A Controlled-Corpus Experiment in Authorship Identification by Cross-Entropy

Patrick Juola
Department of Mathematics and Computer Science
Duquesne University
Pittsburgh, PA 15282 USA
Tel: 412-396-5685
juola@mathcs.duq.edu

Harald Baayen
University of Nijmegen
Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands
Tel: 31-24-3521510
baayen@mpi.nl

Abstract. This paper describes an authorship, and more generally document classification, experiment on a preexisting Dutch corpus of university writings. By measuring linguistic distances using a cross-entropy technique, a technique sensitive not only to the distributions of language features, but also to their relative intersequencing, classification judgments can be made with great sensitivity, significance, confidence, and accuracy. In particular, despite the designed difficulty of the Dutch corpus used, the technique was still able to reliably detect not only authorship, but also subtle features of register, topic, and even the educational attainments of the author. We present evidence suggesting that this technique outperforms more well-known techniques such as function word principal components analysis or linear discriminant analysis, as well as suggest ways in which performance can be improved.

Keywords: authorship attribution, cross-entropy, corpus linguistics

1. Background

Authorship attribution has received significant attention in recent years as a testbed and touchstone for new theories of authorial markers and linguistic statistics; (Holmes, 1994; Holmes, 1998; Rudman, 1998; Holmes and Forsyth, 1995) present brief surveys of some major theories. The basic approach is to identify, by close inspection, a stylistic "fingerprint" characteristic of the author, and then determine whether this fingerprint is also present in a disputed work. Recent scholarship has tended to focus on specifically distributional "fingerprints," to be identified by computerized statistical analysis of the input texts. One of the current front-runners, proposed in (Burrows, 1986; Burrows, 1987; Burrows, 1992b; Burrows, 1992a), suggests that a strong cue to authorship can be gleaned from a principal components' analysis (PCA)

of the most common function words in a document. Like most statistical techniques, a scholar's ability to apply this technique is limited by the usual features of sample size, sample representativeness, and test power. Another popular technique, linear discriminant analysis (LDA), may be able to 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 (Wyner, 1996; Juola, 1997; Juola, 1998a; Juola, 2003) might be a better tool under difficult circumstances because it is capable of extracting more information (and thus distinguish more readily) along perceptually salient lines from a given data set.

With this wide variety of techniques available, it is important and yet very difficult to compare the power and accuracy of different techniques. A fingerprint appropriate to distinguish between Jack London and Rudyard Kipling, for example, may not work to distinguish between Jane Austin and George Eliot. A proper comparison would involve standardized texts of clear provenance, known authorship, on strictly controlled topics, so that the performance of each technique can be measured in a fair and accurate way. Forsyth (1997, http://www.ach.org/abstracts-/1997/p026.html) compiled a first benchmark collection of texts for validating authorship attribution techniques. (Baayen et al., 2002) have developed a more tightly controlled small series of texts produced under strictly controlled conditions. They showed that authorship can be identified even within this corpus produced by very similar authors. In this study, we reanalyze this corpus using a different technique, showing first, that such head-to-head comparisons are practical, useful, and informative, and second, that cross-entropy measures appear to outperform PCA or LDA.

2. Materials

Central to the research described is the availability and previous analysis of an extremely clean test corpus of known provenance.(Baayen et al., 2002) As discussed in (Rudman, 1998), the "integrity and validity" of the primary data is of critical importance in establishing findings of authorship with high confidence. We quote here what we hope will become known as Rudman's Law : *the closest text to the holograph should be found and used.*(Rudman, 2003) Only in such circumstances can one be [relatively] confident that the differences identified come from the author and not an editor, printer, or redactor.

The testing material used in this experiment were originally obtained in 1999 by Baayen et al. at the University of Nijmegen specifically for authorship analysis, and consist of writing samples in Dutch elicited from eight U. Nijmegen students. These students (7 women, 1 man) were all undergraduates in Dutch literature (four in their first year of study, four in their fourth year). Each student was asked to write in three different (broadly-defined) genres (fiction, argumentative writing, and descriptive writing), on nine experimentor-selected topics (fiction: retelling Little Red Riding Hood, a detective story, and a chivalric romance; argumentative writing: European unification, the health risks of smoking, and the 'Big Brother' TV show; descriptive texts: descriptions of football [soccer], the (then impending) new millennium, and a book review [of the book most recently read by the participant]). To minimize possible confounds such as practice effects, all texts were written during the same week, texts were written in random orders, and participants were not allowed to consult each other or written materials such as dictionaries and encyclopediae. Students were paid for their participation; in addition, the best text in each genre was awarded a prize of Hfl 125 (about EUR 55). The resulting 72 texts (8 subjects · 3 genres · 3 topics/genre) varied in length between 630 and 1341 words (3655–7587 characters), averaging 907 words (5235 characters) per text.

