Journal of Law and Policy

Volume 21 | Issue 2 Article 4

2013

Authorship Attribution: What's Easy and What's Hard?

Moshe Koppel, Ph.D.

Jonathan Schler, Ph.D.

Shlomo Argamon, Ph.D

Follow this and additional works at: http://brooklynworks.brooklaw.edu/jlp

Recommended Citation

Moshe Koppel, Ph.D., Jonathan Schler, Ph.D. & Shlomo Argamon, Ph.D, *Authorship Attribution: What's Easy and What's Hard?*, 21 J. L. & Pol'y ().
Available at: http://brooklynworks.brooklaw.edu/jlp/vol21/iss2/4

This Article is brought to you for free and open access by BrooklynWorks. It has been accepted for inclusion in Journal of Law and Policy by an authorized administrator of BrooklynWorks. For more information, please contact matilda.garrido@brooklaw.edu.

AUTHORSHIP ATTRIBUTION: WHAT'S EASY AND WHAT'S HARD?

Moshe Koppel, Jonathan Schler,† and Shlomo Argamon***

INTRODUCTION

The simplest kind of authorship attribution problem—and the one that has received the most attention—is the one in which we are given a small, closed set of candidate authors and are asked to attribute an anonymous text to one of them. Usually, it is assumed that we have copious quantities of text by each candidate author and that the anonymous text is reasonably long. A number of recent survey papers[1] amply cover the variety of methods used for solving this problem.

Unfortunately, the kinds of authorship attribution problems we typically encounter in forensic contexts are more difficult than this simple version in a number of ways. First, the number of suspected writers might be very large, possibly numbering in the many thousands. Second, there is often no guarantee that the true author of an anonymous text is among the known suspects. Finally, the amount of writing we have by each candidate might be very limited and the anonymous text itself might be short.

---

* Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel, moishk@gmail.com (Corresponding Author).

† Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel, schler@gmail.com.

** Department of Computer Science, Illinois Institute of Technology, argamon@iit.edu.

[1] Patrick Juola, *Authorship Attribution*, 1 FOUND. & TRENDS IN INFO. RETRIEVAL 233, 238–39 (2006); Moshe Koppel et al., *Computational Methods in Authorship Attribution*, 60 J. AM. SOC'Y FOR INFO. SCI. & TECH. 9, 9 (2009); Efstathios Stamatatos, *A Survey of Modern Authorship Attribution Methods*, 60 J. AM. SOC'Y FOR INFO. SCI. & TECH. 538, 539 (2009).

This paper considers four versions of the attribution problem that are typically encountered in the forensic context and offers algorithmic solutions for each. Part I describes the *simple authorship attribution problem* described above. Part II considers the *long-text verification problem*, in which we are asked if two *long* texts are by the same author. Part III discusses the *many-candidates problem*, in which we are asked which among thousands of candidate authors is the author of a given text. Finally, Part IV considers the *fundamental problem* of authorship attribution, in which we are asked if two *short* texts are by the same author. Although other researchers have considered these problems, here we offer our own solutions to each problem and indicate the degree of accuracy that can be expected in each case under specified conditions.

I. SIMPLE AUTHORSHIP ATTRIBUTION

The simplest problems arise when, as mentioned above, we have a closed set of candidate authors as well as an abundance of training text[2] for each author. Our objective is to assign an anonymous text to one of the candidate authors. For this purpose, we wish to design automated techniques that use the available training text to assign a text to the most likely candidate author. As a rule, such automated techniques can be divided into two main types: *similarity-based* methods and *machine-learning* methods.[3]

In similarity-based methods, a metric is used to computationally measure the similarity between two documents, and the anonymous document is attributed to that author whose known writing (considered collectively as a single document) is most similar. Research in the similarity-based paradigm has focused on the choice of features for document representation—such as the frequency of particular words or other lexical or

---

[2] Training text is simply a collection of writing samples by a given author that can be used to characterize the author's writing style for purposes of attribution.

