

Author Identification with Simple Statistical Methods

Stephen St.Vincent and Taylor Hamilton

Department of Computer Science

Swarthmore College

Swarthmore, PA

{sstvinc2, thamil1}@swarthmore.edu

Abstract

We present several methods for identifying the author of an unknown text given a set of sample texts by different authors. We have two unique data sets. The first is a group of postings on an academic discussion board by 28 Swarthmore College students. The second is a set of novels written by four professional authors, obtained from the British National Corpus (BNC). Our methods tested punctuation frequencies, type-token ratios, unigram frequencies, average sentence length, and hapax legomena. We also implemented three combiners. We found that our combiners were successful in attributing authorship on both data sets. When tested on the group of professional authors, our individual methods were moderately successful. When applied to the group of students, performance of the individual methods was generally poor, with some notable exceptions.

1 Introduction

Often, especially in historical literature, the author of a work is either ambiguous or unidentified. The classic example is that of the Federalist Papers, which were written independently by Alexander Hamilton, James Madison, and John Jay. The documents were published without identifying a specific author. Later, Hamilton, Madison and Jay

claimed authorship of many of the documents, but the authorship of some were disputed or unknown. Computational linguists used word frequencies and Bayesian statistics to determine the authors of these documents (Mosteller and Wallace, 1964).

Our hypothesis is that there exist other simple methods of identifying the author of an unknown text. We use conventional author identification techniques, such as those outlined in (Chaski, 2001), but we also implement combiners to selectively aggregate the results of these techniques.

To test this hypothesis we have obtained sample data for two different groups of authors: authors of novels included in the British National Corpus (the *BNC* data set), and students in Professor Richard Wicentowski's Natural Language Processing class at Swarthmore College (the *NLP* data set). The texts in the BNC data set have very little in common other than that they are all fictional novels. By contrast, the texts in the NLP data set have many characteristics in common: the authors are all approximately the same age and have similar educational backgrounds, and the texts themselves are all written on the same topics. As a result, we expect that the author of a sample text taken from the BNC data set should be easier to identify than the author of a sample text taken from the NLP data set.

In Section 2 we present in detail the techniques we used to attribute authorship. In Section 3 we present our results. We discuss these results in Section 4.

Author	Works
Terry Pratchett	“Wings”
	“The Colour of Magic”
	“Diggers”
Iain Banks	“Complicity”
	“Walking on Glass”
	“The Wasp Factory”
Roald Dahl	“Matilda”
	“The Minpins”
Tim Vicary	“The Coldest Place On Earth”
	“Grace Darling”
	“Mary Queen of Scots”

Table 1: Authors and texts used in our first data set

2 Methods

2.1 Sample Data

Table 1 shows the authors and texts used in our BNC data set. The NLP data set had 28 authors and a total of 170 documents. Each author wrote between three and seven documents, with most authors writing five or six.

The average number of words per author for the NLP data set was 623, and for the BNC data set was 2,692. This large disparity plays an important role in the effectiveness of our tests and the analysis of our results.

Rather than choosing one document as the questioned document for our experiments, we attempted to identify the author every document using all of the other documents as training data. This gave us the ability to collect comprehensive statistics on the performance of all tests.

2.2 Experiments

Below we describe the methods that we implemented to assign authorship to the unknown documents. Most of these methods are drawn from (Chaski, 2001). For the purposes of our experiments, we conglomerated the texts of each author into a single document before performing the required tests.

2.2.1 Punctuation Frequencies

Our first hypothesis is that different authors use different punctuation marks with distinct but consistent frequencies. To test this hypothesis, we begin

by counting the frequency of each punctuation mark used. We then normalize this count by the number of words in the document, thereby making the test independent of the length of the corpus. The resulting vector of punctuation mark frequencies for each author is then compared to the frequencies observed in the test document using cosine similarity, as described in Section 2.4.

