Caging the Muse:
Metrics for Unconscious Author Markers

Eugene Charniak

ec@cs.brown.edu

Paul McCann
polm@cs.brown.edu

Brown Laboratory for Linguistic Information Processing (BLLIP)

ABSTRACT

Stylometry is usually concerned with finding an authorial invariant, and attempts at authorship identification often model authors based on topics, putting weight on consciously selected content words and their unconsciously controlled frequency. This paper presents two uses of stylometry that rely on factors beyond the control of the author: document dating to a given historical period by detection of neologisms and authorial distinction based on divergence of frequency in usage of various lists of stop words.

1. Introduction

Attempts to attribute anonymous writing are nearly as old as writing itself. Biblical sources, the writings of Shakespeare, political exposés, and other documents all have continuing debates as to their genuine authorship, usually headed by experts in the field who make observations about writing style and historical facts that support their viewpoints. Stylometry attempts to make attributions more convincing by discovering an "author invariant", a value or ratio derived from the text itself that varies only slightly between works by the same author, a sort of literary fingerprint. Methods in stylometry are evaluated based on their ability to correctly classify data from a corpus of works by undisputed authors.

Stylometry has a long history. Mendenhall's "characteristic curve", defined in 1887, still represents a surprisingly useful means of distinguishing authors, and computational applications of stylometry gained significant attention following the 1964 analysis of the Federalist Papers by Mosteller and Wallace. Though conclusions one stylometer comes to are often challenged or refuted by another (Rudman 1998), in cases like that of the Federalist papers statistical analysis has been able to corroborate portions of the historical record in a manner which brings about a general agreement about truth in authorship.

One obvious goal of stylometry is to be resistant to attempted frauds who intentionally try to mimic the style of another author. In some ways this is a more severe case of distinguishing between multiple potential authors with similar styles in the face of conflicting information. One approach to avoiding this kind of confusion is to weigh more heavily facets of writing which are not the result of fully conscious decisions by the writer, often by focusing on the use of function words.

In this paper we examine two areas for stylometric analysis not under the conscious control of the writer: the first is the use of words based on their dates of origin, which we refer to as neologism detection. It seems intuitive that even if when writing a forger might avoid obviously dated words like "telephone", it is highly unusual to consider the earliest attested date of every word used. It is similarly hoped that except in the case of writers living in severest isolation, all will use a few terms invented only a decade or so before the publication of their work.

The second area of analysis is the distribution of function words or the distribution over word-lengths. In dealing with function words, besides their natural tendency to escape conscious consideration by the author, we imagine that a properly chosen set of function words will be resistant to changes in topic, being able to detect a piece of writing by an author without significant regard to genre. In this paper we test the extent to which that is true.

2. Dating

For the purposes of this paper, we examine the utility of discovering neologisms with regards to dating texts. In theory, a work should contain words only up to and including the year of its publication. In reality, both dictionaries and source texts often have issues which make analysis more complicated.

2.1. Building the Corpus

Goals for this corpus included the widest possible breadth of dates and the integrity of the source documents. While material in English is available going back before the dawn of print, it can be difficult to find many works in a condition without modernized spelling or usage or at the very least extensive modern annotation.

The texts in this case were taken from Project Gutenberg from a period spanning approximately 1650 to 1920, concentrated towards the end of this time due to the greater availability of works. Most of the works were novels, though a few early texts (such as Hobbes's Leviathan) were non-fiction.

After collection, Project Gutenberg boilerplate and any obvious annotations were removed. Some annotations were missed and noticed on later analysis, and some introductions written after the the main text but still considerably in the past were also found and removed later. As an example, Pamela has an undated Publisher's Note from after the eighteenth century. Some notes and other textual corruptions may persist, and our method attempts to deal with these if they are not of significant size.

