Using Markov Chains for Identification of Writers

Dmitri V. Khmelev and Fiona J. Tweedie

May 22, 2002

This paper was published in Literary and Linguistic Computing, 2001, vol.16, no.4, pp.299-307

Abstract

In this paper we present a technique for authorship attribution based on a simple Markov chain of letters (i.e., just letter bigrams are used). Many proposed methods of authorship attribution are illustrated on small examples. We show that this technique provides excellent results when applied to over 380 texts from the Project Gutenberg archives, as well as to two previously published data sets.

Contents

1 Introduction **1**

2 Method **2**

3 Application **4**

3.1 Project Gutenberg 4

3.2 Outside the Cave of Shadows 5

3.3 Federalist Papers 6

4 Discussion and Conclusions **7**

A Mathematical Background **9**

B Results from Project Gutenberg texts **9**

1 Introduction

Modern methods of authorship attribution are reviewed for Russian techniques by Milov (1994) and for Western methods by Holmes (1998). Despite the huge variety of methods, none of those described in either paper have been applied to a large number of texts. Often such methods require an element of human intervention which makes their application to large numbers of texts almost impossible. Yet the generalisability of these techniques is of prime importance — can they be used outside the particular case that they were designed for?

One method which has been tested on a number of texts was proposed by Fomenko and Fomenko (1996). They examine the proportion of function words used by an author, and find that this is stable for each author, across a large number of writers in Russian.

In this paper we present a method which has its origins in the early twentieth century in the work of Markov (1916). In criticising the work of Morozov (1915), he recalls a technique used to examine the text of *Eugene Onegin* in 1913. In Markov's paper we find the first application of the idea of the *Markov Chain*, used in many fields today, e.g. speech recognition. We consider a straightforward measure, i.e. the letters that are used in the text. Unlike recent work involving letters by, for example Kjell (1994) and Forsyth and Holmes (1996), we consider not the relative frequency of a letter bigram, but rather the probabilities of the subsequent letter, e.g. given that a particular letter is an 'f' which letters are most likely to follow it?

In the following section we describe the method in detail. Subsequent sections put the idea in to practice with a large selection of texts from the Project Gutenberg archives, data from Baayen et al. (1996)'s investigation of the use of syntactic information, and from The *Federalist Papers*. We finish with the conclusions drawn from these examples.

2 Method

If we wanted to find a probabilistic model for natural language text, we might consider a simple model where letters and spaces were generated independently according to their probabilities of occurrence in the language. Of course, letters do not occur at random, and are dependent on the letters which occur before them. The simplest such model would have each letter being dependent only on the preceding letter. This gives rise to a first-order, or simple, Markov chain model. We will show that this model can be used to determine authorship in a wide variety of examples.

As an example of a first-order Markov chain, we can consider a sequence of a reduced number of letters, for example a section of DNA. The section below is taken from the start of the X-chromosome.

```
GATCATTGATATGTTGCTAGAACTATGAGTGTTAAAGGTGCTTGTGGTGAGTTATCAGAC
AGAAACGCAGAAGATGTTATTGGAAGCTTGAGGAAAAGTGATCCTGGATTTACAGTGCCA
AGAATTGGCCTGTATTGTGTTCTCAATGTTTTTGAGGAAGGTAGAAACTGTAAGTGATGA
```

In this case we have four possible characters, A, C, G and T. If we count the number of times in this section of DNA that A occurs, we find that of the 53 occurrences, A is followed by another A 17 times, or 32.1% of the time, while a C follows only 5 times (9.4%), a G, 17 times (32.1%) and a T 14 times or 26.4%. We can then construct a full *transition matrix*:

		Second char			
		A	C	G	T
	A	17	5	17	14
First	C	7	3	1	8
char	G	20	6	8	17
	T	10	5	24	17

This can be turned into a matrix of probabilities by dividing each number by the total for that row, e.g. for A followed by another A we have $17/(17+5+17+14) = 0.320755$:

		Second char			
		A	C	G	T
	A	0.320755	0.094340	0.320755	0.264151
First	C	0.368421	0.157895	0.052632	0.421053
char	G	0.392157	0.117647	0.156863	0.333333
	T	0.178571	0.089286	0.428571	0.303571

While in studies of DNA the alphabet consists of 4 characters, with texts in English we have 26 characters, plus the space character, to deal with. The theory remains the same, with a 27×27 transition matrix replacing the 4×4 one illustrated above.