2.2.2 Type-Token Ratios

The type-token ratio (TTR) test measures the diversity of an author’s vocabulary. It is defined to be the ratio of the number of unique words in a document (types) to the total number of words in the document (tokens).

2.2.3 Hapax Legomena

Hapax legomena are words used only once in a document. Similarly to the TTR test, this test measures the proportion of words in a document that occur only once, providing some insight into the richness of an author’s vocabulary. This is the only case where each of an author’s documents was treated separately.

2.2.4 Unigram Frequencies

For each author, we counted the number of times that the author used each word. We then normalized these counts by dividing by the total number of words in the sample. This resulted in a vector of word frequencies which was compared to the questioned document using cosine similarity.

2.2.5 Bigram Frequencies

This is similar to the unigram frequencies test, with the exception that the frequency of each pair of words in a document was treated as a single entity for comparison with other authors.

2.2.6 Average Sentence Length

In this test we compare the average sentence lengths of the different authors.

2.3 Combiners

Each of the above tests provides different information about an author’s tendencies. As such circumstance will dictate which of the tests better identifies authorship. For any given unknown document it is not *a priori* known which test will produce the most

accurate results. Therefore, it should be helpful to combine all of the individual tests’ information in some intelligent way. Because these tests are largely independent of each other, a combined test should theoretically contain more information than any individual test alone.

Below we present three combiners that use different techniques for integrating the results of the experiments outlined in Section 2.2.

2.3.1 Cosine Combiner

The results of the experiments described above can be combined together to form a vector for each author whose elements can be compared to those of the questioned document using cosine similarity.

2.3.2 Voting Combiner

Each experiment assigns a rank to each author based on the order in which the author would have been selected for authorship attribution. For example, if author 8 was deemed to be the most likely author of a questioned document by a particular test, then that test would assign a rank of 1 to author 8. Similarly, if author 12 was the third most likely author, he or she would receive a rank of 3.

For the voting combiner, we sum the ranks assigned to each author by each experiment. The author with the lowest total score is selected by the combiner for authorship attribution.

2.3.3 Weighted Voting Combiner

After analyzing preliminary experimental results, we determined that some experiments are much more likely to correctly attribute authorship than others. We incorporate this observation into our weighted voting combiner.

Rather than simply summing the ranks of each experiment, we weight the ranks based on the observed accuracy of each test before adding them to the combined rank. We use the average rank given to the correct author by each test as the weighting factor that indicates the test’s accuracy. For example, one can see from Table 3 that the average rank of the correct author for the punctuation test for the NLP data set was 5.743. The value contributed by the punctuation test to the weighted combiner score is therefore the rank divided by 5.743. It should be clear that the unigram test, with an average rank of

2.83, will contribute most heavily to the weighted voting combiner.

2.4 Analysis Techniques

For all of the simple tests described in section 2.2 except for punctuation frequency, unigram frequency, and bigram frequency, we simply guess that the author of the unknown text is the one with the most similar value to that of the questioned document. For example, if the questioned document had an average sentence length of 15, we would choose the author that had the closest average sentence length to 15.

For the punctuation frequency, unigram, and bigram tests, we are able to use the more sophisticated measure of cosine similarity. The cosine of two vectors v_1 and v_2 is defined as follows:

$$\cos(\theta) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} \quad (1)$$

Intuitively, the cosine measures the similarity of two vectors: the more similar the vectors are, the closer the cosine is to unity. As such, we assign authorship to the author with the highest cosine between the author’s vector (v_1) and the questioned document’s vector (v_2).

3 Results

Test	% Correct	Avg Rank
Punctuation	27.27	5.000
TTR	45.45	2.091
Unigram	45.45	2.091
Bigram	54.55	4.455
Sen. Length	54.55	2.273
Hapax Lego.	36.36	5.000
Cosine Combiner	54.55	1.909
Voting Combiner	81.82	1.182
Weighted Combiner	72.73	1.273

Table 2: Test results for the BNC data set. Avg Rank indicates the average ranking assigned to the correct author.