Our corpus was divided in to twenty-five year buckets named for the first year, with no empty buckets, and containing a total of ninety-five works. The "date" of a work corresponds to its first publication; all works are prose and selected to be either an author's representative work or from roughly the middle of their career.

2.2. Gathering Neologisms

Neologisms, for our purposes, refer to words in English associated with their earliest attested appearance. By finding the words with the latest attestation dates in a document we expect to be able to guess the date of writing; for example, any book containing the word "telephone" was not written before the invention of the device.

To compile a list of words to acquire dates for, the corpus was part-of-speech tagged and lists were made of each noun, verb, adjective, or adverb, on the assumption that words outside these categories either never acquire novelty (pronouns and determiners) or may not have meaningful attestation dates (expletives or foreign words). Tree Tagger[1] was used for the tagging, and the total vocabulary was 29742 words.

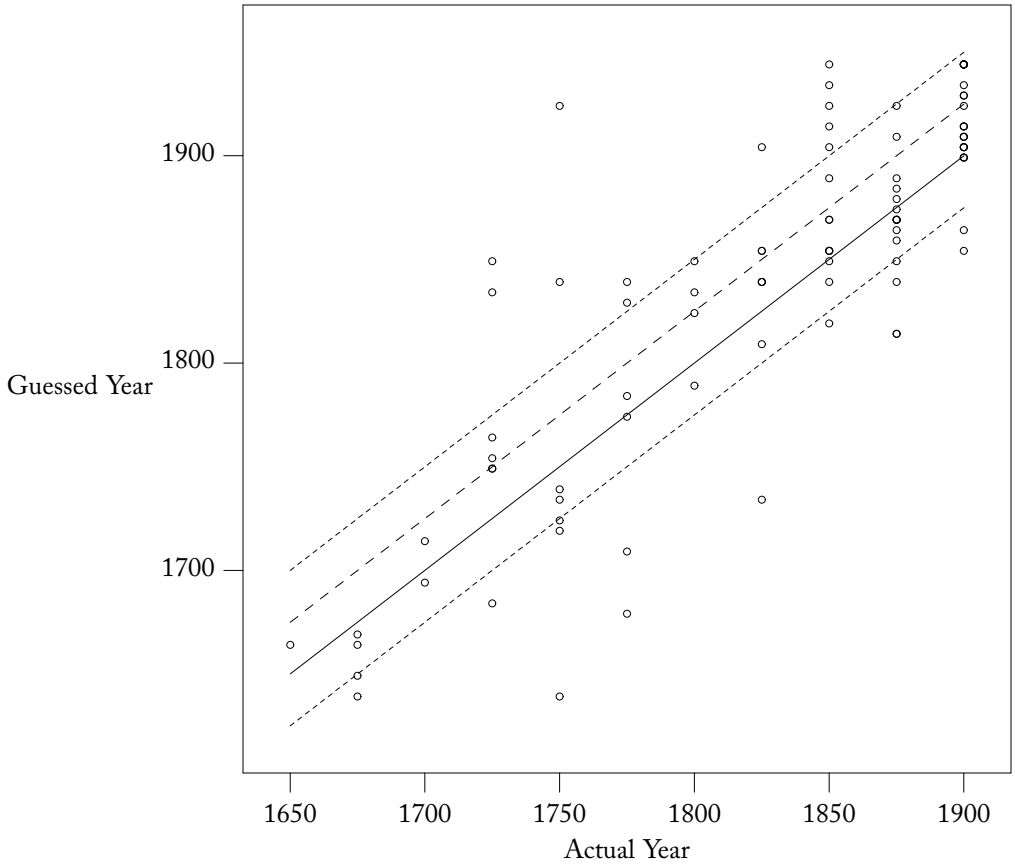
Dates were acquired by using information from the Random House Dictionary as supplied by Dictionary.com. In most cases the attestation date as presented was rounded to a year divisible by five; words with dates before 1300 were ignored.