Some pre-processing of the texts is carried out before the transition matrices are calculated. All punctuation and formatting are removed. Khmelev (2000) shows that better results are obtained when capitalised words such as proper nouns and sentence-initial words are ignored and so words beginning with a capital letter are also omitted. Formatting is reduced to a single space between words, and also at the beginning and end of the text. A mathematical description of the technique can be found in Appendix A while the main text will describe it in general terms.

A transition matrix, comparable to the one above, although much larger, is then calculated for each text. The transition matrix for an author is produced by averaging the elements of the matrices from each text by that author.

In order to predict the authorship of a new text, assuming that it was written by one of the known authors, we consider the probability of the text being generated by each of the transition matrices. Each author will thus be assigned a probability. These probabilities are then ranked, with the highest probability having rank 0, the next rank 1 and so on, and the author with the highest probability and thus lowest rank is deemed to have written the text.

Khmelev (2000) presents this technique and applies it to authors writing in Russian. He shows that it gives substantially better results than the analysis of individual letters. In the present paper we apply the method to English texts including two published data sets.

3 Application

We will consider three data sets to illustrate the technique described above:

1. Authors of texts in English, obtained from the Project Gutenberg archives,[1]

2. Data from the Baayen et al. (1996) paper: 'Outside the Cave of Shadows',

[1] Project Gutenberg web site: `http://promo.net/pg/`

3. The *Federalist Papers*.

Each section will present the problem, describe the division into training and test sets, indicate the levels of cross-validation accuracy, and present and discuss the results.

3.1 Project Gutenberg

In order to consider as wide a variety of writers in English as possible, texts were obtained from the Project Gutenberg archives. A total of 387 texts were obtained from 45 authors who had more than one text included in the archive; the details of authors and classifications are given in Appendix B. One randomly chosen text from each author was held out to make up an initial test set. The results from comparing these texts with the 45 authors are given in the first two columns of Table 1. Thirty-three texts were correctly classified, while another 5 had rank 1, i.e. the correct author was the second choice. The mean of the ranked values is $\mathrm{E}(R_k) = 1.681$. Given that there are 45 texts to be assigned, the correct assignment of 73.3% of them represents an error rate of just 0.687%.[2]

One Text		All Texts	
Rank	Number	Rank	Number
0	33	0	288
1	5	1	26
2	0	2	10
3	1	3	5
4	2	4	9
> 4	5	> 4	49

Table 1: Results from cross-validation

A full cross-validation was also carried out, where each text in turn is left out of the analysis, then the authorship of this text is predicted from the other data. The third column of the table in Appendix B details the average

[2] If we assume that each pairwise comparison of a text and author is independent and error rate in each comparision is p, then the probability of a correct classification is $(1 - p)^{45}$. For error rate $p = 0.05$ we have $(1 - 0.05)^{45} \approx 0.099$, and solving for p gives $(1 - 0.006868)^{45} = 0.7333$.

rank obtained for texts from each author. Of the 387 texts and 45 possible authors, 288 texts are correctly classified. The full results are presented in the second part of Table 1. The average rank for the texts is $\mathrm{E}(R_k) = 2.100$.

3.2 Outside the Cave of Shadows

In a study of how authorial discrimination may be improved by the use of syntactic data, Baayen et al. (1996) examined ten samples of text from each of two known authors, Innes and Allingham. Of the twenty samples, the provenance of fourteen was known to the experimenters, the remaining six had to be assigned to one of the two authors. Baayen et al. use principal components analyses of frequently occurring words, measures of lexical richness, and rare constructions to identify the authors of the six text samples. While Baayen et al. apply their methods to both the syntactic and lexical vocabulary, we will just consider the lexical data since a transition matrix of syntax rules would be too large to deal with. The two sets of attributed text samples will form the training set, while the unassigned samples will form the test set.