[3] Stamatatos, *supra* note 1, at 551.

syntactic features in the document—and on the choice of distance metric.[4]

In machine-learning methods, the known writings of each candidate author (considered as a set of distinct training documents) are used to construct a classifier that can then be used to categorize anonymous documents. The idea is to formally represent each of a set of training documents as a numerical vector and then use a learning algorithm to find a formal rule, known as a classifier, that assigns each such training vector to its known author. This same classifier can then be used to assign anonymous documents to (what one hopes is) the right author. Research in the machine-learning paradigm has focused on the choice of features for document representation and on the choice of learning algorithm.[5]

This section of the paper focuses on machine-learning methods. Here we consider and compare a variety of learning algorithms and feature sets for three authorship attribution problems that are representative of the range of classical attribution problems. The three problems are as follows:

1. A large set of emails between two correspondents (M. Koppel and J. Schler, co-authors of this paper), covering the year 2005. The set consisted of 246 emails from Koppel and 242 emails from Schler, each stripped of headers, named greetings,

---

[4] *See generally* Ahmed Abbasi & Hsinchun Chen, *Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace*, 26 ACM TRANSACTIONS ON INFO. SYS. 7:1 (2008); Shlomo Argamon, *Interpreting Burrows's Delta: Geometric and Probabilistic Foundations*, 23 LITERARY & LINGUISTIC COMPUTING 131 (2007); John Burrows, *'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship*, 17 LITERARY & LINGUISTIC COMPUTING 267 (2002); Carole E. Chaski, *Empirical Evaluations of Language-Based Author Identification Techniques*, 8 INT'L J. SPEECH LANGUAGE & L. 1 (2001); David L. Hoover, *Multivariate Analysis and the Study of Style Variation*, 18 LITERARY & LINGUISTIC COMPUTING 341 (2003).

[5] Abbasi & Chen, *supra* note 4, at 7:10; Koppel et al., *supra* note 1, at 11–12; Ying Zhao & Justin Zobel, *Effective and Scalable Authorship Attribution Using Function Words*, 3689 INFO. RETRIEVAL TECH. 174, 176 (2005); Rong Zheng et al., *A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques*, 57 J. AM. SOC'Y FOR INFO. SCI. & TECH. 378, 380 (2006).

signatures, and quotes from previous posts in the thread. Some of the texts were as short as a single word. Messages sent prior to July 1 were used as training data. The task is to classify messages sent after July 1 as having been written by either Schler or Koppel.

2. Two books by each of nine late nineteenth- and early twentieth-century authors of American and English literature (Hawthorne, Melville, Cooper, Shaw, Wilde, C. Bronte, A. Bronte, Thoreau, and Emerson). One book by each author was used for training. The task is to determine the author of each 500-word passage from the other books.

3. The full set of posts of twenty prolific bloggers, harvested in August 2004. The number of posts of the individual bloggers ranged from 217 to 745 with an average of just over 250 words per post. All but the last thirty posts of each blogger were used for training. The task is to determine the author of each of the 600 (20 authors * 30 posts) remaining blog posts.

These corpora differ along a variety of dimensions, including most prominently the size of the candidate sets (2, 9, 20) and the nature of the material (emails, novels, blogs).

For each corpus, we ran experiments comparing the effectiveness of various combinations of feature types—measurable properties of a text, such as frequencies of various words, that can be used to characterize the text—and machine-learning methods. The feature types and machine-learning methods that we used are listed in Table 1. Each document in each corpus was processed to produce a numerical vector, each of whose elements represents the relative frequency of some feature in the selected feature set. Models learned on the training sets were then applied to the corresponding test sets to estimate generalization accuracy. Table 2 shows the results for each combination of features and learning method for the email corpus. Table 3 shows the results for the literature corpus. Table 4 shows the results for the blog corpus.

As can be seen, a feature set consisting of common words and character *n*-grams (sequences of *n* characters), used in conjunction with either Bayesian logistic regression or support vector machines (SVM) as a learning algorithm, yields accuracy near or above 80% for each problem. More broadly, the results

suggest that large sets of very simple features are more accurate than small sets of sophisticated features for this purpose. Many other experiments on more straightforward problems indicate that for two-author problems and ample training text, accuracy is very close to 100%.

II. LONG-TEXT AUTHORSHIP VERIFICATION

Next, we consider the authorship verification problem for long, book-length texts. Specifically, we seek to determine whether two specific books, *A* and *X*, were written by the same author. The "unmasking" method (described below) can be used to answer this question.[6] Broadly speaking, unmasking is a technique for measuring the depth of the differences between two documents.

A naïve starting point might be to apply the methods described above to learn a model for *A* vs. *X* and assess the extent of the difference between *A* and *X* by evaluating generalization accuracy through cross-validation. (That is, we use part of the available data for training and test on the rest, repeating this process according to a specific protocol, the details of which we omit here.) This intuitive model asserts that if cross-validation accuracy is high, one should conclude that the author of *A* did not write *X*; however, if cross-validation accuracy is low (i.e., we fail to correctly classify test examples better than chance), one should conclude that the author of *A* did write *X*. This intuitive method does not actually work well at all.

Examining a real world example helps us consider exactly why the last method fails. Suppose we are given known works by Herman Melville, James Fenimore Cooper, and Nathaniel Hawthorne. For each of the three authors, we are asked if that author was or was not also the author of *The House of the Seven Gables*.[7] Using the method described and using a feature set consisting of the 250 most frequently used words in *Gables* and

---

[6] *See generally* Moshe Koppel et al., *Measuring Differentiability: Unmasking Pseudonymous Authors*, 8 J. MACH. LEARNING RES. 1261 (2007).

[7] NATHANIEL HAWTHORNE, THE HOUSE OF THE SEVEN GABLES (Project Gutenberg ed., 2008), http://www.gutenberg.org/catalog/world/readfile?fk_files=1441383.

in the known works of each of the three candidate authors, respectively, we find that we can distinguish *Gables* from the works of each author with cross-validation accuracy of above 98%. If we were to conclude, therefore, that none of these authors wrote *Gables*, we would be wrong: Hawthorne, in fact, wrote it.

If we look closely at the models that successfully distinguish *Gables* from one of Hawthorne's other works (in this case, *The Scarlet Letter*), we find that only a small number of features distinguish between them. These features include "he," which appears more frequently in *The Scarlet Letter*, and "she," which appears more frequently in *Gables*. The situation in which an author will use a small number of features in a consistently different way between works is typical. These differences might result from thematic differences between the works, differences in genre or purpose, chronological stylistic drift, or deliberate attempts by the author to mask his or her identity.

Our main point is to show how this problem can be overcome by determining not only if *A* is distinguishable from *X*, but also how great the depth of difference between *A* and *X* is.[8] To do this, we use a technique that we call "unmasking."[9] The idea is to remove, by stages, those features that are most useful for distinguishing between *A* and *X* and to gauge the speed with which cross-validation accuracy degrades as more features are removed. Our main hypothesis is that if *A* and *X* are by the same author, then whatever differences are between them will be reflected in only a relatively small number of features, despite possible differences in theme, genre, and the like. Thus, for example, we expect that when comparing *Gables* to works by other authors, the degradation as we remove distinguishing features from consideration is slow and smooth but when comparing it to another work by Hawthorne, the degradation is sudden and dramatic.

Formally, our algorithm works as follows:

1. Determine the accuracy results of a ten-fold cross-validation experiment (using SVM as a learning algorithm and

---

[8] This material is adapted from an earlier work, Koppel et al., *supra* note 6.

[9] *Id.* at 1263–64.

the 250 most common words in the corpus as a feature set) for A *against* X.

2. For the model obtained in each fold, eliminate the k *most strongly weighted positive features and the* k *most strongly weighted negative features.*

3. Go to step 1.

In this way, we construct degradation curves for the pair $<A,X>$.

In Figure 1, we show degradation curves obtained from comparing *Gables* to known works of Melville, Cooper, and Hawthorne, respectively. This graph bears out our hypothesis. Indeed, when comparing *Gables* to another work by Hawthorne, the degradation is far more severe than when comparing it to works by the other authors. Once a relatively small number of distinguishing markers are removed, the two works by Hawthorne become nearly indistinguishable.

This phenomenon is actually quite general. In fact, we have shown elsewhere[10] that we can distinguish same-author degradation curves from different-author degradation curves with accuracy above 90% in a variety of genres and languages. Unfortunately, unmasking does not work for short documents.[11] Below, we turn to the short-document problem.

III. THE MANY-CANDIDATES PROBLEM FOR SHORT DOCUMENTS

Next, we consider cases in which there may be a very large number of candidate authors, possibly in the thousands. While most work has focused on problems with a small number of candidate authors, there has been some recent work on larger candidate sets.[12]

---

[10] *Id.* at 1264–67.

[11] Conrad Sanderson & Simon Guenter, *Short Text Authorship Attribution Via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation*, PROC. INT'L CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING, 2006, at 490, *available at* http://itee.uq.edu.au/~conrad/papers.html.

[12] *See, e.g.*, Moshe Koppel et al., *Authorship Attribution with Thousands of Candidate Authors*, PROC. 29TH ANN. ACM & SIGIR CONF. ON RES. & DEV. ON INFO. RETRIEVAL, 2006, at 1–2, *available at*

We report here on a method we introduced in a previous paper.[13] The key insight is that a similarity-based approach can be used to identify the most likely authors, but the robustness of the similarity must be taken into account in order to filter false positive identifications.

We use a set of 10,000 blogs harvested in August 2004 from blogger.com.[14] The corpus is balanced for gender within each of a number of age intervals. In addition, each individual blog is predominantly in English and contains sufficient text, as will be explained. For each blog, we choose 2,000 words of known text and a *snippet*, consisting of the last 500 words of the blog, such that the posts from which the known text and the snippet are taken are disjoint. Our object is to determine which—*if any*—of the authors of the known texts is the author of a given snippet.

We begin by representing each text (both known texts and snippets) as a vector representing the respective frequencies of each *space-free character 4-gram*. For our purposes, a space-free character 4-gram is either (a) a string of characters of length four that includes no spaces or (b) a string of four or fewer characters surrounded by spaces. In our corpus, there are just over 250,000 unique (but overlapping) space-free character 4-grams. We select the 100,000 such features most frequent in

---

http://www.csie.ntu.edu.tw/~r95038/paper/paper%20WebIR/p659-koppel.pdf (demonstrating experiment with 10,000 authors); Kim Luyckx & Walter Daelemans, *Authorship Attribution and Verification with Many Authors and Limited Data*, PROC. 22ND INT'L CONF. ON COMPUTATIONAL LINGUISTICS, 2008, at 513, *available at* http://www.clips.ua.ac.be/~kim/publications.php (145 authors); David Madigan et al., *Author Identification on the Large Scale*, PROC. MEETING CLASSIFICATION SOC'Y N. AM., 2006, at 9, *available at* http://dimacs.rutgers.edu/Research/MMS/PAPERS/authorid-csna05.pdf (114 authors); Arvind Narayanan et al., *On the Feasibility of Internet-Scale Author Identification*, PROC. 33RD CONF. ON IEEE SYMP. ON SECURITY & PRIVACY, 2012, *available at* http://www.cs.berkeley.edu/~dawnsong/papers/2012%20On%20the%20Feasibility%20of%20Internet-Scale%20Author%20Identification.pdf (100,000 authors).

[13] Moshe Koppel et al., *Authorship Attribution in the Wild*, 45 LANGUAGE RESOURCES & EVALUATION 83, 86–87 (2011).

[14] This material is adapted from an earlier work, Moshe Koppel et al., *The "Fundamental Problem" of Authorship Attribution*, 93 ENG. STUD. 284, 286–88 (2012).

the corpus as our feature universe. Character *n*-grams have been shown to be effective for authorship attribution[15] and have the advantage of being measurable in any language without specialized background knowledge.