2.3. Analysis and Tuning

A few words were manually removed from the list due to apparent errors in dating. As an example, the attestation date provided for "publicly" is 1925 despite the word's appearance in Shakespeare. The other words manually removed were "fore", "leaded", "stupendous", and "lib". All of these were discovered because their reported attestation dates were well outside the boundaries of the corpus, and were removed because their date errors either stuck out as very bad or their entries had contradictory dates.

Running tests based solely on the single latest attested word had a very high rate of error, caused mostly by either dubious word dates or corruptions of the corpus with later text. Since these problems will always be encountered in any similar experiment, this was controlled for by reporting the latest date attested that had words from two other years in the fifty years preceding. In the future, it would be ideal to

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/



On this graph the horizontal axis indicates the actual year of the work by bucket, while the vertical axis indicates the year guessed by our software. The points on this graph used a dictionary of neologisms from all four parts of speech considered. Guesses below but not on the dashed line and on or above the solid line are considered correct. Guesses between the dotted lines (again, including the bottom one but not the top one) are considered close.

model this noise as coming from a distribution of some kind, but as most errors were caused by single word outliers this measure was more than sufficient to produce better results.

2.4. Results

Using all neologisms found we correctly guessed the bucket of a given work 31.58% of the time. As we had ten bucket periods, chance for this application would have been 10%. However, if we relax our definition of correctness to

include neighboring buckets, thus taking into account guesses at most up to fifty years off, we have 71.58% accuracy. Of the not exactly correct guesses, thirty were early and thirty-five were late.

Doing analysis using only words discovered for a given part of speech resulted in results inferior to using them all together. However, using nouns or adjectives gave only slightly inferior performance while using verbs or adverbs was worse than chance.

Of the errors, many are caused by surprising texts - The Adventures of Peregrine Pickle, a particularly misclassed work (from 1751, classified 1840) contains an unusual quantity of Latin, including words like "auto", "persona", and even "laser" - in this case referring to a an extremely obscure rare herb, *laser Syriacum* - and words that have changed significantly in meaning from their use in the novel, like "medallist", used to mean someone who works with metals rather than a competition winner. The 1840 date specifically comes from the use of the word

"abominate", which is an issue with the date from the dictionary, as other sources[2] put it at 1640. Many words, such as the latin previously mentioned, postdate 1900, but their placement was sufficiently sparse that our classifier ignored them.

It should also be noted that though some works were dated as occurring after the latest work in the corpus, none were classified as coming significantly before it despite the distribution of words by year being weighted heavily towards years before 1500.

3. Stop Word Distribution

In this section we examine the effectiveness of distributions over word lengths or stop words given variations in topical content. Though it is commonly believed that due to their lack of specific meaning the use of stop words by an author would change only between dramatically different styles of writing (like poetry and prose), this does not always hold up under testing. To examine this more directly we assemble two distinct kinds of function word lists and use them in testing.

3.1. Building the Corpus

Existing corpora with divergent topics often lack a recognized set of authors; the interesting thing about the corpus is the uncertainty of its authorship despite any resemblances to well-known figures. Thus, to evaluate the topic dependency of some authorship attribution methods we will apply them to a minimally differentiated corpus: plays of William Shakespeare. Taking five each from the histories, tragedies, and comedies, and treating them each as though they were one author, we aim to see what list of words diverges minimally between each set. To evaluate the extent to which genre influences stop word distributions, we will compare each of these groupings to each other as well as to five further plays by Shakespeare, five superficially similar comedic plays by his contemporary Ben Jonson, and an anonymous 1912 translation of the Count of Monte Cristo.

The Shakespeare texts are based on the versions available on the MIT Shakespeare page [3]. The Jonson plays are taken from the Holloway Pages editions[4], and the Count of Monte Cristo is from Project Gutenberg. For all documents all non-textual elements such as stage directions, dramatis personae, speaker names and act or scene headings were removed: in other words everything except dialog. Punctuation was stripped but case was left intact.