Because these materials were independently gathered and analyzed, the statistical results of our investigation are directly comparable to the previous analysis done by Baayen et al., and thus can be used to evaluate head-to-head the effectiveness of cross-entropy as an authorship attribution technique against the more traditional techniques studied in the earlier work.

3. Methods

3.1. Cross-entropy

In information theory (Shannon, 1948; Shannon, 1951), "entropy" is simply a measure of the unpredictability of a given event, given all relevant background information that could be brought to bear. "Cross-entropy" is a measure of the unpredictability of a given event, given a specific (but not necessarily best) model of events and expectations. A person completely familiar with 20th century English may still find Shakespeare somewhat daunting, an effect of three centuries of language drift, but will be more comfortable than a German speaker with no English knowledge whatsoever. This difference can be quantified and measured as a "distance" between two samples.

This general technique has several advantages over other available work such as (Burrows, 1992a; Holmes, 1998; Baayen et al., 1996). First, it seems to be widely applicable to a variety of linguistic and text-analysis problems. Second, the word "distance" is here used in its exact sense as a numerical measure that can be compared with other similarly-scaled "distances" measured from unrelated documents. Third, the method is relatively parsimonious of input text, enough so that measurements of useful precision can be made from small samples; as will be discussed, the authorship of a disputed document can be determined using less than a page of data. Fourth, this technique is sensitive to all levels and aspects of language variation.

The mathematics describing the work are not difficult, but are somewhat involved. A full explanation of the relevant algorithm can be found in (Juola, 1997; Juola, 1998a; Juola, 2003). As a quick summary, the first document is used as a sample from which one can make informed guesses about the next letter, word, grammatical construct, topic, etc. in the second document. The more closely linked the two documents are, the more accurate the guesses will be — just as someone who knows English well can predict that for any document written in English, the word "the" will be more common than the word "gryphon." This notion of accurate guessing can be accurately computed as a linguistic distance, where a low number implies accurate guessing, and therefore two documents close in all aspects. By contrast, a high number implies inaccurate guesses, implying in turn a substantial and significant difference in some way, be it language, authorship, topic, style, genre, date, or other aspect.

Algorithmic details aside, it should be noticed that there are two major "parameters" involved in the execution of this process. First, the size of the first document — or more exactly, the size of the sample drawn from the first document — will determine the amount of "information" available. More importantly, the implementation of the algorithm will control what sort of guess is made : the challenge "guess the next word" is similar in spirit but not in detail to "guess the next letter," or "guess the next part of speech." (See for yourself : "My sister gave me a" predicts an object of some sort, perhaps a concrete noun or an adjective/noun combination, but doesn't specify the exact gift.) Depending upon the problem, predicting the next word may be more or less informative than predicting (e.g.) the next letter.

3.2. Experimental framework

As observed previously, the cross-entropy can be treated mathematically as a "distance" between two documents, where a low distance

describes two similar documents. To the extent that any literary similarity exists between two documents, whether similarity in author, topic, genre, or even characteristics of the author, this should be reflected in a lower distance. Thus, we predict (and test) the following basic claim : the average within-group distance between two texts sharing a given text/author property (excluding of course the effectively zero self-distance) should be greater than the average without-group distance betweeen two texts that do not share that property.

In order to test this claim, we calculated distances between every pair of documents, treating each document as a stream of characters (i.e., every character was a separate and separately predicted "event"), using a 1024-character sample (a size approximately equivalent to this paragraph and the one immediately above) from the first document to predict characters from the second document. These distances were collected in a 72 by 72 matrix, symmetrized, and used as the basis for appropriate T-tests.

For our second and more stringent test, we perform direct pairwise comparisons between each pair of (possible) authors. Specifically, for every text in the corpus, we calculate the "distance" between all other texts by the correct author as well as eight texts by one of the seven possible distractor authors. The single excluded text is the topic-matched text by the distractor author, to prevent similarity of topic from dominating perceived authorial or stylistic similarity. This results in 504 "typical" authorship attribution tasks, where a single document of unknown provenance must be assigned to one of two authors for which a known and validated body of work exists. In this framework, the authorship of the "disputed" text can be assigned to the (known) author of the closest measured document.

Finally, to allow a more direct comparison of the assumptions that underly authorship attribution techniques, all distances were recomputed using redacted samples with attention restricted solely to a list of 164 function words (out of the 10,752 types in the total corpus). Pairwise authorship tests were performed using these new distances. This gives a more direct measurement both of the degree of importance of function words as well as the comparability of cross-entropy with other histogram-based techniques.

4. Results

A preliminary cluster analysis, as expected, correctly identified the overwhelming significance and similarity in content words and produced a collection of nine clusters, grouped by topic. Analysis of the

Table I. Within-group/without-group mean difference comparisons by text property

Property	t-value	p-value
Genre	-23.0217 (df=2458)	$p < 2.2 \cdot 10^{-16}$
Author	-5.1714 (df=1913)	$p < 2.513 \cdot 10^{-7}$
Author's Education	-2.6398 (df=5110)	$p < 0.00832$

within-group/without-group differences clearly show that, as predicted, average within-group distances were significantly smaller than average without-group distance. The numeric findings are summarized as table I.

In the direct pairwise comparisons, 73.2% (368/504) of all trials resulted in the "disputed" text being attributed to the correct author instead of the distractor author when comparisons were done using a character-based event model and distance.

When using word-based models and distances, within-group/without-group yielded comparable results, although the mean group distances for author's educations were no longer significant (t-value = -1.3226, df = 5108, $p < 0.1860$). Authorship attribution was substantially more accurate on pairwise comparisons, correctly assigning authorship in 86.9% (438/504) of the trials.

5. Discussion and Future Work

In direct comparison to the two primary methods analyzed in (Baayen et al., 2002), cross-entropy shows a substantial improvement. In their analysis, function word PCA (i.e., the Burrows technique) shows "no authorial structure," although "some structure for education level" was revealed. The linear discriminant analysis technique yielded results from 55% to 57%, depending upon the number of function words tabulated. To boost performance further, an entropy-based vector weighting scheme (inspired by (Landauer and Dumais, 1997; Landauer et al., 1998)) could increase performance by about 20 additional percentage points, into the 72–82% range. The most directly comparable cross-entropy test, using word-based events from function words, could achieve 87% accuracy. We thus conclude that using cross-entropy for this task can reduce misattributions by nearly one third.

Practical advantages aside, to what can we attribute the performance improvement, and what are the implications for authorship attribution and for humanities scholarship in general? (Juola, 1998b) has argued elsewhere that one way to evaluate the importance of a

particular aspect of language is to remove or distort it from a large sample, and to see how the solution(s) inferred from the distorted sample differ. A similar line of argument suggests that our experimental framework makes use of different (and more informative) information than traditional PCA/LDA.

There are several possible candidates for this difference. As we use 164 function words (instead of the 40–60), the difference may be due to a larger-than-expected informativeness in moderately common function words. A more likely explanation is that PCA/LDA operate only on an unordered probability distribution over a "bag of words", while the mathematics of cross-entropy takes into account ordering and inter-word sequential dependencies. Function words have long been known to inform specifically about syntactic structures [see also (Baayen et al., 1996)], even to the extent of being used as a primary learning technique by humans and computers (Morgan, 1986; Mori and Moeser, 1983; Juola, 1995); attribution techniques based on function words are at least partially focusing on the presence/absence of favored structures. It is reasonable to expect that sequence information would provide a better cue to inherently sequential syntax structures.

Framing the argument the other way, this provides evidence that a significant part of the hoped-for "authorial fingerprint" may lie in the author's choice of favored syntactic constructions, which are relatively topic-independent. However, this view contrasts somewhat with Burrows' analysis (Burrows, 2003) of the semantics of common lexical words — does his category of "temporal/modal, including auxiliary verbs and appropriate adverbs" reflect, as he suggests, an orientation "where the present is either embraced or else avoided in favor of reminiscence or desire" [p. 29], or does it reflect a personal committment to specific syntactic constructions that [might] require specific modal lexical items? Or is this orientation present and expressed through such a committment? Much additional work is needed to tease apart issues of lexical choice from syntactic choice as a marker of authorial style.

Less abstractly, the framework of the current investigation shows the importance of a set of agreed-upon and standardized test suites so that the effectiveness of a given method can be compared directly to other proposed methods, in an effort to determine both "best practices" in technology as well as to determine what areas of information are likely to be signal or noise. The score presented here of 87% can be regarded as a target, hopefully to be beaten by the next set of researchers.

Finally, there are several independent reasons to suggest that at this state of technology, Juola/Wyner cross-entropy should be regarded as the current "best practice," at least among the candidates studied. In raw performance terms, the accuracy score is comparable or better

to other metrics. Furthermore, like function word PCA (but unlike either conventional or "enhanced" LDA), cross-entropy can be used as an unsupervised, exploratory technique for investigating questions of authorship or authorial style, without the necessity for a validated "training corpus." It is extremely parsimonious of input text, allowing it to be used in situations where thousands or tens of thousands of words are unavailable, or where the disputed text itself is small. It is relatively fast to perform, and the algorithm itself is short and easy to understand.

Of course, several additional improvements are possible, if not downright likely. The fundamental underlying task performed by cross-entropy is to determine the "distance" between two documents in a high-dimensional metric space. In combination with techniques such as multi-dimensional scaling, it may be possible to determine the locations of such documents, instead of relying on blind PCA to perform such an embedding. The application of a technique such as ["enhanced"] LDA within this entropy space might be expected to outperform either technique used alone. Similarly, as other important aspects of authorship are identified, it may be possible to adapt cross-entropy to use other event models or frameworks to incorporate these aspects. One misattributed text out of seven, even on a difficult corpus, should not and probably will not be an acceptable standard.

6. Conclusions

The research presented here builds on our existing knowledge of the authorial structure of a previously collected Dutch corpus, and compares the results of cross-entropy as an authorship inference technique to the results presented in the cited paper. The original claim that "there is considerable authorial authorial structure in written texts even when the authors of these texts come from very similar background"(Baayen et al., 2002) is supported, and even strengthened, by our improved accuracy in authorship attribution. We show that cross-entropy, particularly where the events are the individual words and attention is restricted to common function words, can perform this task more accurately than even "enhanced" PCA/LDA techniques and attribute this improvement to the importance of ordering information within the sequence of function words tokens.

References

Baayen, R. H., H. Van Halteren, A. Neijt, and F. Tweedie: 2002, 'An experiment in authorship attribution'. In: *Proceedings of JADT 2002*. St. Malo, pp. 29–37.

Baayen, R. H., H. Van Halteren, and F. Tweedie: 1996, 'Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution'. *Literary and Linguistic Computing* **11**, 121–131.

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

Burrows, J. F.: 1986, 'Modal verbs and moral principles: An aspect of Jane Austen's style'. *Literary and Linguistic Computing* **1**, 9–23.

Burrows, J. F.: 1987, 'Word-patterns and story-shapes: The statistical analysis of narrative style'. *Literary and Linguistic Computing* **2**, 61–70.

Burrows, J. F.: 1992a, 'Computers and the Study of Literature'. In: C. S. Butler (ed.): *Computers and Written Texts*. Oxford: Blackwell, pp. 167–204.

Burrows, J. F.: 1992b, 'Not unless you ask nicely: The interpretative nexus between analysis and information'. *Literary and Linguistic Computing* **7**, 91–109.

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

Holmes, D. I.: 1998, 'Authorship attribution'. *Literary and Linguistic Computing* **13**(3), 111–117.

Holmes, D. I. and R. S. Forsyth: 1995, 'The Federalist revisited: new directions in authorship attribution'. *Literary & Linguistic Computing* **10**, 111–127.

Juola, P.: 1995, 'Learning to Translate : A Psycholinguistic Approach to the Induction of Grammars and Transfer Functions'. Ph.D. thesis, University of Colorado at Boulder.

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.: 1998a, 'Cross-Entropy and Linguistic Typology'. In: D. Powers (ed.): *Proceedings of New Methods in Language Processing 3*. Sydney, Australia.

Juola, P.: 1998b, 'Measuring Linguistic Complexity : The Morphological Tier'. *Journal of Quantitative Linguistics* **5**(3), 206–13.

Juola, P.: 2003, 'The Time Course of Language Change'. *Computers and the Humanities* **37**(1), 77–96.

Landauer, T. K. and S. Dumais: 1997, 'A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge.'. *Psychological Review* **104**, 211–240.

Landauer, T. K., P. W. Foltz, and D. Laham: 1998, 'Introduction to Latent Semantic Analysis'. *Discourse Processes* **25**, 259–284.

Morgan, J. L.: 1986, *From Simple Input to Complex Grammar*. Cambridge, MA: MIT Press.

Mori, K. and S. D. Moeser: 1983, 'The Role of Syntax Markers and Semantic Referents in Learning an Artificial Language'. *Journal of Verbal Learning and Verbal Behavior* **22**, 701–18.

Rudman, J.: 1998, 'The State of Authorship Attribution Studies: Some Problems and Solutions'. *Computers and the Humanities* **31**, 351–365.

Rudman, J.: 2003, 'On Determining a Valid Text for Non-Traditional Authorship Attribution Studies : Editing, Unediting, and De-Editing'. In: *Proc. 2003 Joint International Conference of the Association for Computers and the Humanities*

and the Association for Literary and Linguistic Computing (ACH/ALLC 2003). Athens, GA.

Shannon, C. E.: 1948, 'A Mathematical Theory of Communication'. *Bell System Technical Journal* **27**(4), 379–423.

Shannon, C. E.: 1951, 'Prediction and Entropy of Printed English'. *Bell System Technical Journal* **30**(1), 50–64.

Wyner, A. J.: 1996, 'Entropy Estimation and Patterns'. In: *Proceedings of the 1996 Workshop on Information Theory.*