The methods we describe in Part I for authorship attribution were not designed for large numbers of classes, certainly not for 10,000 classes. Instead, we use a similarity-based method. Specifically, we use a common, straightforward information retrieval method to assign an author to a given snippet. Using cosine similarity as a proximity measure, we simply return the author whose known writing (considered as a single vector of space-free character 4-gram frequencies) is most similar to the snippet vector. Testing this rather naïve method on 1,000 snippets selected at random from among the 10,000 authors, we find that 46% of the snippets are correctly assigned. While this accuracy is perhaps surprisingly high, it is certainly inadequate for forensic applications. To remedy this problem, we adopt a previously devised approach,[16] which permits a response of "*Don't Know*" in cases where attribution is uncertain. The objective is to obtain high precision for those cases where an answer is given, while trying to offer an answer as often as possible.

The key to our new approach is the same as the underlying principle of unmasking. The known text of a snippet's actual author is likely to be the text most similar to the snippet, even as we vary the feature set that we use to represent the texts. Another author's text might happen to be the most similar for one or a few specific feature sets, but it is highly unlikely to be consistently so over many different feature sets.

This observation suggests using the following algorithm:

Given: snippet of length L_1; known-texts of length L_2 for each of C candidates

Repeat k_1 times

Randomly choose some fraction k_2 of the full feature set

Find top match using cosine similarity

---

[15] Efstathios Stamatatos et al., *Computer-Based Authorship Attribution Without Lexical Measures*, 35 COMPUTERS & HUMAN. 193, 207–08 (2001).

[16] Koppel et al., *supra* note 13; Koppel et al., *supra* note 14.

For each candidate author *A*,
Score(A) = proportion of times *A* is top match
Output: arg $\max_A$ Score(*A*) **if** max Score(*A*) > σ*; **else** *Don't Know*

The idea is to check if a given author proves to be most similar to the test snippet for many different randomly selected feature sets of fixed size. The number of different feature sets used (k_1) and the fraction of all possible features in each such set (k_2) are parameters that must be selected. The threshold σ*, which serves as the minimal score an author requires to be deemed the actual author, is a parameter that we vary for recall-precision tradeoff. We choose a high threshold if we wish to be cautious and avoid incorrect attributions, at the price of frequently returning *Don't Know*. We set the number of iterations (k_1) to 100, the snippet length (L_1) to 500, the known-text length for each candidate (L_2) to 2000, and the fraction of available features used in the feature set (k_2) to 40%. We consider how the number of candidate authors affects precision and recall. Figure 2 shows recall-precision curves for various numbers of candidate authors. Note that, as expected, accuracy increases as the number of candidate authors diminishes. The point σ* = .90 is marked on each curve. For example, for 1,000 candidates, at σ* = .90, we achieve 93.2% precision at 39.3% recall.

IV. THE "FUNDAMENTAL PROBLEM" OF AUTHORSHIP ATTRIBUTION

The above method can serve as the basis for solving what we call the "fundamental problem" of authorship attribution: determining the authorship of two (possibly short) documents written by either the same or two different authors. Plainly, if we can solve this problem, we can solve the standard attribution problems considered above, as well as many other authorship attribution problems.

Our approach[17] to solving the fundamental problem is as follows: Given two texts, *X* and *Y*, we generate a set of

[17] Koppel et al., *supra* note 14.

impostors ($Y_1,\ldots,Y_n$) and then use the above method to determine if X was written by the author of Y or any of the impostors or by none of them. If and only if we obtain a result that X was written by the author of Y with a sufficiently high score, we say that the two documents are by a single author. (Clearly, we can additionally, or alternatively, generate impostors $X_1,\ldots,X_n$ and compare them to Y.)

The crucial issues we must consider in order to adapt the above method to our problem are the following: How many impostors should be used? How should the impostors be chosen? What score should we require in order to conclude that two documents are by a single author?

We consider a test set consisting of 500 pairs of blog posts written by a single author and 500 pairs written by two different authors. Each post is truncated to exactly 500 words.

For each test pair $<X,Y>$, we proceed as follows: Choosing from a very large universe of blog posts, we identify the 250 most similar blog posts to Y (to ensure that impostors at least roughly resemble Y) and then randomly choose from among them 25 blog posts to serve as our impostors, $Y_1,\ldots,Y_n$. We assign $<X,Y>$ to a single author if and only if Y is selected from among the set $\{Y,Y_1,\ldots,Y_n\}$ as most similar to X in at least 11 trials out of 100. (The threshold 11 was determined on a separate development set.)