3.2. Stop Word Selection

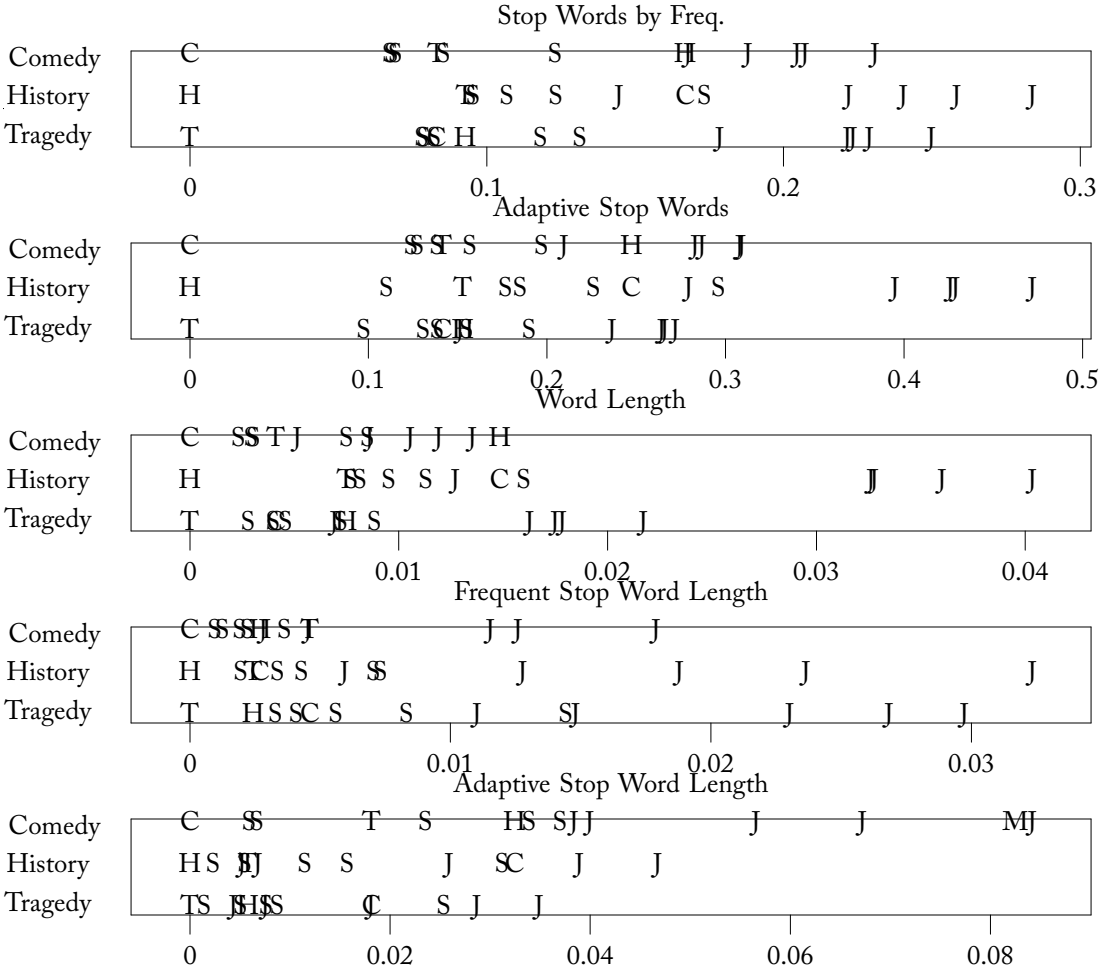
In Probability of Two Noriegas (2000), Church outlines a novel method for discovering topic words based on adaptability; that is, the tendency of a word mentioned early in a document to be mentioned again later in that document given a corpus of many documents. The reasoning as applied to newspaper articles is that if an article is about a famous person, for example Manuel Noriega, he will be mentioned throughout the article. Words of similar frequency unrelated to the topic, such as unusual non-content words like "except" or "naturally", will not occur concentrated in single articles, and thus will have low adaptability scores.