Cross-validation gives perfect results; all of the texts known to be by author A (Innes) are classified as being by Innes, and the samples from author B (Allingham) are all classified as being by Allingham. The allocation of the six samples to be classified and their correct attributions, shown in parentheses, are A (A), B (B), A (B), A (A), B (B) and A (A) respectively. Five of the six samples have been correctly assigned to their authors, which compares favourably with similar results reported by Baayen et al. when using measures of lexical richness (four out of six classified correctly) and frequently occurring words (five out of six). Inspection of the probabilities realised by the transition matrices shows that the difference between the attributions has an average of 0.0060 and a standard deviation of 0.0028. The third text, which was mis-classified, gives rise to a difference of only 0.00038. This technique therefore performs as well as any of the methods applied to the lexical data by Baayen et al.

3.3 Federalist Papers

The *Federalist Papers* were written in 1787 and 1788 to persuade the citizens of New York State to adopt the nascent Constitution of the United States. From their initial use by Mosteller and Wallace in 1964, through research by

McColly and Weier (1983) to recent work by Holmes and Forsyth (1995) and Tweedie et al. (1996), they have become somewhat of a test case for new methods of authorship attribution.

There are 85 texts, of which 52 were written entirely by Hamilton and 14 entirely by Madison, with 12 papers that are disputed between these two authors. A further three texts were written jointly by Hamilton and Madison. The remaining texts were written by Jay and we shall not consider these texts here. The texts known to be by either Hamilton or Madison will form the training set, while the disputed and joint papers will make up the test set.

When individual texts are held out for cross-validation purposes, we find that four out of the fourteen Madison papers are classified as being by Hamilton, while only two out of the 52 Hamilton papers are misclassified. This overall misclassification rate of 9% is quite acceptable for cross-validation.

All of the disputed papers are assigned by the Markov chain to Madison, a result consistent with that of Mosteller and Wallace and subsequent researchers. In addition, the joint papers, i.e. numbers 18 to 20, are classified as being by Madison, Hamilton and Madison, respectively.

Mosteller and Wallace (1964) also assign paper 18 to Madison, although Tweedie et al.'s neural network assigns it to Hamilton. The latter note that very few of their eleven function words actually occur in the text. A technique such as the one presented here which deals with letter bigrams will not have this problem, and hopefully give rise to more accurate results.

Our technique assigns paper 19 to Hamilton, although the probabilities differ only by 0.0007, in comparison with the average difference of 0.0048 for the undisputed papers. Mosteller and Wallace conclude that the majority of the paper was written by Madison, and Tweedie et al's neural network also assigns the text to Madison.

Finally, for paper 20, Tweedie et al. (1996) cite Bourne (1901) writing

> Fully nine-tenths of it is drawn from Madison's own abstract of Sir William Temple's *Observations upon the United Provinces* and of Fetice's *Code de l'Humanité* ... Sir William Temple's claim to be recognized as joint author of No. 20 is far stronger than Hamilton's.

They conclude that Temple's influence is the reason that the paper is assigned to Hamilton; our method assigns the text, perhaps more accurately, to Madison.

4 Discussion and Conclusions

In the section above we have considered three data sets which illustrate the versatility of our proposed technique. Many studies of authorship attribution are limited by the small number of texts that are considered, with valid questions about their generalisability. To address this, our first data set was made up of 387 texts from 45 authors, 74.42% of which were correctly classified. If we treat as correct an author being in the top three selected, the success rate goes up to 83.72%, a quite remarkable result. Khmelev (2000) reports similar results with texts written in Russian.

To test the method on data in the literature we considered two previously published cases. Our technique performs at least as well as any used by Baayen et al. (1996) on the lexical data, and correctly assigns the disputed *Federalist Papers*. This sucess is particularly reassuring given the change in magnitude of the sample sizes, from hundreds of thousands of words in the Project Gutenberg archive data to around ten thousand in the Cave of Shadows data to around one thousand words in the *Federalist Papers*.

The data used for the Markov Chain can perhaps be described as linguistically microscopic - the unit is too small for meaningful conclusions to be reached regarding characteristics of the texts by the individual authors. Comparison of transition matrices may allow the researcher to comment that Hamilton uses ‘p’ followed by ‘a’ more than Madison, for example, but this does not add to the stylistic interpretation of the texts.

