Chapter 10

AUTHORSHIP ATTRIBUTION FOR ELECTRONIC DOCUMENTS

Patrick Juola

Abstract Forensic analysis of questioned electronic documents is difficult because the nature of the documents eliminates many kinds of informative differences. Recent work in authorship attribution demonstrates the practicality of analyzing documents based on authorial style, but the state of the art is confusing. Analyses are difficult to apply, little is known about error types and rates, and no best practices are available. This paper discusses efforts to address these issues, partly through the development of a systematic testbed for multilingual, multigenre authorship attribution accuracy, and partly through the development and concurrent analysis of a uniform and portable software tool that applies multiple methods to analyze electronic documents for authorship based on authorial style.

Keywords: Authorship attribution, stylometrics, text forensics

1. Introduction

The forensic importance of questioned documents is well-understood: did Aunt Martha really write the disputed version of her will? Document examiners can look at handwriting (or typewriting) and determine authorship with near miraculous sophistication from the dot of an "i" or the cross of a "t." Electronic documents do not contain these clues. All flat-ASCII "A" characters are identical. How can one determine who made a defamatory, but anonymous, post on a blog, for example? Whether the authorship of a purely electronic document can be demonstrated to the demanding standards of a Daubert [25] hearing is an open, but important, research question.

2. Problem Statement

With the advent of modern computer technology, a substantial amount of "writing" today never involves pen, ink or paper. This paper is a good example—born as a PDF file, the first time these words see paper is in this bound volume. If my authorship of these words were challenged, I have no physical artifacts for specialists to examine.

Furthermore, the nature of electronic documents makes it substantially easier to "publish" or misappropriate them tracelessly or even to commit forgery with relative impunity. A network investigation at best only reveals the specific computer on which the document was written. It is almost impossible to figure out who was at the keyboard—who wrote it.

Chaski [6] describes three scenarios where it is both necessary to pierce the GUI and impossible to do so with traditional network investigations. In all three cases, there was no question about which computer the documents came from. Instead, the question was whether the purported authorship could be validated. The key question thus can be structured in terms of the message content. Can the authorship of an electronic document be inferred reliably from the message content?

3. Related Work

This section discusses research in authorship attribution, and the development of a test corpus for authorship attribution.

3.1 Authorship Attribution

Recent studies suggest that inferring the authorship of a document from its content is possible, but further research is necessary to meet the stringent Daubert criteria. The question of determining authorship by examining style has a long history. For example, Judges 12:5–6 describes the inference of tribal identity from the pronunciation of a specific word. Such *shibboleths* could involve specific lexical or phonological items; a person who writes of sitting on a "Chesterfield" is presumptively Canadian [7]. Wellman [27] describes how an idiosyncratic spelling of "toutch" was used in court to validate a document.

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 [20]), and it is unlikely to be found independently in two different samples. 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

judgments. The discussion of methods to do this is an active research area: 70,400 hits turned up on May 4, 2006 on a Google search for "authorship attribution." The increase from November 13, 2005 (49,500 hits) illustrates part of the continuing activity in this area in just six months.

Recent research suggests that statistical distributions of common patterns, such as the use of prepositions, may be universal enough to be relied upon, while still being informative. For this reason, scholars have focused on more sophisticated and reliable statistical tests. Specifically, Burrows [3–5] demonstrated that a statistical analysis of common words in large samples of text could group texts by author. Since then, many additional methods [1, 2, 6, 8, 10–12, 22–24] have been proposed. The current state of the art is an *ad hoc* mess of disparate methods with little cross comparison to determine which methods work and which do not. Or more accurately, because they all work at least reasonably well: under conditions discussed below, 90% accuracy is fairly typical for "good" methods. See [17] for details about which methods work the best.

Authorial analysis can even show more subtle aspects, such as the dates of documents. Figure 1 shows such an analysis [15] for a single author (Jack London), clearly dividing works written before 1912 from works that came later. The apparent division is a vertical line at about 3.14 on Dimension 1. Finding that a newly-discovered Jack London manuscript would be placed on the left-hand side of the diagram is strong evidence that it was written after 1912 as well.

3.2 Test Corpus Development

With the wide variety of techniques available, it is important but difficult to compare their power and accuracy. A fingerprint that can distinguish between Jack London and Rudyard Kipling, for example, may not work for Jane Austin and George Eliot. A proper comparison would involve standardized texts of clear provenance and known authorship on strictly controlled topics, so that the performance of each technique can be measured in a fair and accurate way. Forsyth [9] compiled the first benchmark collection of texts for validating authorship attribution techniques. Baayen [2] has developed a tighter series of texts produced under strictly controlled conditions.

To establish testing material, Baayen and co-workers at the University of Nijmegen elicited writing samples in Dutch from eight university students. The resulting 72 texts (8 subjects $\times$ 3 genres $\times$ 3 topics/genre) varied in length between 630 and 1,341 words (3,655–7,587 characters), averaging 907 words (5,235 characters) per text.

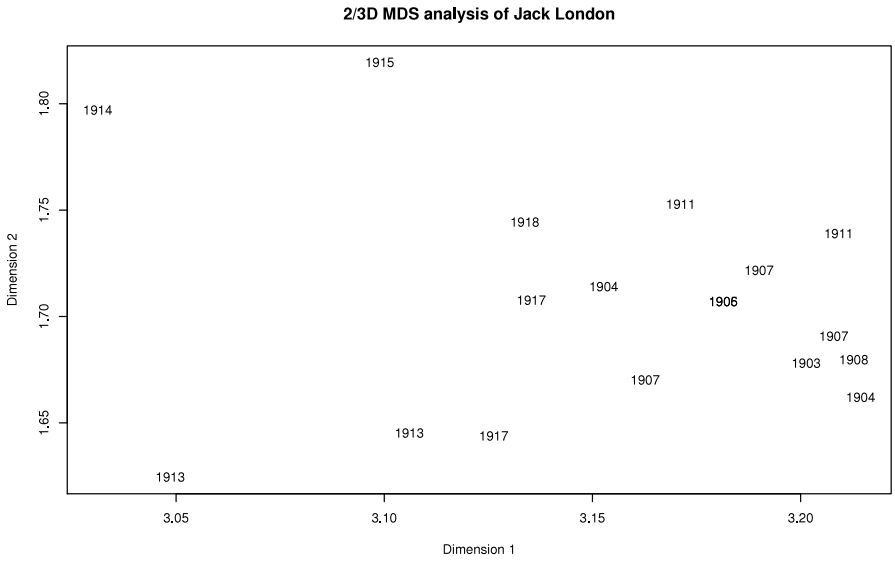

Figure 1. Spatial analysis of time development of Jack London's style.

This corpus has been comparatively analyzed using several different techniques. One of the most well-known authorship attribution techniques, proposed in [3] and later extended, is a principal components analysis (PCA) of the most common function words in a document. Another popular technique, linear discriminant analysis (LDA) [2], can distinguish among previously chosen classes, but as a supervised algorithm, it has so many degrees of freedom that the discriminants it infers may not be clinically significant. An alternative technique using measurements of cross-entropy has been independently proposed [12].

The question of which method is most accurate in this circumstance is easily answered: simply use all methods and compare the results. In particular, these methods have been tested [16] on the Baayen corpus. The software was presented with repeated trials consisting of triples containing all possible author pairs and disputed documents. Using this framework, function word PCA performed at essentially chance level, while function word LDA achieved 55% to 57% accuracy, depending upon the number of function words tabulated. Cross-entropy achieved up to 73% accuracy using a character-based model, and 87% accuracy across all pairwise comparisons using a word-based model.

From these results it can be concluded that under the circumstances of this test, cross-entropy and, in particular, word-based cross-entropy,

Table 1. Competition partipants, affiliations and methods.

Name	**Affiliation**	**Method**
Baronchelli, *et al.*	Rome	Entropy-based informatic distance
Coburn	Middlebury	Contextual network graph
van Haltern	Nijmegen	"Linguistic Profiling"
Hoover	NYU	Cluster analysis of word frequencies
Hoover	NYU	Google search for distinctive phrases
Juola	Duquesne	Match length within a database
Lana and Amisano	UNIPMN	Common N-grams (two variants)
Kešelj and Cercone	Dalhousie	CNG with weighted voting
Kešelj and Cercone	Dalhousie	CNG-wv with reject
O'Brien and Vogel	Trinity/Dublin	Chi by degrees of freedom
Rudner	GMAC	Multinomial Bayesian Model/BETSY
Koppel and Schler	Bar-Ilan	SVM with linear kernel function
Stamatatos	Patras	Meta-classifiers via feature selection

is a more accurate technique for assessing authorship. However, the chance of a false assignment is an unacceptably high 13%.

4. Ad-hoc Authorship Attribution Competition

The authorship attribution studies raise an important follow-up question about the role of the test circumstances themselves. In particular, the test data was all in Dutch, the topics were very tightly controlled, and about 8,000 words of sample data per author were available. Would the results have been substantially different if the authors had written in English? If there had been 800,000 words per author, as might be the case in a copyright dispute involving a prolific author? Can the results of an analysis involving expository essays be generalized across genres, for example, to personal letters?

To answer these questions, the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH) hosted an Ad-hoc Authorship Attribution Competition (AAAC) [13] (see Table 1). A standardized test corpus would not only allow researchers to test the ability of statistical methods to determine authorship, it would also allow "successful" methods to be distinguished from "very successful" methods. From a forensic standpoint, this would validate the science while establishing the standards of practice and creating information about error rates as Daubert requires.

Table 2. Detailed results (Problems A–G).

Team	A	B	C	D	E	F	G
baronchelli	3/13	3/13	8/9	3/4	1/4	9/10	2/4
coburn	5/13	2/13	8/9	3/4	4/4	9/10	1/4
halteren	9/13	3/13	9/9	3/4	3/4	9/10	2/4
hoover1	4/13	1/13	8/9	2/4	2/4	9/10	2/4
hoover2	4/13	2/13	9/9	4/4	4/4	10/10	2/4
juola	9/13	7/13	6/9	3/4	2/4	9/10	2/4
keselj1	11/13	7/13	8/9	3/4	2/4	9/10	3/4
keselj2	9/13	5/13	7/9	2/4	1/4	9/10	2/4
lana-amisano1	0/13	0/13	3/9	2/4	0/4	0/10	0/4
lana-amisano2	0/13	0/13	0/9	2/4	0/4	0/10	0/4
obrien	2/13	3/13	6/9	3/5	2/4	7/10	2/4
rudner	0/13	0/13	6/9	3/4	1/4	0/10	3/4
schler	7/13	4/13	9/9	4/4	4/4	10/10	2/4
stamatatos	9/13	2/13	8/9	2/4	2/4	9/10	2/4

4.1 Competition Setup

Competition materials included thirteen problems (see [13, 17] for details). These included a variety of lengths, styles, genres and languages, mostly gathered from the web but including some materials specifically gathered for the purpose. The participants (see Table 1) downloaded the anonymized materials and returned their attributions to be evaluated against the known correct answers.

4.2 Competition Results

The competition results (see Tables 2 and 3) were surprising at many levels. Some researchers initially refused to participate given the admittedly difficult tasks included among the corpora. Indeed, not all groups submitted results for all test problems. Problems for which no results were received were scored as $0/N$.

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 1,000 words) makes statistical inference difficult. Despite this apparent difficulty, almost all the groups were able to score 90% or better on this problem.

Similarly, Problem A was a realistic exercise in the analysis of student essays gathered in a first-year writing class—as is typical, no essay

Table 3. Detailed results (Problems H–M).

Team	**H**	**I**	**J**	**K**	**L**	**M**
baronchelli	3/3	2/4	1/2	2/4	4/4	5/24
coburn	2/3	2/4	1/2	2/4	3/4	19/24
halteren	2/3	3/4	1/2	2/4	2/4	21/24
hoover1	2/3	3/4	1/2	2/4	4/4	7/24
hoover2	3/3	4/4	2/2	2/4	4/4	7/24
juola	3/3	2/4	1/2	2/4	4/4	11/24
keselj1	1/3	3/4	1/2	2/4	4/4	17/24
keselj2	0/3	2/4	0/2	1/4	3/4	15/24
lana-amisano1	3/3	0/4	0/2	0/4	1/4	0/24
lana-amisano2	0/3	0/4	0/2	0/4	3/4	0/24
obrien	1/3	1/4	1/2	3/4	4/4	5/24
rudner	3/3	3/4	1/2	0/4	1/4	0/24
schler	2/3	3/4	2/2	1/4	4/4	4/24
stamatatos	1/3	3/4	1/2	2/4	3/4	14/24

exceeded 1200 words. From a standpoint of literary analysis, this may be regarded as an unreasonably short sample, but from a standpoint of a realistic test of forensic attribution and the difficult problem of testing the sensitivity of the techniques, these are legitimate.

Overall results from this competition were heartening. The highest scoring team (keselj1) had an average success rate of approximately 69%. In particular, Kešelj'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 unsolvable. As a side note, Hoover identified a weakness in the problem structure. Since much of the data was taken from the web, a search engine such as Google could be used to identify many of the documents and, therefore, the authors. Hoover himself admits that this solution neither generalizes nor addresses the technical questions of stylometry.

All the participants scored significantly above chance on the problems for which they submitted solutions. Perhaps because most research focuses on English, performance on English problems tended to be better than those in other languages. More surprisingly, the availability of large documents was not as important to accuracy as the availability of a large number of smaller documents, possibly because they are more representative samples of an author's writing. Finally, methods based on simple lexical statistics performed substantially worse than methods based on N-grams or similar measures of syntax in conjunction with lexical statistics.

With regard to generalization and confidence issues, the findings are very good for the field as a whole. In general, algorithms that were successful under one set of conditions tended to be successful under other conditions. In particular, the average performance of a method on English samples (Problems A–H) correlated significantly ($r = 0.594$, $p < 0.05$) with that method's performance on non-English samples. Correlation between large-sample problems (problems with more than 50,000 words per sample) and small sample problems was still good, although no longer strictly significant ($r = 0.3141$). This suggests that the problem of authorship attribution is at least somewhat a language- and data-independent problem, and one for which we may be able to find wide-ranging technical solutions for the general case, instead of (e.g., in machine translation) having to tailor solutions with detailed knowledge of the problem/texts/languages at hand.

In particular, we offer the following challenge to researchers who are developing new forensic analysis methods: If you cannot get 90% correct on the Paston letters (Problem F), then your algorithm is not competitively accurate. Every well-performing algorithm studied in the competition had no difficulty achieving this standard. Statements from researchers that their methods do not work on small training samples should be regarded with some suspicion.

Unfortunately, another apparent result is that the high-performing algorithms appear to be mathematically and statistically (although not necessarily linguistically) sophisticated. The good methods have names that appear fearsome to the uninitiated: linear discriminant analysis [2, 26], orthographic cross-entropy [16], common byte N-grams [18], SVM with a linear kernel function [19]. Indeed, it may be difficult to explain to explain the underlying analysis techniques to a jury.

5. Future Developments

Because authorship attribution methods can be difficult to implement (and use) we cannot expect a casual user to apply these new methods without technical assistance. At the same time, the number of techniques proposed has exploded, which also limits the pool of available users.

This issue was addressed by Juola [14], who proposed a computational framework in which the different methods could be unified, cross-compared, cross-fertilized and evaluated to achieve a well-defined "best of breed." During the past year, a proof of concept framework has been developed [17].

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

- Canonicization: No two physical realizations of events will ever be identical. Similar realizations are considered to be identical to restrict the event space to a finite set.
- Event Set Determination: The input stream is partitioned into individual non-overlapping events. At the same time, uninformative events are 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 to 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. 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 well-known distribution of English letters, 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 [20]. The question of the comparative accuracy of these methods can be judged empirically. This framework allows researchers to focus on the important differences between methods and to mix and match techniques to achieve the best results.

The usefulness of this framework is verified by our prototype user-level authorship attribution tool. Currently, this tool coordinates and combines four different technical approaches to authorship attribution [4, 5, 12, 21]. The Java program combines a GUI atop the three-phase approach defined above. Users may select a set of sample documents (with labels for known authors) and a set of testing documents by unknown authors. Users are also able to select from a menu of event selection/preprocessing options and technical inference mechanisms. Three choices are currently supported: 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 have been written to infer the similarity between

two different vectors. Authorship of the test document is assigned to the author of the most similar document.

As a specific example of application, we note that many of the AAAC methods relied on inferential statistics applied to N-grams. But N-grams of what? Juola's method was explicitly applied to N-grams of letters, van Halteren's to words or word "classes," Stamatatos' to "common words," and Koppel/Schler's to "unstable words." Therefore, we can, in theory, code Koppel's method for identifying unstable words as a separate instance of the event set class, then calculate inferential statistics using van Halteren's or Juola's method (as an instance of the inference class) possibly resulting in an improvement over any component method.

While this program is being refined, new methods are also being developed and improved. The AAAC data is still available on-line to permit people to test their methods, and we hope to incorporate new practices into our continuing study of best practices. At the same time, we will extend the functionality and user-friendliness of the system with the hope of making it more than a research prototype.

The AAAC corpus itself has some limitations that need to be addressed. For example, the mere fact that the data is on the web (in many cases, gathered from web-accessible public archives) gives an unfair advantage to any method that searches the web. Similarly, the multilingual coverage is unbalanced. The coverage of different genres is spotty and there are probably important issues that have not been addressed at all. We hope to create and offer a follow-up competition with an improved test corpus and more stringent analysis parameters.

6. Conclusions

Authorship attribution of electronic documents is an important problem in digital forensics. Recent developments in authorship attribution, including large-scale empirical experiments, are helping establish a set of best practices for analyzing questioned documents. Implicit in these experiments are an enhanced ability to create toolsets for analysis and the requirement to create new and more accurate experiments that validate the best practices.

References

[1] S. Argamon and S. Levitan, Measuring the usefulness of function words for authorship attribution, *Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2005.

[2] R. Baayen, H. van Halteren, A. Neijt and F. Tweedie, An experiment in authorship attribution, *Proceedings of JADT 2002: Sixth International Conference on Textual Data Statistical Analysis*, pp. 29-37, 2002.

[3] J. Burrows, Word-patterns and story-shapes: The statistical analysis of narrative style, *Literary and Linguistic Computing*, vol. 2, pp. 61-70, 1987.

[4] J. Burrows, "an ocean where each kind. . . :" Statistical analysis and some major determinants of literary style, *Computers and the Humanities*, vol. 23(4-5), pp. 309-321, 1989.

[5] J. Burrows, Questions of authorships: Attribution and beyond, *Computers and the Humanities*, vol. 37(1), pp. 5-32, 2003.

[6] C. Chaski, Who's at the keyboard: Authorship attribution in digital evidence invesigations, *International Journal of Digital Evidence*, vol. 4(1), 2005.

[7] G. Easson, The linguistic implications of shibboleths, presented at the *Annual Meeting of the Canadian Linguistics Association*, 2002.

[8] J. Farringdon, *Analyzing for Authorship: A Guide to the Cusum Technique*, University of Wales Press, Cardiff, United Kingdom, 1996.

[9] R. Forsyth, Towards a text benchmark suite, *Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 1997.

[10] D. Holmes, Authorship attribution, *Computers and the Humanities*, vol. 28(2), pp. 87-106, 1994.

[11] D. Hoover, Delta prime? *Literary and Linguistic Computing*, vol. 19(4), pp. 477-495, 2004.

[12] P. Juola, The time course of language change, *Computers and the Humanities*, vol. 37(1), pp. 77-96, 2003.

[13] P. Juola, Ad-hoc authorship attribution competition, *Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2004.

[14] P. Juola, On composership attribution, *Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2004.

[15] P. Juola, Becoming Jack London, to appear in *Journal of Quantitative Linguistics*.

[16] P. Juola and H. Baayen, A controlled-corpus experiment in authorship attribution by cross-entropy, *Literary and Linguistic Computing*, vol. 20, pp. 59-67, 2005.

[17] P. Juola, J. Sofko and P. Brennan, A prototype for authorship attribution studies, to appear in *Literary and Linguistic Computing*, 2006.

[18] V. Kešelj and N. Cercone, CNG method with weighted voting, presented at the *Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2004.

[19] M. Koppel and J. Schler, Ad-hoc authorship attribution competition approach outline, presented at the *Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, 2004.

[20] H. Kučera and W. Francis, *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island, 1967.

[21] O. Kukushkina, A. Polikarpov and D. Khmelev, Using literal and grammatical statistics for authorship attribution, *Problemy Peredachi Informatii*, vol. 37(2), pp. 96-198, 2000; translated in *Problems of Information Transmission*, MAIK Nauka/Interperiodica, Moscow, Russia, pp. 172-184, 2000.

[22] T. Merriam, An application of authorship attribution by intertextual distance in English, *Corpus*, vol. 2, 2003.

[23] J. Rudman, The state of authorship attribution studies: Some problems and solutions, *Computers and the Humanities*, vol. 31, pp. 351-365, 1998.

[24] E. Stamatatos, N. Fakotakis and G. Kokkinakis, Automatic authorship attribution, *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 158-164, 1999.

[25] Supreme Court of the United States, Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579, no. 92-102, 1993.

[26] H. van Halteren, R. Baayen, F. Tweedie, M. Haverkort and A. Neijt, New machine learning methods demonstrate the existence of a human stylome, *Journal of Quantitative Linguistics*, vol. 12(1), pp. 65-77, 2005.

[27] F. Wellman, *The Art of Cross-Examination*, MacMillan, New York, 1936.