To expand on lists of stop words used in other efforts at non-topical authorship identification, a list of these non-adaptive or anti-Noriega words was compiled by splitting each of the Shakespeare plays into five equal-sized sections and treating the first hundred lines of each section as the "history" as described in Church's paper. The expectation is that words that appeared in concentrated bursts would tend to be both in the history and non-history, or test section of any given chunk. Words that appeared with similar frequency but without this kind of concentration would receive high negative adaptation scores, which we used as our criteria for selection, taking the hundred words with the highest negative adaptation scores. The table below has the first ten of these words as compared to the ten most frequent words in the corpus:

[2] http://www.etymonline.com

[4] http://shakespeare.mit.edu/

[4] http://www.hollowaypages.com/Jonson.htm



This graph presents the results for each of the five methods. Each row on a graph represents a different base distribution: Comedy, Tragedy, or History. The letters represent the KL distance of each classified work, with C, T, and H standing for the respective base distributions, S standing for other Shakespeare, and J standing for Jonson. M is for Monte Cristo, but except for with Comedies under the last model all distances for this were far in excess of others. Notice the graphs are not all to the same scale. Ideally, the five right-most characters should all be J.

First words of Stop Word Lists

Frequency	Adaptive
I	mine
the	did
and	never
to	He
of	This
you	we
a	ll
my	say
in	much
is	let

Note that while neither set of words give us any hints as to the actual contents of the documents, the adaptive words seem more interesting than the more frequent words and can suggest a certain style or focus in writing.

Specifically, the negative adaptation used to select these words is the ratio of documents where the word appears only in the "test" section of the document (in our case, after the first hundred lines) to this number plus the number of documents where it does not appear.

If normal function words are truly unrelated to topic, then it seems reasonable that

non-adaptive words should be at least as independent of topic, seeing as their necessarily widespread distribution through the corpus implies they are both less universal than common stop words and at least as unrelated to any specific topic.

For comparison we also took the one hundred most common words by raw frequency, corresponding to a more usual stop word list.

All of these words were selected with case but not including punctuation.

3.3. Testing Methods

To test for similarity between texts in our corpus we use symmetric Kullback-Liebler divergence. The KL divergence of two distributions of equal length P and Q is defined as:

$$KL(P \mid\mid Q) = \sum_i P(i)\ log(\frac{P(i)}{Q(i)})$$

The symmetric KL divergence is KL(P||Q) + KL(Q||P).

If the theory that stop words are genuinely topic-neutral is true, then regardless of the genre used as the gold-standard for building a distribution for Shakespeare we should observe that the Shakespeare plays have smaller divergences than the Jonson plays, which should have smaller divergence than The Count of Monte Cristo. The reasoning is that the obvious difference between Shakespeare and The Count of Monte Cristo should be maintained while the less obvious differences between Shakespeare and Ben Jonson, particularly in the case of the comedies, should also be made clear.

For testing we consider five distributions: the distribution of usage over the most frequent one hundred words in the training corpus, the one hundred least adaptive words in the training corpus, the general distribution of word lengths (Mendenhall's Characteristic Curve), and the distributions of word lengths for both sets of stop words. Words of length greater than fifteen letters are not considered. To avoid issues with zero counts we use add-one smoothing for all distributions.

3.4. Results

Results for this classification attempt were at once encouraging and disappointing. Despite the small size of our data set, works by Shakespeare were generally found to be closer to works by Shakespeare than the other authors; distinguishing The Count of Monte Cristo proved as easy as we would have hoped. Unfortunately, errors were still reasonably common and varied considerably by genre.

Across methods Jonson's Poetaster, a satire about poets, was always closer to the histories than Shakespeare's own As You Like It. This was in fact the only mistake made by the basic word frequency distribution method, or by the histories with any method except the distribution of adaptive stop word lengths.