Such letter sequences may also be dependent on the subject of the texts. Improved results may be obtained by removing context-dependent words and performing the analysis only on function words, or very frequent words. Another possibility, along the lines of Baayen et al. (1996), would be to examine the transitions between parts of speech used, thus tapping in to the syntactic structure of the text and avoiding any dependence on context. This research is ongoing and some results are presented in Kuskushkina et al. (2001).

A Mathematical Background

Given W writers each of which has N_w texts, where $w = 0, \ldots, W-1$, we have Q_{ij}^{wn} which is the number of transitions from letter i to j, for text n ($n = 0, \ldots N_w - 1$) from writer w ($w = 0, \ldots W-1$). To find the predicted author for text $\hat{n}$ of author $\hat{w}$ we have

$$Q_{ij}^{k} = \sum_{n=0}^{N_w - 1} Q_{ij}^{kn}$$

for $k \neq \hat{w}$, and

$$Q_{ij}^{\hat{w}} = \sum_{n \neq \hat{n}} Q_{ij}^{\hat{w}n}.$$

We then have

$$\Lambda_k(\hat{w}, \hat{n}) = -\sum_{i,j} Q_{ij}^{\hat{w}\hat{n}} \ln \left(\frac{Q_{ij}^{k}}{Q_i^{k}} \right)$$

and

$$\Lambda_{\hat{w}}(\hat{w}, \hat{n}) = -\sum_{i,j} Q_{ij}^{\hat{w}\hat{n}} \ln \left(\frac{Q_{ij}^{\hat{w}}}{Q_i^{\hat{w}}} \right)$$

If $Q_{ij}^{k} = 0$ then we do not evaluate that part of the sum.

We also define ranks $R_k(\hat{w}, \hat{n})$ to be the rank of $\Lambda_k(\hat{w}, \hat{n})$ in $\{\Lambda_k(\hat{w}, \hat{n}), k = 0, \ldots, W-1\}$ where $R_k(\hat{w}, \hat{n}) \in \{0, \ldots, W-1\}$. If the text is assigned to the correct author, then $R_{\hat{w}}(\hat{w}, \hat{n}) = 0$.

B Results from Project Gutenberg texts

Author	rank of held-out text	average rank in c-v	number of texts
Austen, Jane, 1775-1817	0	0	8
Bronte, Anne, 1820-1849	0	0	2
Bronte, Charlotte, 1816-1855	1	5.25	4
Burroughs, Edgar Rice, 1875-1950	0	0	25
Carroll, Lewis, 1832-1898	0	7.67	6
Cather, Willa Sibert, 1873-1947	0	0	5
Christie, Agatha, 1891-1976	0	0	2
Conrad, Joseph, 1857-1924	0	0.32	22
Cooper, James Fenimore, 1789-1851	0	0.83	6
Crane, Stephen, 1871-1900	0	2	2
Defoe, Daniel, 1661?-1731	3	7.12	8
Dickens, Charles, 1812-1870	4	2.07	57
Doyle, Arthur Conan, Sir, 1859-1930	7	2.45	20
Eliot, T. S., 1888-1965	0	0	3
Fielding, Henry, 1707-1754	1	1	2
Fitzgerald, F. Scott, 1896-1940	0	0	2
Hardy, Thomas, 1840-1928	0	0	7
Hawthorne, Nathaniel, 1804-1864	0	0.5	12
Henry, O., 1862-1910	0	0	8
Irving, Washington, 1783-1859	0	5.43	7
James, Henry, 1843-1916	0	1.36	22
Kilmer, Joyce, 1886-1918	0	0	2
Kipling, Rudyard, 1865-1936	0	2.14	14
Lamb, Charles, 1775-1834	0	0	3
Lewis, Sinclair, 1885-1951	0	0	2
London, Jack, 1876-1916	1	1	28
Longfellow, Henry Wadsworth, 1807-1882	0	0	3
Marlowe, Christopher, 1564-1593	0	0	7
Maugham, W. Somerset, 1874-1965	0	0	2
Melville, Herman, 1819-1891	0	2	2
Millay, Edna St. Vincent, 1892-1950	0	0	2
Milton, John, 1608-1674	0	6.59	6
Poe, Edgar Allan, 1809-1849	0	0.75	8
Potter, Beatrix, 1866-1943	0	0	2
Shaw, George Bernard, 1856-1950	10	4.57	7
Shelley, Mary Wollstonecraft, 1797-1851	0	0	2
Sinclair, Upton, 1878-1968	25	29.67	3
Stoker, Bram, 1847-1912	11	7	2
Swift, Jonathan, 1667-1745	0	0.5	4
Tennyson, Alfred, Baron, 1809-1892	0	0	3
Thoreau, Henry David, 1817-1862	0	0	3
Trollope, Anthony, 1815-1882	4	6	5
Wells, H. G., 1866-1946	1	0.67	18
Wharton, Edith, 1862-1937	0	0.11	9
Wilde, Oscar, 1854-1900	1	7.25	20

