide this development.

References

Burrows, J. (2003). Questions of authorships : Attribution and beyond. *Computers and the Humanities*, 37(1):5–32.

Burrows, J. F. (1989). 'an ocean where each kind. . . ' : Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23(4-5):309–21.

Easson, G. (2002). The linguistic implications of shibboleths. In *Annual Meeting of the Canadian Linguistics Association*, Toronto, Canada.

Farringdon, J. M. (1996). *Analyzing for Authorship : A Guide to the Cusum Technique*. University of Wales Press, Cardiff.

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

Holmes, D. I. (1998). The evolution of stylometry in humanities computing. *Literary and Linguistic Computing*, 13(3):111–7.

Juola, P. (1997). What can we do with small corpora? Document categorization via cross-entropy. In *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*, Edinburgh, UK. Department of Artificial Intelligence, University of Edinburgh.

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, 37(1):77–96.

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

Juola, P. (2004b). On composership attribution. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden.

Juola, P. and Baayen, H. (2003). A controlled-corpus experiment in authorship attribution by cross-entropy. In *Proceedings of* ACH/ALLC-2003, Athens, GA.

Kukushkina, O. V., Polikarpov, A. A., and Khmelev, D. V. (2000). Using literal and grammatical statistics for authorship attribution. *Problemy Peredachi Informatii*, 37(2):96–198. Translated in "Problems of Information Transmission," pp. 172–184.

Kučera, H. and Francis, W. N. (1967). *Computational Analysis of Present-day American English*. Brown University Press, Providence.

Ladefoged, P. (1993). *A Course in Phonetics*. Harcourt Brace Jovanovitch, Inc., Fort Worth, 3rd edition.

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

Nerbonne, J. (2004). The data deluge. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, Sweden. To appear in *Literary and Linguistic Computing*.

Stinson, D. R. (2002). *Cryptography: Theory and Practice*. Chapman & Hall/CRC, Boca Raton, 2nd edition.

Wellman, F. L. (1936). *The Art of Cross-Examination*. MacMillan, New York, 4th edition.