Comedy proved less resilient, as the surface resemblance of Jonson's plays placed all of them closer to Shakespeare's comedies than Shakespeare's histories under Mendenhall's Characteristic Curve, though signs of similar confusion were seen with adaptive stop words and frequent stop word lengths leaving little extra distance.

In general, the adaptive words did more poorly than the standard frequency-based stop words.

4. Conclusion

Surprisingly, we were unable to find evidence of previous attempts to arbitrarily classify the year a document was written. Previous efforts at computational document dating rely on patterns of word frequency within a corpus and focus on organizing such a corpus on a small time scale, such as nine years (Dalli 2006), or for specific corpora like Biblical texts (Holmes 1994). "Document Dating" for physical documents generally refers to analysis of the physical document, including paper and ink.

Though somewhat sensitive, the general accuracy of the neologism-based classifier is encouraging, and its use with other methods merits further investigation. Future developments of the work have many angles available: documents should be pre-processed to remove foreign words and other misleading tokens; misreadings of words should be modeled as noise from a distribution; a better source for attestation dates should be acquired. We were very disappointed to discover several *dictionaries* of English etymology but not one *database*. To help further research, the list of words used in this experiment with their dates is available online. [5]

[5] http://dampfkraft.com/projects.

The stop word analysis was less rewarding. The hope that words that lack inherent content remain constant in some fashion across genres did not outweigh the basic differences in the styles of writings between genres. In this our results concur with the analysis of Mikros et al. on a Greek corpus (2007), particularly in that this does not necessarily degrade the general ability of methods to distinguish authors. The hoped advantage of words tailored to the corpus, which would find those writerly tics and make a genuine authorial fingerprint, did not come using the metric for non-adaptive words. More research is required into topic-neutral methods of authorship attribution, but for the moment it is clear a difference in genre can make a great deal of difference when attempting to classify a work even when the relationship of the genre to the feature set is inobvious.

Ultimately, our analysis here shows that even with features where it is unlikely or impossible that an author has control over their ultimate expressions their consistency is far from given and their relationship to an author's individual style is still not clear.

5. Future Work

As previously mentioned, the list of neologisms should be cleaned, expanded, and generally improved. By testing it against a known corpus words with faulty dates can be found and fixed or removed or corruptions in a work can be discovered. That said, it may be better to find a superior source of attestation dates. Additionally, the guess provided by the neologism detector, or the entire distribution of words by year it discovers, should prove a useful feature in combination with clustering methods for different eras.

With stop word lists, the basic techniques are currently being used in an experiment examining the authorship of works traditionally attributed to Shakespeare. Different word lists are being used, even some consisting of topic words; it will be seen the extent to which the opposite of supposedly topic-neutral methods is effective. Using a metric for distance similar to that used in this paper a play is attributed to the closest author.

6. References

Church, KW. (2000). Empirical Estimates of Adaptation: The Chance of Two Noriegas is closer to p/2 than p^2. *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*. Stroudsburg, PA: Association for Computational Linguistics.

Dalli, Angelo & Wilks, Yorick. (2007). Automatic Dating of Documents and Temporal Text Classification. *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, p. 17-22. Sydney: Association for Computational Linguistics.

Holmes, David I. (1994). Authorship Attribution. *Computers and the Humanities*. Netherlands: Springer.

Mendenhall, T. C. (1887). The Characteristic Curves of Composition. *Science*.

Mikros, George K. & Argiri, Elini K. (2007). Investigating Topic Influence in Authorship Attribution. *Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*. Amsterdam: ACM SIGIR.

Mosteller, F. & Wallace, D. (1964). Inference and Disputed Authorship: The Federalist. *Center for the Study of Language and Information*.

Rudman, Joseph. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities*. Netherlands: Kluwer Academic Publishers.