Table 2: Authors sampled from Project Gutenberg, with the number of texts examined, the rank of a single held-out text and the sum of the ranks when

References

[Baayen et al. 1996] Baayen, R. H., van Halteren, H., and Tweedie, F. J. (1996). Outside the cave of shadows. Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–131.

[Bourne 1901] Bourne, E. G. (1901). The authorship of *The Federalist*. In *Essays in historical criticism*, pages 113–145. Charles Scribner's Sons, New York.

[Fomenko and Fomenko 1996] Fomenko, V. P. and Fomenko, T. G. (1996). Avtorskij invariant russkikh literaturnykh tekstov. [Predislovie A. T. Fomenko] (Authors' quantitative invariant for Russian literary texts [Commentary by Academician A. T. Fomenko]). In Fomenko, A. T., editor, *Novaja khronologija Gretsii: Antichnost' v srednevekov'e*, pages 768–820. Izd-vo MGU.

[Forsyth and Holmes 1996] Forsyth, R. S. and Holmes, D. I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, 11(4):163–174.

[Holmes 1998] Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

[Holmes and Forsyth 1995] Holmes, D. I. and Forsyth, R. S. (1995). The *Federalist* revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2):111–127.

[Khmelev 2000] Khmelev, D. V. (2000). Disputed authorship resolution through using relative entropy for Markov chains of letters in human language texts. *Journal of Quantitative Linguistics*, 7.

[Kjell 1994] Kjell, B. (1994). Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing*, 9(2):119–124.

[Kukushkina et al.] Kukushkina, O. V., Polikarpov, A. A., and Khmelev, D. V. (2001). Opredelenie avtorstva teksta ispol'zovaniem bukvennoi i grammaticheskoi informacii. *Probl. peredachi inform.*, 37(2): to appear. Translated under the title Using Letters and Grammatical Statistics for

Authorship Attribution. *Problems of Information Transmission* 37(2), *to appear.*

[Markov 1913] Markov, A. A. (1913). Primer statisticheskogo issledovanija nad tekstom 'Evgenija Onegina' illjustrirujuschij svjaz' ispytanij v tsep (An example of statistical study on the text of 'Eugene Onegin' illustrating the linking of events to a chain). *Izvestija Imp. Akademii nauk, serija VI*, 3:153–162.

[Markov 1916] Markov, A. A. (1916). Ob odnom primenenii statisticheskogo metoda (On some application of statistical method). *Izvestija Imp. Akademii nauk, serija VI*, 4:239–242.

[McColly and Weier 1983] McColly, W. B. and Weier, D. (1983). Literary attribution and likelihood ratio tests — the case of the Middle English *Pearl*-poems. *Computers and the Humanities*, 17:65–75.

[Milov 1994] Milov, L. V., editor (1994). *Ot Nestora do Fonvizina. Novye metody opredelenija avtorstva. (From Nestor to Fonvizin. New methods of determining authorship.* Progress Publishing.

[Morozov 1915] Morozov, N. A. (1915). Lingvisticheskie spektry (Linguistic spectra). *Izvestija Akademii Nauk (Section of Russian Language)*, XX(1–4).

[Mosteller and Wallace 1964] Mosteller, F. and Wallace, D. L. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* Addison-Wesley, Reading.

[Tweedie et al. 1996] Tweedie, F. J., Singh, S., and Holmes, D. I. (1996). Neural network applications in stylometry: The *Federalist Papers*. *Computers and the Humanities*, 30:1–10.