Using this method, 87.3% of our 1,000 test pairs are correctly identified as *same-author* or *different-author*.

V. Discussion

To summarize, four distinct problems have been considered in this paper, roughly in order of difficulty. The ordinary attribution problem with a small, closed set of candidates is well understood and solvable with established machine-learning techniques. Authorship verification, in which we wish to determine if two documents are by the same author, can be solved using unmasking provided that the documents in question are sufficiently long. The case in which there are many candidate authors can be handled using feature randomization techniques with fairly high precision, but for many cases this

method will simply respond with "*Don't Know*." Finally, authorship verification for short documents can be handled by assembling an impostor set and then invoking the method used for the many-candidates problem. This method remains somewhat speculative.

In addition to the four problems discussed above, methods have been developed by the authors of this paper for profiling authors (in terms of gender, age, native language, and personality type).[18] Moreover, it has been shown by the authors that multi-author documents can be segmented into distinct authorial threads.[19]

Although in all these cases accuracy results on out-of-sample test sets have been provided, many methodological questions that are crucial in forensic contexts are left open. Are our test corpora comparable to the kinds of cases that arise in forensic contexts? Do we make hidden assumptions about the data that are not realistic? Do our methods allow us to tell a good enough story to persuade a judge or jury of the reliability of our conclusions?

These questions are probably best answered in cooperation with legal experts and are left open for discussion.

---

[18] Shlomo Argamon et al., *Automatically Profiling the Author of an Anonymous Text*, COMM. ACM, Feb. 2009, at 119.

[19] Moshe Koppel et al., *Unsupervised Decomposition of a Document into Authorial Components*, PROC. 49TH ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS, 2011, *available at* http://www.aclweb.org/anthology-new/P/P11/P11-1136.pdf.

FW	a list of 512 function words, including conjunctions, prepositions, pronouns, modal verbs, determiners, and numbers	Stylistic
POS	38 part-of-speech unigrams and 1,000 most common bigrams using the Brill (1992) part-of-speech tagger	Stylistic
SFL	all 372 nodes in SFL trees for conjunctions, prepositions, pronouns and modal verbs	Stylistic
CW	the 1,000 words with highest information gain (Quinlan 1986) in the training corpus among the 10,000 most common words in the corpus	Content
CNG	the 1,000 character trigrams with highest information gain in the training corpus among the 10,000 most common trigrams in the corpus (cf. Keselj 2003)	Mixed content and style

NB	WEKA's implementation (Witten and Frank 2000) of Naïve Bayes (Lewis 1998) with Laplace smoothing
J4.8	WEKA's implementation of the J4.8 decision tree method (Quinlan 1986) with no pruning
RMW	our implementation of a version of Littlestone's (1988) Winnow algorithm, generalized to handle real-valued features and more than two classes (Schler 2007)
BMR	Genkin et al.'s (2006) implementation of Bayesian multi-class regression
SMO	WEKA's implementation of Platt's (1998) SMO algorithm for SVM with a linear kernel and default settings

Table 1: *Feature types and machine-learning methods used in our experiments.*

features/learner	NB	J4.8	RMW	BMR	SMO
POS	61.0%	59.0%	66.1%	66.3%	67.1%
FW+POS	65.9%	61.6%	68.0%	67.8%	71.7%
SFL	57.2%	57.2%	65.6%	67.2%	62.7%
CW	67.1%	66.9%	74.9%	78.4%	74.7%
CNG	72.3%	65.1%	73.1%	80.1%	74.9%
CW+CNG	73.2%	68.9%	74.2%	83.6%	78.2%

Table 2: *Accuracy on test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the email corpus.*

features/learner	NB	J4.8	RMW	BMR	SMO
FW	51.4%	44.0%	63.0%	73.8%	77.8%
POS	45.9%	50.3%	53.3%	69.6%	75.5%
FW+POS	56.5%	46.2%	61.7%	75.0%	79.5%
SFL	66.1%	45.7%	62.8%	76.6%	79.0%
CW	68.9%	50.3%	57.0%	80.0%	84.7%
CNG	69.1%	42.7%	49.4%	80.3%	84.2%
CW+CNG	73.9%	49.9%	57.1%	82.8%	86.3%

Table 3: *Accuracy on test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the literature. corpus.*

features/learner	NB	J4.8	RMW	BMR	SMO
FW	38.2%	30.3%	51.8%	63.2%	63.2%
POS	34.0%	30.3%	51.0%	63.2%	60.6%
FW+POS	47.0%	34.3%	62.3%	70.3%	72.0%
SFL	35.4%	36.3%	61.4%	69.2%	71.7%
CW	56.4%	51.0%	62.9%	72.5%	70.5%
CNG	65.0%	48.9%	67.1%	80.4%	80.9%
CW+CNG	69.9%	51.6%	75.4%	86.1%	85.7%

Table 4: *Accuracy test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the blog corpus.*

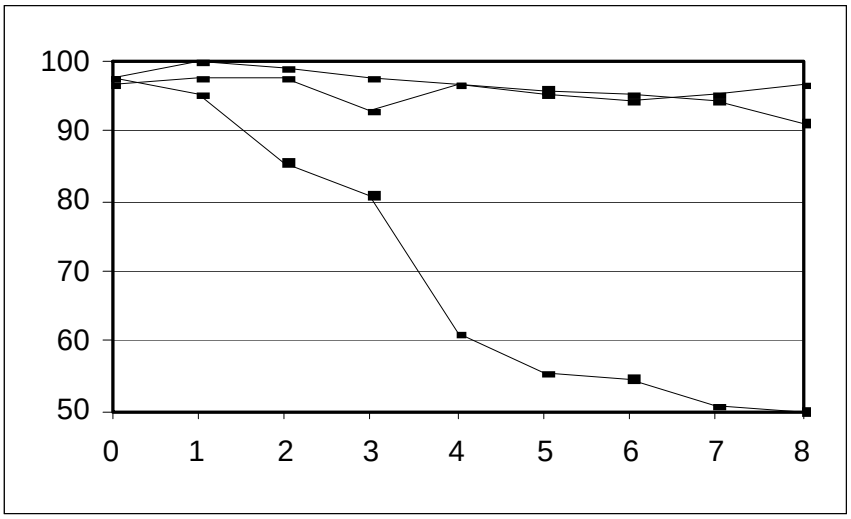

Figure 1. *Ten-fold cross-validation accuracy of models distinguishing* The House of the Seven Gables *from each of Hawthorne, Melville, and Cooper. The x-axis represents the number of iterations of eliminating best features at previous iteration. The curve well below the others is that of Hawthorne, the actual author.*

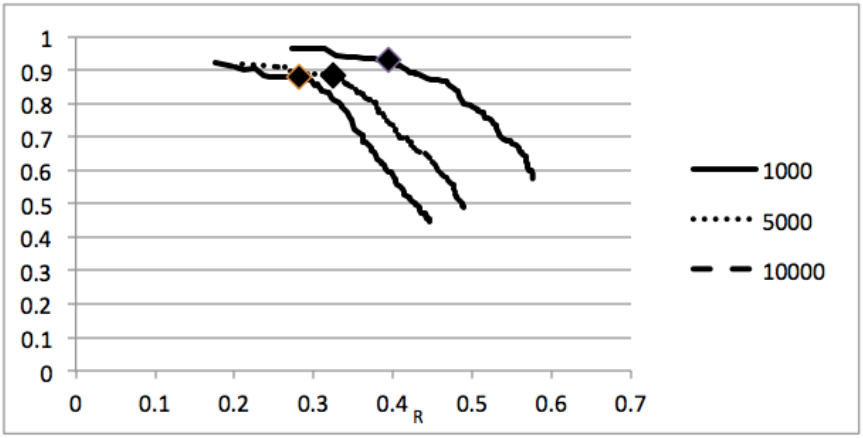

Figure 2 *Recall-precision for the many-candidates experiment (for various candidates set sizes).*