Tables 2, 3, and 4 show the results for each test for the BNC, NLP, and combined data sets.

From tables 3 and 4 it is clear that the performance of the bigram test was particularly underwhelming.

Test	% Correct	Avg Rank
Punctuation	29.82	5.743
TTR	2.34	13.029
Unigram	59.06	2.830
Bigram	1.75	31.456
Sen. Length	11.70	8.550
Hapax Lego.	9.36	9.526
Cosine Combiner	9.36	9.550
Voting Combiner	42.11	3.263
Weighted Combiner	51.46	2.743

Table 3: Test results for the NLP data set

Test	% Correct	Avg Rank
Punctuation	29.67	5.698
TTR	4.95	12.368
Unigram	58.24	2.786
Bigram	4.95	29.824
Sen. Length	14.29	8.170
Hapax Lego.	10.99	9.253
Cosine Combiner	12.09	9.088
Voting Combiner	44.51	3.137
Weighted Combiner	52.75	2.654

Table 4: Combined NLP and BNC Test Results

This is because authors in the NLP data set shared almost no bigrams, leading to cosine similarities that were frequently zero. As such, we attribute very little meaning to any rankings derived from this test for the NLP data set.

4 Discussion

From the tables above it can be seen that the weighted voting combiner and the unigram test were the most successful at correctly attributing authorship. For the NLP data set, the unigram test achieved the highest accuracy, but the weighted voting combiner had the lowest average rank of the correct author. For the BNC test the simple voting combiner performed best in both categories, but the weighted combiner was the next most accurate test by a significant margin.

There were two tests that had notably different results for the two data sets. The first such test was the punctuation test. Although the test appears to perform similarly on the two data sets (29.82% correct for the NLP data set and 27.27% for the BNC data

set), the BNC data set had only four authors, so the punctuation test performed only slightly better than random chance (25%). On the other hand, for the NLP data set, selecting an author randomly would have resulted in 3.6% accuracy (28 authors). Thus, the punctuation test performed significantly better on the NLP data set.

We believe that the difference observed is due in part to the difference in average sample size for the two data sets, and in part to the unedited, colloquial nature of the NLP samples. In a small sample of writing, even one unusual punctuation mark is a significant percentage of the total punctuation used in the document. A document that is only ten sentences long may use one exclamation point and not seem unusual, but a longer document in which one out of every ten sentences contains an exclamation point would be rather jarring! Additionally, unusual punctuation is more likely to show up in unedited and casual texts, in which authors are less likely to conform to strict stylistic guidelines.

The other test which performed differently in the two data sets was the type-token ratio test. While the TTR test did fairly well on the BNC data set, it performed worse than random chance on the NLP data set. We believe that this result is also due to the small average sample size of the NLP data set. In small samples word repetition is less likely than in large samples, so type-token ratios are more similar.

There were three tests for which the average rank of the correct author in the BNC data set was greater than four. Because there were only four authors in the BNC data set, these tests consistently indicated that an author from the NLP data set was more likely to be the correct author of a questioned document than was the actual author. Despite this observation, all three of these tests performed better than chance, and therefore contributed some useful information to the combiners.

5 Conclusions

On the NLP data set, the weighted voting combiner successfully identified the correct author of the questioned document 51% of the time, with an average rank of 2.743. On the BNC data set, the regular voting combiner achieved 82% accuracy, with an average rank of 1.182. From these results, we conclude

that combining simple methods of analysis that individually provide only marginal results can produce a very effective means of author identification. Given an average document size of 623 words per author and 28 authors from which to choose, our weighted voting combiner could correctly determine authorship on more than half of its trials. We find this to be a pleasing result with clear implications for improvement through the inclusion of additional simple methods of analysis and/or combiner techniques.

References

- Carole Chaski. 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*.
- Frederick Mosteller and David Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA.