Computational Methods in Authorship Attribution

Moshe Koppel and Jonathan Schler
Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel.
E-mail: moishk@gmail.com; schler@gmail.com

Shlomo Argamon
Department of Computer Science, Illinois Institute of Technology, 10 W. 31st Street, Chicago, IL 60616.
E-mail: argamon@iit.edu

Statistical authorship attribution has a long history, culminating in the use of modern machine learning classification methods. Nevertheless, most of this work suffers from the limitation of assuming a small closed set of candidate authors and essentially unlimited training text for each. Real-life authorship attribution problems, however, typically fall short of this ideal. Thus, following detailed discussion of previous work, three scenarios are considered here for which solutions to the basic attribution problem are inadequate. In the first variant, the *profiling* problem, there is no candidate set at all; in this case, the challenge is to provide as much demographic or psychological information as possible about the author. In the second variant, the *needle-in-a-haystack* problem, there are many thousands of candidates for each of whom we might have a very limited writing sample. In the third variant, the *verification* problem, there is no closed candidate set but there is one suspect; in this case, the challenge is to determine if the suspect is or is not the author. For each variant, it is shown how machine learning methods can be adapted to handle the special challenges of that variant.

Introduction

The task of determining or verifying the authorship of an anonymous text based solely on internal evidence is a very old one, dating back at least to the medieval scholastics, for whom the reliable attribution of a given text to a known ancient authority was essential to determining the text's veracity. More recently, this problem of *authorship attribution* has gained greater prominence due to new applications in forensic analysis, humanities scholarship, and electronic commerce, and the development of computational methods for addressing the problem.

In the simplest form of the problem, we are given examples of the writing of a number of candidate authors and are asked to determine which of them authored a given anonymous text. In this straightforward form, the authorship attribution problem fits the standard modern paradigm of a text categorization problem (Lewis & Ringuette, 1994, Sebastiani, 2002). The components of text categorization systems are by now fairly well understood: Documents are represented as numerical vectors that capture statistics of potentially relevant features of the text, and machine learning methods are used to find classifiers that separate documents that belong to different classes.

However, real-life authorship attribution problems are rarely as elegant as straightforward text categorization problems, in which we have a small closed set of candidate authors and essentially unlimited training text for each. A number of varieties of attribution problems fall short of this ideal. For example, we may encounter scenarios such as:

1. There is no candidate set at all. In this case, the challenge is to provide as much demographic or psychological information as possible about the author. This is the *profiling* problem.
2. There are many thousands of candidates for each of whom we might have a very limited writing sample. This is the *needle-in-a-haystack* problem.
3. There is no closed candidate set but there is one suspect. In this case, the challenge is to determine if the suspect is or is not the author. This is the *verification* problem.

Our goal in this article is to survey the history of methods used for the basic authorship attribution scenario and to discuss some recent solutions for the more complex variants mentioned earlier.

In the following section, we offer a brief history of analytical approaches to authorship attribution, from 19th-century work on statistical authorial invariants to recent application of machine learning techniques. These modern techniques, together with recent advances in natural language processing, have enabled the development of a plethora of potential markers of authorial style (discussed later). Next, we describe the results of a systematic comparison of learning algorithms and feature sets on several representative test beds to determine the best combinations for authorship attribution.

Received April 30, 2008; revised August 8, 2008; accepted August 8, 2008

© 2008 ASIS&T • Published online 26 September 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20961

We then turn to consideration of variant scenarios where we do not have a small closed candidate set. After giving an overview of the problems and approaches, we consider the profiling problem, the needle-in-a-haystack problem, and the verification problem. Our findings are summarized in the final section.

History of Methods

Over the last century and more, a great variety of methods has been applied to authorship attribution problems of various sorts (cf. Juola, 2008). For convenience, we divide them into three classes of approach: (a) the earliest, *unitary invariant*, approach, in which a single numeric function of a text is sought to discriminate between authors; (b) the *multivariate analysis* approach, in which statistical multivariate discriminant analysis is applied to word frequencies and related numerical features; and (c) the most recent, the *machine learning* approach, in which modern machine learning methods are applied to sets of training documents to construct classifiers that can be applied to new anonymous documents.

Unitary Invariant Approach

A scientific approach to the authorship attribution problem was first proposed in the late 19th century in the work of Mendenhall (1887), who studied the authorship of texts attributed to Bacon, Marlowe, and Shakespeare, and of Mascol (1888a,1888b), who studied the authorship of the gospels of the New Testament. The key idea was that the writing of each author could be characterized by a unique curve expressing the relationship between word length and relative frequency of occurrence; these characteristic curves thus would provide a basis for author attribution of anonymous texts. This early work was put on a firmer statistical basis in the early 20th century with the search for invariant properties of textual statistics (Zipf, 1932). The existence of such invariants suggested the possibility that some related feature might be found that was at least invariant for any given author, though possibly varying among different authors. Thus, for example, Yule (1944) considered sentence length as a potential method for authorship discrimination, though he determined that this method was not reliable. A number of other measures have been proposed as authorial markers (discussed later), but for the most part, this approach has not proved stable (Burrows, 1992b; Grieve, 2007; Sichel, 1986) and has given way to multivariate methods.

Multivariate Analysis Approach

Mosteller and Wallace's work (1964) on the authorship of the *Federalist Papers* augured in a new set of methods for stylometric authorship attribution, based on combining information from multiple textual clues. Mosteller and Wallace applied a then-novel method of Bayesian classification to the papers (essentially what is now called "Naïve Bayes" classification), using as features the frequencies of a set of a few dozen function words (FWs); that is, words with primarily grammatical functions (e.g., *the*, *of*, and *about*). The fundamental insight was that a rigorous Bayesian methodology, applied to the frequencies of a set of topic-independent words, could yield a measurably reliable method for attributing authorship. This opened up the field to the exploration of new types of textual features and new modeling techniques.

A basic intuition behind these methods is that finding the most probable attribution can be viewed as taking documents as points in some space, and assigning a questioned document to the author whose documents are "closest" to it, according to an appropriate distance measure. This simple notion is quite powerful, so such distance measures have continued to be used in recent studies examining the efficacy of different metrics and feature sets. One such method is Burrows's (2002a) Delta, which has been extended and used for a variety of attribution problems (Burrows, 2002b; Hoover, 2004a, 2004b) and is equivalent to an approximate probabilistic ranking based on a multidimensional Laplacian distribution over frequently appearing words (Argamon, 2008; Stein & Argamon, 2006). A number of other similarity functions computed as distance measures for authorship attribution have been applied to different feature sets as well (Burrows, 2007; Chaski, 2001; Craig, 1999; Keselj, Peng, Cercone, & Thomas, 2003; Stamatatos, Fakotakis, & Kokkinakis, 2001; van Halteren, Baayen, Tweedie, Haverkort, & Neijt, 2005). Recently, Grieve (2007) ran an exhaustive battery of tests using this type of method.

A related class of techniques was developed earlier by Burrows (1987, 1989), who applied principal components analysis on word frequencies to analyze authorship. The idea is to visualize the differences between texts written by different authors by projecting high-dimensional word-frequency vectors computed for those text onto the two-dimensional subspace spanned by the two principal components; if good separation is seen between documents known to be written by different authors, then new texts may be attributed by seeing which authors' comparison documents are closest to them in this space. This method was elaborated on by Binongo and Smith (1999), and has been used to resolve several outstanding authorship problems (Binongo, 2003; Burrows, 1992a; Holmes, 2003). A related method is ANOVA as applied, for example, by Holmes and Forsyth (1995) to the *Federalist*. From the probabilistic standpoint, these methods take into account, to some extent, the statistical dependence of different words' frequencies.

Another form of dependence between words is taken into account by methods that model the sequencing of words in a document. This may be accounted for by using a probabilistic distance measure such as K-L divergence between Markov model probability distributions of the texts (Juola, 1998; Juola & Baayen, 2005; Khmelev, 2001; Khmelev & Tweedie, 2002; Sanderson & Guenter, 2006), possibly implicitly in the context of compression methods (Benedetto, Caglioti, & Loreto, 2002; Khmelev & Teahan, 2003; Kukushkina, Polikarpov, & Khmelev, 2001; Marton, Wu, & Hellerstein, 2005).

DOI: 10.1002/asi

Machine Learning Approach

The emergence of text categorization techniques rooted in machine learning marked an important turning point in authorship attribution studies. The application of such methods is straightforward: Training texts are represented as labeled numerical vectors, and learning methods are used to find boundaries between classes (i.e., authors) that minimize some classification loss function. The nature of the learned boundaries depends on the learning method used, but in any case, these methods facilitate the use of classes of boundaries that extend well beyond those implicit in methods that minimize distance.

Among the earliest methods to be applied were various types of neural networks, typically using small sets of FWs as features (Hoorn, Frank, Kowalczyk, & van der Ham, 1999; Kjell, 1994a; Lowe & Matthews, 1995; Matthews & Merriam, 1993; Merriam & Matthews, 1994; Tweedie, Singh, & Holmes, 1996; Waugh, Adams, & Tweedie, 2000). More recently, Graham, Hirst, and Marthi (2005) and Zheng, Li, Chen, and Huang (2006) used neural networks on a wide variety of features. Other studies have used *k*-nearest neighbor (Hoorn et al., 1999; Kjell, Woods, & Frieder, 1995; Zhao & Zobel, 2007), Naïve Bayes (Hoorn et al., 1999; Kjell, 1994a; Peng, Schuurmans, & Wang, 2004), rule learners (Abbasi & Chen, 2005; Argamon-Engelson, Koppel, & Avneri, 1998; Holmes, 1998; Holmes & Forsyth, 1995; Koppel & Schler, 2003; Zheng et al., 2006), support vector machines (Abbasi & Chen, 2005; de Vel et al., 2001; Diederich et al., 2003; Koppel & Schler, 2003; Koppel et al., 2005; Zheng et al., 2006), Winnow (Argamon, Koppel, Fine, & Shimoni, 2003; Koppel, Argamon, & Shimoni, 2002; Koppel, Akiva, & Dagan, 2006), and Bayesian regression (Argamon, Koppel, Pennebaker, & Schler, 2008; Genkin, Lewis, & Madigan, 2006; Madigan et al., 2006). Further details regarding these studies can be found in the Appendix.

Comparative studies on machine learning methods for topic-based text categorization problems (Dumais, Platt, Heckerman, & Sahami, 1998; Joachims, 1998; Yang, 1999) have shown that in general, support vector machine (SVM) learning is at least as good for text categorization as any other learning method, and the same has been found for authorship attribution (Abbasi & Chen, 2005; Zheng et al., 2006). Some recent studies (Genkin et al., 2006; Koppel, Argamon, & Shimoni, 2002) have shown that some variations of Winnow and Bayesian regression also are very promising. Next, we compare the performance of several representative learning methods for authorship attribution. As we will see, however, the choice of the learning algorithm is no more important than the choice of the features by which the texts are to be represented. We discuss this issue in the next section.

Features for Authorship Attribution

One of the advantages of modern machine learning methods is that they permit us to consider a wide variety of potentially relevant features without suffering great degradation in accuracy if most of these features prove to be irrelevant. In this section, we consider a number of feature types that have been, or might be, used for the attribution problem. A number of earlier works that have surveyed and/or compared various types of feature sets include Forsyth and Holmes (1996), Holmes (1998), McEnery and Oakes (2000), Love (2002), Zheng et al. (2006), Abbasi and Chen (2008), and Juola (2008). Note that in addition to the cited work dealing with attribution of texts in a variety of genres, there also has been a fair amount of work on attribution of programming code, music, art, and other media; such work is beyond the scope of this article.

Complexity Measures

As noted previously, early work on authorship focused on the search for a single feature that remained invariant for a given author, but varied among different authors. The search for such invariants centered on measures of text complexity. These measures included average word length (or more generally, word length distribution) in terms of syllables (Fucks, 1952) or letters (Brinegar, 1963; Mendenhall, 1887) and average number of words in sentence (Morton, 1965; Yule, 1944). When these measures proved inadequate, more sophisticated measures were invented, involving type-token ratio and the number of words appearing with given frequency in a text (e.g., *hapax legomena*). Among the better known of these are Yule's (1944) K-measure, Sichel's (1975) S-measure, and Honore's (1979) R-measure. Ultimately, none of these measures has proved especially useful on its own (Burrows, 1992b; Grieve, 2007), though it may be that these features have marginal value as additional inputs together with the features that we consider next (Abbasi & Chen, 2005, 2008; Corney, de Vel, Anderson, & Mohay; de Vel, Anderson, Corney, & Mohay, 2001; Li, Zheng, & Chen, 2006; Zheng et al., 2006).

Function Words

The search for a single invariant measure of textual style was natural in the early stages of stylometric analysis, but with the development of more sophisticated multivariate analysis techniques, larger sets of features could be considered. Among the earliest studies to use multivariate approaches was that of Mosteller and Wallace (1964), who considered distributions of FWs. The reason for using FWs in preference to others is that we do not expect their frequencies to vary greatly with the topic of the text, and hence, we may hope to recognize texts by the same author on different topics. It also is unlikely that the frequency of FW use can be consciously controlled, so one may hope that use of FWs for attribution will minimize the risk of being deceived (Chung & Pennebaker, 2007).

Many studies since that of Mosteller and Wallace (1964) have shown the efficacy of FWs for authorship attribution in different scenarios (Argamon & Levitan, 2005; Argamon-Engelson et al., 1998; Baayen, van Halteran, Neijt, & Tweedie, 2002; Binongo, 2003; Burrows, 1987; de Vel et al., 2001; Holmes, 1998; Holmes, Gordon, & Wilson, 2001;

DOI: 10.1002/asi

Holmes, Robertson, & Paez, 2001; Juola & Baayen, 2005; Karlgren & Cutting, 1994; Kessler, Nunberg, & Schütze, 1997; Koppel, Akiva, & Dagan, 2006; Koppel, Schler, & Zigdon, 2005; Merriam & Matthews, 1994; Morton, 1978; Zhao & Zobel, 2005), confirming the hypothesis that different authors tend to have different characteristic patterns of FW use.

Typical modern studies using FWs in English use lists of a few hundred words, including pronouns, prepositions, auxiliary and modal verbs, conjunctions, and determiners. Numbers and interjections are usually included as well since they are essentially independent of topic, although they are not, strictly speaking, FWs. Results of different studies using somewhat different lists of FW have been similar, indicating that the precise choice of FW is not crucial. Discriminators built from FW frequencies often perform at levels competitive with those constructed from more complex features.

Syntax and Parts-of-Speech

A different type of feature set is based on relative frequencies of different syntactic constructions, made possible by development of fast and reliable statistical natural-language-processing techniques. A number of studies have used the output of syntactic text chunkers and parsers to create features and have found that they could considerably improve results based on traditional word-based analysis alone (Baayen, van Halteren, & Tweedie, 1996; Chaski, 2005; Gamon, 2004; Hirst & Feiguina, 2007; Stamatatos, Fakotakis, & Kokkinakis, 2000, 2001; Uzuner & Katz, 2005; van Halteren, 2004). Many studies have used the frequencies of short sequences of parts-of-speech (POS), or combinations of POS and other classes of words, as a simple method for approximating syntactic features for this purpose (Argamon-Engelson et al., 1998; Chaski, 2005; de Vel et al., 2001; Koppel et al., 2002; Koppel et al., 2005; Koppel, Akiva, & Dagan, 2006; Koppel & Schler, 2003; Kukushkina et al., 2001; van Halteren et al., 2005; Zhao, Zobel, & Vines, 2006; Zheng et al., 2006).

Functional Lexical Taxonomies

FWs and some POS features can be subsumed by considering taxonomies, based on Systemic Functional Linguistics (SFL; Halliday & Matthiessen, 2003), which represent grammatical and semantic distinctions between classes of FWs at different levels of abstraction (Matthiessen, 1992). Such taxonomies are represented as trees whose roots are labeled by sets of POS (e.g., articles, auxiliary verbs, conjunctions, prepositions, pronouns). Each node's children are labeled by meaningful subclasses of the parent node (e.g., the various sorts of personal pronouns). This bottoms out at the leaves, which are labeled by sets of individual words. These taxonomies can be used to construct features for stylistic text classification, as has been done for authorship attribution on texts in English (Argamon et al., 2008; Argamon et al., 2007; Whitelaw, Herke-Couchman, & Patrick, 2004) and in Portuguese (Pavelec, Justino, & Oliveira 2007).

Such feature sets might include most FWs and some POS unigrams as well as features at intermediate levels of abstraction. Note that the features so constructed are all closed sets of words so that no POS tagging is required for identifying such features in a text.

Content Words

There are aspects of authorial identity that are not easily captured by the sorts of stylistic features described previously. For example, one author may prefer to use the words *start* and *large* whereas another may prefer *begin* and *big* (Koppel, Akiva, & Dagan, 2006; Mosteller & Wallace, 1964). Such patterns of lexical choice can be represented by modeling the relative frequencies of content words (Argamon et al., 2008; Craig, 1999; Diederich, Kindermann, Leopold, & Paass, 2003; Hoover, 2004a, 2004b; Martindale & McKenzie, 1995; Waugh et al., 2000). Typically, very rare words and those with near-uniform distribution over the corpus of interest can be omitted (Forman, 2003) so that a reasonable set of perhaps several thousand words may used. Sequences and collocations of content words also can be useful (Hoover, 2002, 2003a, 2003b).

As noted earlier, the use of content-based features for authorship studies can be problematic. Content markers might just be artifacts of a particular writing situation or experimental setup and might thus produce overly optimistic results that will not be borne out in real-life applications. Thus, if one author's training documents are all on a particular topic, the trained classifier may do very poorly at identifying documents by that author on a different topic. We are therefore careful in this article to distinguish results that exploit content-based features from those that do not.

Character N-Grams

Several authors have proposed that the frequencies of various character n-grams might be useful for capturing lexical preferences—and even grammatical and orthographic preferences—without the need for linguistic background knowledge (making application to different languages trivial). Kjell (1994a, 1994b) and Kjell et al. (1995) used relative frequencies of character n-grams for attribution of the Federalist papers, and others have used character n-grams for authorship attribution of texts in English (Clement & Sharp, 2003; Houvardas & Stamatatos, 2006; Ledger &Merriam, 1994; Stamatatos, 2008), Dutch (Hoorn, Frank, Kowalczyk, & van der Ham, 1999), Russian (Kukushkina, et al. 2001), Italian (Benedetto et al., 2002) and Greek (Keselj et al., 2003; Peng et al., 2004). Grieve (2007) found that character bigrams work surprisingly well for attribution of newspaper opinion columns. Chaski (2005, 2007) found character n-grams to work well for attribution in a forensic context. Character n-grams also have been shown useful for related stylistic classification tasks such as document similarity (Damashek, 1995) or determining the native language of the writer (Zigdon, 2005), though Graham et al.

(2005) found that character n-grams did not work as well as syntax-based features for stylistic text segmentation. Zhang and Lee (2006) found clusters of character n-grams that prove useful for a variety of text categorization problems.

The caveats regarding content words also apply to the use of character n-grams, as many will be closely associated to particular content words and roots.

Other Specialized Features

Some other features have been found useful for authorship and stylistic classification in particular cases. Morphological analysis has been shown to be useful for authorship attribution in languages with richer morphology than English, such as Greek (Stamatatos et al., 2001) and Hebrew (Koppel, Mughaz, & Akiva, 2006), where many FWs are represented by prefixes and suffixes.

For unedited texts, we might identify an author according to frequency of distinctive punctuation habits (Chaski, 2001; O'Donnell, 1966) or orthographic/syntactic errors and idiosyncrasies (de Vel et al., 2001, Koppel & Schler, 2003). Thus, Koppel and Schler (2003) analyzed e-mail texts by running them through the Microsoft Word spelling and grammar checker, automatically assigning each error found an "error type" such as *repeated letter* (e.g., *remmit* instead of *remit*), *letter substitution* (e.g., *firsd* instead of *first*), *letter inversion* (e.g., *fisrt* instead of *first*), or *conflated words* (e.g., *stucktogether*). This approximates methods used in manual analyses of authorship whose goal is to identify idiosyncratic characteristics of the author that can be recognized in a questioned text (Foster, 2000).

Finally, for documents such as e-mail, blogs, and other online content, formatting and other structural features also can be profitably exploited for authorship attribution (Abbasi & Chen, 2008; Corney et al., 2002; de Vel et al., 2001).

Summary

We have described a wide variety of feature sets and analysis methods that have been applied to various authorship attribution problems over the years. In principle, any feature set can be used with nearly any classification method, provided proper methodology is followed in study design (cf. Rudman, 1997). In practice, however, certain combinations have been more often applied and studied. As a reference, the Appendix contains a summary of methods and feature sets that have been used in different authorship studies.

Comparison Studies

In this section, we consider and compare methods and features applied to three authorship attribution problems representative of the range of classical attribution problems. Results will show that in most cases, the more informative functional lexical features perform just as well as do larger feature sets containing FWs and POS, though both are typically outperformed by content-bearing feature sets.

The corpora used in the study are as follows:

1. A large set of e-mail messages between two of the authors of this article (Koppel & Schler), covering the year 2005. The set consisted of 246 e-mail messages from Koppel and 242 e-mail messages from Schler, each stripped of headers, named greetings, signatures, and quotes from previous posts in the thread. Some of the texts were as short as a single word. The messages prior to July 1 were used for training and the second half for testing.
2. Two books by each of nine late 19th- and early 20th-century authors of American and English literature (Hawthorne, Melville, Cooper, Shaw, Wilde, C. Bronte, A. Bronte, Thoreau, Emerson). One book of each was used for training and the other for testing. Each 500-word chunk in the test books was tested separately.
3. The full set of posts of 20 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. The last 30 posts of each blogger were used as a test corpus.

As can be seen, these corpora differ along a variety of dimensions, including—most prominently—the size of the candidate sets (i.e., 2, 9, 20) and the nature of the material (i.e., e-mail messages, novels, blogs).

For each corpus, we ran experiments comparing the effectiveness of various combinations of feature types and machine learning methods. The feature types and machine learning methods that we used are given in Table 1. Note that for each feature type we consider, there are parameters that need to be chosen. It is beyond the scope of this article to determine the optimal parameter settings in each case. We show results for plausible settings that earlier work or our own preliminary tests have suggested work reasonably well. Thus, for POS, we use all unigrams, sufficiently frequent bigrams, and no trigrams (Koppel & Schler, 2003). We show results for SFL alone, but not in combination with FW and/or POS since the overlap of SFL with each of the other types is very large. We consider only character trigrams since these are long enough to capture morphology without mapping too obviously to specific words. For both content words (CW) and character n-grams (CNG), we choose the 1,000 features with the highest infogain from those that are among the 10,000 most frequent in the corpus. Note that FWs, POS, and SFL are purely stylistic feature sets while both CW and CNG encode aspects of document content in addition to style.

Each document in each corpus is 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 are then applied to the corresponding test sets to estimate generalization accuracy. Table 2 shows results for each combination of features and learning method for the e-mail corpus, Table 3 shows the results for the classic authors corpus, and Table 4 shows results for the blog corpus.

As can be seen, Naïve Bayes, which we use as a representative for multivariate methods, performs very poorly for all feature sets. Moreover, SVM and Bayesian regression are far

DOI: 10.1002/asi

TABLE 1. Feature types and machine learning methods used in our experiments.

FW	A list of 512 function words, including conjunctions, prepositions, pronouns, modal verbs, determiners, and numbers (purely stylistic)
POS	Thirty-eight part-of-speech unigrams and 1,000 most common bigrams using the Brill (1992) part-of-speech tagger (purely stylistic)
SFL	All 372 nodes in SFL trees for conjunctions, prepositions, pronouns, and modal verbs (purely 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
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)
NB	WEKA's implementation (Witten & 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 multiclass regression
SMO	Weka's implementation of Platt's (1998) SMO algorithm for SVM with a linear kernel and default settings

TABLE 2. Accuracy on test set attribution for a variety of feature sets and learning algorithms applied to authorship classification for the e-mail corpus.

Features/learner	NB (%)	J4.8 (%)	RMW (%)	BMR (%)	SMO (%)
FW	60.2	58.7	66.1	68.2	63.8
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

Note. See features/learner descriptions in Table 1.

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	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

Note. See features/learner descriptions in Table 1.

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

Note. See features/learner descriptions in Table 1.

superior to the other learning algorithms, for all feature sets. Moreover, for these learners, SFL features are approximately as good as FWs and POS together.

In the corpora we consider here, content words prove to be very useful; in no case do they lead us astray. Especially surprising is the effectiveness of the character n-gram feature set. Note that character n-grams perform almost identically to content words for the first two corpora and significantly outperform content words for the blog corpus. Consideration of some examples of useful character n-grams suggests that character n-grams serve as proxies for content words (e.g., *dsh* for *spreadsheet*) as well as for FWs, POS, and even formatting (e.g., the string *colon–newline–1* suggests a numbered list). In the case of the blog corpus, character n-grams have the additional benefit of capturing acronyms and abbreviations characteristic of blog writing.

We tentatively conclude, therefore, that when the context indicates that purely stylistic features are appropriate, the combination of POS and FWs constitutes a reasonable choice of feature set and that SFL features can be used as an efficient proxy for this combination. In cases where using content features is appropriate, properly chosen unigrams are a good choice, with similarly chosen character trigrams an efficient and language-independent proxy. Using these features, as appropriate, in conjunction with either Bayesian regression or SVM, constitutes a convenient and effective method for ordinary authorship attribution.

Variations on the Basic Attribution Problem

The attribution problem we have considered thus far is the standard one in which we are given a relatively small closed set of candidate authors and are asked to determine which of them is the author of a given document. In the sections that follow, we consider three variations in which no small closed candidate set is available.

First, we consider the case in which no candidate set is available at all so that the best we can hope to do is to profile the anonymous author. We will see later that essentially the same methods that we used earlier for distinguishing individual authors can be used to distinguish between classes of authors, such as males and females or writers of different ages. The discussion is drawn from that in Argamon et al. (in press).

Next, we consider the case in which the candidate set consists of many thousands of authors so that learning a classifier to distinguish them is infeasible. We will see later that this problem can be solved if we are willing to accept *Don't Know* as an answer for those cases where the document to be attributed is not sufficiently distinct to permit attribution. We use meta-learning to identify such cases and find that in the remaining cases, where the system believes attribution is reliable, we are able to provide highly accurate results. The discussion is an expansion of that given in Koppel, Schler, Argamon, and Messeri (2006).

Finally, we consider the case where there is a single candidate author, and our task is to determine if the anonymous document was written by that author. Subsequently, we show that this problem is solvable if the anonymous text is sufficiently long. The method used entails measuring the "depth" of the differences between the known texts of the candidate author and the anonymous text. In particular, we check how accurately we can distinguish between the two because the best features for doing so are iteratively eliminated. The discussion in is drawn from Koppel, Schler, and Bonchek-Dokow (2007).

Profiling

As noted previously, even in cases where we have an anonymous text and no candidate authors, we would like to say something about the anonymous author. That is, we wish to exploit the sociolinguistic observation that different groups of people speaking or writing in a particular genre and in a particular language use that language differently (cf. Chambers, Trudgill, & Schilling-Estes, 2004). More specifically, we wish to use the features and methods employed earlier to distinguish between individual authors to distinguish between classes of authors.

As in Argamon et al. (in press), we consider the following profile dimensions: author *gender* (Argamon, Koppel, Fine, & Shimoni, 2003; Koppel et al., 2002), *age* (Burger & Henderson, 2006; Schler, Koppel, Argamon, & Pennebaker, 2006), *native language* (Koppel et al., 2005), and *neuroticism level* (Pennebaker & King, 1999; Pennebaker, Mehl, & Niederhoffer, 2003). For each of these, we assemble an appropriately labeled corpus and proceed exactly as described previously. Thus, for example, we learn a classifier to distinguish between male and female writers using the same procedure we used earlier to distinguish between individual authors. Other authors have considered dimensions we do not consider here, such as education level (Corney et al., 2002).

Following our earlier observations, here we use SFL as our stylistic feature set. For comparison, we also consider content features alone and stylistic features and content features together. The content features used are the CW feature set described previously. We use Bayesian regression as our learning algorithm. For each of the three feature sets, we run 10-fold cross-validation tests to test the extent to which each profiling problem is solvable. We also present the most discriminating features for each category within each of the four problems.

TABLE 5. Classification accuracy for profiling problems using different feature sets.

	Baseline	Style	Content	Style + Content
Gender (2 classes)	*50.0*	72.0	75.1	**76.1**
Age (3 classes)	*42.7*	66.9	75.5	**77.7**
Language (5 classes)	*20.0*	65.1	**82.3**	79.3
Neuroticism (2 classes)	*50.0*	**65.7**	53.0	63.1

Gender

Our corpus for both gender and age, first described by Schler et al. (2006), was assembled by taking as an initial set all 47,000 blogs in blogger.com (as of August 2004) that self-reported both age and gender, and included at least 200 occurrences of common English words. After dividing the set into age intervals, we selected equal numbers of male and female bloggers in each age interval by randomly eliminating surplus. The final corpus consists of the full set of postings of 19,320 blog authors (Each text is the full set of posts by a given author.) ranging in length from several hundred to tens of thousands of words, with a mean length of 7,250 words per author.

Classification results for gender are shown in the first line of Table 5. As is evident, all feature sets give effective classification while the content features are slightly better than are style features.

In the first line of Table 6, we show the most discriminating style and content features, respectively, for gender. As can be seen, the style features most useful for gender discrimination are determiners and prepositions (markers of male writing) and pronouns (markers of female writing). The content features most useful for gender discrimination are words related to technology (male) and words related to personal life and relationships (female). Earlier studies (Argamon et al., 2003) on author gender in both fiction and nonfiction have shown that the style features found here to be useful for blogs are strong discriminators in other types of text as well.

Age

Based on each blogger's reported age, we label each blog in our corpus as belonging to one of three age groups: 13–17 (42.7%), 23–27 (41.9%), and 33–47 (15.5%) years. Intermediate age groups were removed to avoid ambiguity since many of the blogs were written over a period of several years. Our objective is to identify to which of these three age intervals an anonymous author belongs.

Accuracy results for age classification are shown in the second line of Table 5. Both style and content features give us over 76% accuracy for this three-way classification problem while the baseline majority-class classifier would give an accuracy of just 42.7%.

TABLE 6. Most important style and content features (by information gain) for each class of texts in each profiling problem.

Class	Style features	Content features
Female	**personal pronoun,** *I, me, him, my*	*cute, love, boyfriend, mom, feel*
Male	**determiner,** *the, of,* **preposition-matter,** *as*	*system, software, game, based, site*
Teens	*im, so, thats, dont, cant*	*haha, school, lol, wanna, bored*
20s	**preposition, determiner,** *of, the, in*	*apartment, office, work, job, bar*
30s+	**preposition,** *the,* **determiner,** *of, in*	*years, wife, husband, daughter, children*
Bulgarian	**conjunction-extension, pronoun-interactant,** *however,* **pronoun-conscious,** *and*	*bulgaria, university, imagination, bulgarian, theoretical*
Czech	**personal pronoun,** *usually, did, not, very*	*czech, republic, able, care, started*
French	*indeed,* **conjunction-elaboration,** *will,* **auxverb-future, auxverb-probability**	*identity, europe, european, nation, gap*
Russian	*can't, i, can, over, every*	*russia, russian, crimes, moscow, crime*
Spanish	**determiner-specific,** *this, going_to, because, although*	*spain, restoration, comedy, related, hardcastle*
Neurotic	*myself,* **subject pronoun, reflexive pronoun, preposition-behalf, pronoun-speaker**	*put, feel, worry, says, hurt*
Non-neurotic	*little,* **auxverbs-obligation, nonspecific determiner,** *up,* **preposition-agent**	*reading, next, cool, tired, bed*

The style features most useful for age classification (Table 6) are contractions without apostrophes (younger writing), and determiners and prepositions (older writing). Note that the strongest style features for those authors in their 20s and 30s are identical; they are those that distinguish both of these classes from teenagers. The content features that prove to be most useful for discrimination are words related to school and mood for teens, to work and social life for those in their 20s, and to family life for those in their 30s.

Native Language

For the problem of determining an author's native language, we use a portion of the *International Corpus of Learner English* (Granger, Dagneaux, & Meunier, 2002). All writers in the corpus are university students (mostly in their 3rd or 4th year) studying English as a second language and assigned to the same proficiency level in English. We consider 1,290 texts in five subcorpora, comprising 258 writers from Russia, the Czech Republic, Bulgaria, France, and Spain, respectively. All texts in the resulting corpus are between 579 and 846 words long. Our objective is to determine which of the five languages is the native tongue of an anonymous author writing in English.

Accuracy results are shown in the third line of Table 5. Both style and content features give results above 65%, well above the baseline accuracy of 20%.

In Table 6, we can see some consistent patterns of usage in the style features. For example, as might be expected, native speakers of Slavic languages (i.e., Russian, Bulgarian, Czech) tend to omit the definite article *the*, which does not exist in these languages (Since we list only features that are *overrepresented* in a given class, this feature is seen by examining the list of features for Spanish. Indeed, many of the most discriminating features are those that are *underrepresented* for particular languages.) Furthermore, those words with commonly used analogs in a given language are used with greater frequency by native speakers of that language, such as *indeed* (French), *over* (Russian), and *however* (Bulgarian).

Elsewhere (Koppel et al., 2005), we have shown that for determining native language, features that measure stylistic idiosyncrasies and errors are particularly useful. Using such features together with the style features considered in this section yields classification accuracy of over 80% for this task.

Regarding content words, it should be noted that unlike the text collections used in the other experiments described in this article, writers in the learner corpus did not necessarily freely choose their writing topics, so that differences in content word usage here are plausibly artifacts of the experimental setup.

Personality

To examine the extent to which personality type can be determined from writing style, we use a corpus of essays written by psychology undergraduates at the University of Texas at Austin. Students were instructed to write a short "stream of consciousness" essay wherein they tracked their thoughts and feelings over a 20-min free-writing period. The essays range in length from 251 to 1,951 words. Each writer also filled out a questionnaire testing for the "Big Five" personality dimensions: *neuroticism*, *extraversion*, *openness*, *conscientiousness*, and *agreeableness* (John, 1990). To illustrate personality profiling, we consider just the dimension of neuroticism; methods and results for other personality factors are qualitatively similar. To formulate this as a classification problem, we define "positive" examples to be the participants with neuroticism scores in the upper third and "negative" examples to be those with scores in the lowest third. The rest of the data is ignored; the final corpus consists of 198 writing samples.

Accuracy results are shown in the fourth line of Table 5. Notably, style features give a great deal of information about personality. An accuracy rate of 65.7% in detecting neuroticism is surprisingly high; independent studies of individuals who attempted to guess others' neuroticism levels have given an average accuracy of 69%, even among people who have known each other for several years (Vazire, 2006).

As shown in Table 6, the most discriminating style features for this task suggest that neurotics tend more to refer to themselves, to use pronouns for subjects rather than as objects in a clause, to use reflexive pronouns, and to consider explicitly who benefits from some action (through prepositional phrases involving, e.g., *for* and *in order to*); nonneurotics, on the other hand, tend to be less concrete and use less precise specification of objects or events (determiners and adjectives, e.g., *a* or *little*), and to show more concern with how things are or should be done (via prepositions, e.g., *by* or *with*, and modals, e.g., *ought to* or *should*).

In fact, classifiers learned using only the 10 style features shown in Table 6 give classification accuracy of 63.6%. More surprisingly, although the results in Table 5 indicate that content words overall are useless for classifying texts by neuroticism, using as features the 10 most informative content features (i.e., those in Table 5) gives an accuracy of 68.2%. Apparently, the vast majority of content is irrelevant to this classification problem and masks a small number of features involving worry about personal problems (neurotics) and relaxation activities (nonneurotics) that are quite useful for this task.

Finding a Needle-in-a-Haystack

Consider now the scenario where we seek to determine the specific identity of a document's author, but there are many thousands of potential candidates. We call this the needle-in-a-haystack attribution problem. In this case, standard text-classification techniques are unlikely to give reasonable accuracy and may require excessive computation time to learn classification models; however, we will show in this section that if we are willing to tolerate our system telling us it does not know the answer, we can achieve high accuracy for the cases where the system does give us an attribution it considers reliable.

The blogosphere forms a convenient test bed for this problem, as it provides us with text written by an essentially unlimited number of authors. For this study, we use the blog corpus described earlier, choosing the 20,000 longest blogs in our initial set. We took 10,000 blogs randomly to create a test set of "snippets," each snippet comprising enough of the most recent posts of a blog to total at least 500 words; the remainder of that blog is termed the author's "known work." The other 10,000 blogs are held out for training purposes, as is described next.

The goal then is to determine, for each snippet, to which of the 10,000 blogs it belongs, by comparison to the various known works. We address the problem independently for each of the snippets in the test set; that is, we do not make use of the fact that there is a one-to-one correspondence between the blogs and the snippets.

A Simple Attribution Method

Learning classification models for a 10,000-class problem with thousands of features is impractical. So, as a first approximation, we apply the standard information-retrieval technique in which we define some distance measure over meaningful textual features and attribute each snippet to the closest blog in that feature space. Related approaches to authorship problems have been considered by Novak, Raghavan, and Tomkins (2004) and Abbasi and Chen (2008).

We represent each text in four different ways: three varieties of *tf-idf* representations based on the 1,000 most frequent content features in the text and another based on a *tf-idf* representation based on style features. For each of these representation methods, we use the standard cosine measure (Salton & Buckley, 1988) to quantify the similarity of each author's known work with a given snippet. The various authors can then be ranked according to the similarity between their known works and the snippet under consideration, with the hope that the highest-ranked author is the author of the snippet. The idea is that some distinctive feature(s) might render the snippet particularly similar to just one of the candidate authors.

This simple approach to the problem actually works surprisingly well. The three content representations assign the snippet to the actual author between 52 and 56%, respectively, of the time while the style representation lags behind with only 6% of snippets assigned to the actual author. Sixty-four percent of the snippets are most similar to their actual authors' known works in at least one of the four representation methods. Thus, there is a great deal of useful information here.

Meta-Learning

While 56% may seem to be quite a high level of accuracy, given the large number of candidates and the simplicity of the method, it also is quite useless in the sense that we are still unable to confidently assert that a given snippet was written by a given author; after all, the system is still wrong almost half of the time. Thus, we would like to automatically determine which attributions by which representation schemes have a high likelihood of being correct; when none of them do, the system will report that results are inconclusive. The goal is to return specific attributions as often as possible while ensuring a tolerably high level of accuracy for those cases.

To accomplish this, we apply a meta-learning scheme, using the holdout set of 10,000 blogs (i.e., those not included in the test set) set aside for this purpose. We consider each pair consisting of a snippet and an author ranked most similar to that snippet for at least one representation method, in a given blog set (holdout or test). We call the pair a successful pair if the candidate author is in fact the actual author. The pairs over the holdout set are used as training to learn a model that distinguishes successful pairs from unsuccessful pairs. Each example (pair) is represented in terms of a set of meta-features reflecting, for each representation, the similarity of the author to the snippet, both absolutely and relative to other authors, and the author's rank in similarity relative to other authors.

A linear SVM is used for each representation method to learn a "meta-model" that decides whether a given pair is reliable. To do this, we use the meta-model to compute a *reliability score*, which is a monotonic function with range [0,1]

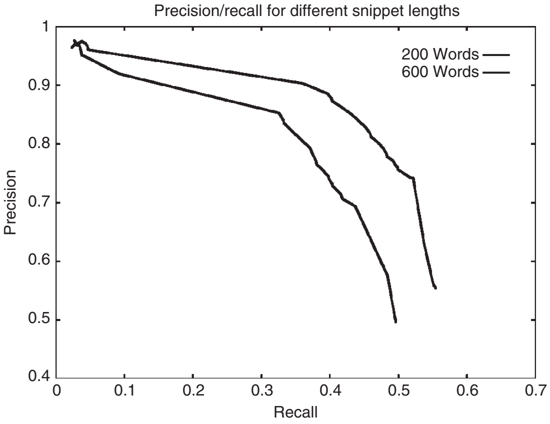

FIG. 1. Precision/recall curves for attribution, adjusting the SVM threshold for deciding whether the highest scoring attribution should in fact be made. Upper curve is for snippets limited to 600 words, and lower is for snippets limited to 200 words. Recall (percentage of possible attributions correctly made) is on the *x* axis, and precision (percentage of actual attributions correctly made) is on the *y* axis.

of the distance of the pair's representation from the SVM margin.

Given reliability scores for each of the representations, the system chooses the attribution of the highest scoring representation, provided its reliability score is above a predetermined threshold. Otherwise, the output is *Don't Know*. Varying this threshold will change the number of attributions made and the accuracy of those attributions. This enables us to plot recall/precision curves (Figure 1, upper curve), where recall is defined as the fraction of possible attributions (i.e., number of authors represented by snippets and by known works; in this case, 10,000) that were correctly attributed, and precision is defined as the fraction of attempted attributions that were in fact correct. Note that, for example, we can achieve recall of 40% with precision of 87%, but if we can settle for recall of 30%, we can get precision of 94%.

To test the sensitivity of these results for snippet length, we ran the experiment for snippets limited to 200 words. In this case (see Figure 1, lower curve), at a recall level of 30%, we achieve precision of 86%; at recall of 40%, we get precision of 73%.

Thus, provided we are willing to live with the response *Don't Know* in a number cases, we can achieve reasonably reliable authorship attribution even where the number of candidate authors is in the many thousands and where the texts are rather short.

Unattributable Texts

In the real world, however, we cannot assume that the author of a questioned text will in fact be contained in our candidate set, even if that set is very large. To evaluate the performance of our method in such a scenario, we randomly discarded 5,000 of the known works from the candidate set and evaluated performance on the original 10,000 snippets. Now, half of our test cases ought to result in an output of

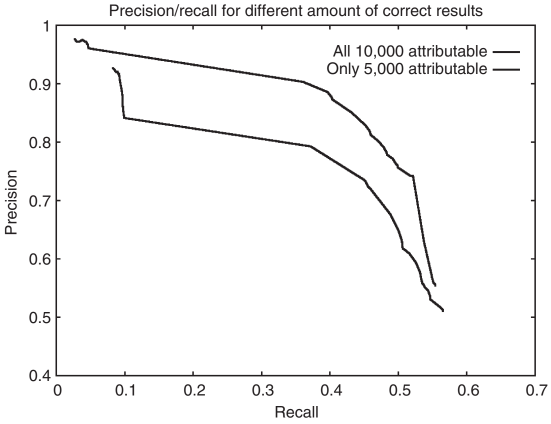

FIG. 2. Precision/recall curves (as in Figure 1) for attribution of 10,000 snippets where all 10,000 are theoretically attributable (upper curve) and where only 5,000 are theoretically attributable.

Don't Know in the best case since their actual authors are not in the candidate set. The precision/recall curve for this case is the lower curve in Figure 2 (Note that recall here is defined as the fraction of the 5,000 possible attributions that are correctly made.), shown along with the original 600-word curve for comparison. In this case, at a recall level of 30%, we achieve precision of 81%; at recall of 40%, we get precision of 72%. Clearly, performance is noticeably degraded relative to the case where all snippets have authors in the candidate set, although useful accuracy levels are still attainable.

Note that as the number of alternative candidates becomes much smaller, the problem might, somewhat counterintuitively, become more difficult. This is because our method implicitly leverages the fact that if a document is much more similar to one author's writing than to those of all others, it is very likely the document was written by that author. As the number of alternative authors decreases, the reliability of such a conclusion will similarly decrease. Thus, in the extreme case of *authorship verification*, where we are faced with a single candidate author, we need an entirely different method. It is to this problem that we now turn.

Authorship Verification

Consider the case in which we are given examples of the writing of a single author and are asked to verify that a given target text was or was not written by this author. As a categorization problem, verification is significantly more difficult than basic attribution, and virtually no work has been done on it (but see van Halteren, 2004), outside the framework of plagiarism detection (Clough, 2000; Meyer zu Eissen, Stein, & Kulig, 2007). If, for example, all we wished to do is to determine if a text had been written by Shakespeare or by Marlowe, it would be sufficient to use their respective known writings, to construct a model distinguishing them, and to test the unknown text against the model. If, on the other hand, we need to determine if a text was written by Shakespeare, it is difficult to assemble a representative sample of non-Shakespeare texts.

The situation in which we suspect that a given author may have written some text, but do not have an exhaustive list of alternative candidates, is a common one. The problem is complicated by the fact that a single author may vary his or her style from text to text or may unconsciously drift stylistically over time, not to mention the possibility of conscious deception. Thus, we must learn to somehow distinguish between relatively shallow differences that reflect conscious or unconscious changes in an author's style and deeper differences that reflect styles of different authors.

Verification can be thought of as a one-class classification problem (Manevitz & Yousef, 2001; Scholkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001; Tax, 2001). But perhaps a better way to think about authorship verification is that we are given two example sets and are asked whether these sets were generated by the same process (i.e., author) or by two different processes. This section, drawn from Koppel et al. (2007), describes a method for adducing the depth of difference between two example sets, which may have far-reaching consequences for determining the reliability of classification models. The idea is to test the extent to which the accuracy of learned models degrades as the most distinguishing features are iteratively removed from the learning process.

This method provides a robust solution to the authorship verification problem that is independent of language, period, and genre and already has been used to settle at least one outstanding literary attribution problem (Koppel & Schler, 2004; Koppel et al., 2007).

Authorship Verification: Naïve Approaches

Let us begin by considering two naïve approaches to the problem. Although neither of them will prove satisfactory, each will contribute to our understanding of the problem.

One possibility that suggests itself is what we will call the "impostors" method: Assemble a representative collection of works by other authors and to use a two-class learner, such as SVM, to learn a model for A versus not-A. Then chunk the mystery work X and run the chunks through the learned model. If the preponderance of chunks of X are classed as A, then X is deemed to have been written by A; otherwise, it is deemed to have been written by someone else.

This method is straightforward, but suffers from a conceptual flaw. While it is indeed reasonable to conclude that A is not the author if most chunks are attributed to not-A, the converse is not true. Any author who is neither A nor represented in the sample not-A, but who happens to have a style more similar to A than to not-A, will be falsely determined by this method to be A. Despite this flaw, we will see later that this approach can be used to augment other methods.

Another approach, which does not depend on negative examples, is to learn a model for A versus X and assess the extent of the difference between A and X by evaluating generalization accuracy by cross-validation. If cross-validation accuracy is high, then conclude that A did not write X; if cross-validation accuracy is low (i.e., we fail to correctly classify test examples better than chance), then conclude that A did write X. This intuitive method does not actually work well at all.

Let us consider exactly why the last method fails by examining a real-world example. Suppose we are given known works by three of the authors considered earlier, 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 Seven Gables* (henceforth, *Gables*). Using the method just described and using a feature set consisting of the 250 most frequently used words in A and X, 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.

A New Approach: Unmasking

If we look closely at the models that successfully distinguish *Gables* from Hawthorne's other work (in this case, *The Scarlet Letter*), we find that only a small number of features are doing all the work of distinguishing between them. These features include *he* (more frequent in *The Scarlet Letter*) and *she* (more frequent 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, from differences in genre or purpose, from chronological stylistic drift, or from 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. To do this, we use a technique we call "unmasking." 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 there are between them will be reflected in only a relatively small number of features, despite possible differences in theme, genre, and the like.

unmasking
theory

In Figure 3, we show the result of unmasking when comparing *Gables* to known works of Melville, Cooper, and Hawthorne. This graph illustrates our hypothesis: 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. Once a relatively small number of distinguishing markers are removed, the two works by Hawthorne become nearly indistinguishable.

This phenomenon is actually quite general, as we will show later. As we also will see, the suddenness of the degradation can be quantified in a fashion optimal for this task. Thus, by taking into account the depth of difference between two works, we can determine if they were authored by the same person or by two different people.

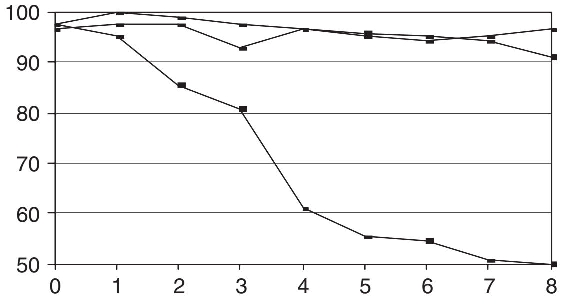

FIG. 3. Tenfold cross-validation accuracy of models distinguishing *House of Seven Gables* from each of Hawthorne, Melville, and Cooper. The *x* axis represents the number of iterations of eliminating best features at previous iteration. The curve well below the others is that of Hawthorne, the actual author.

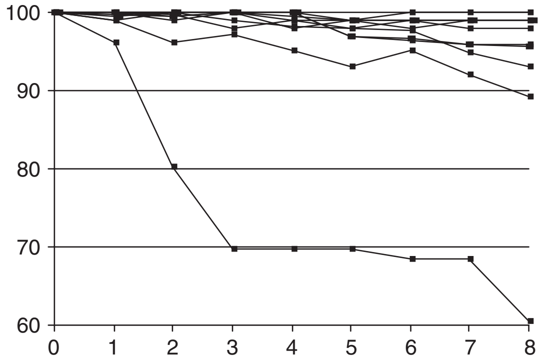

FIG. 4. Unmasking *An Ideal Husband* against each of the 10 authors ($n = 250$, $k = 3$). The curve below all the authors is that of Oscar Wilde, the actual author (Several curves are indistinguishable.)

Test Corpus

We use as our corpus the collection of classic 19th- and early 20th-century books considered earlier. To break up the two-books-per-author pattern in the corpus, we add to the corpus one additional work by Melville and one by Hawthorne as well as a work by Emily Bronte, who has no other work in the corpus.

Our objective is to run 209 independent authorship verification experiments representing all possible author/book pairs (21 books × 10 authors, but excluding just the pair Emily Bronte/*Wuthering Heights*, which cannot be tested since it is the author's only work).

As previously, we partitioned each book into approximately equal-length sections of at least 500 words without breaking up paragraphs. For each author A and each book X, let A_X consist of all the works by A in the corpus, unless X is in fact written by A, in which case A_X consists of all works by A except X. Our objective is to assign to each pair <A_X,X> the value *same-author* if X is by A and the value *different-author* otherwise.

Unmasking Applied

Now let us introduce the details of our new method based on our earlier observations regarding iterative elimination of features. We choose as an initial feature set the *n* words with highest average frequency in A_X and X (i.e., the average of the frequency in A_X and the frequency in X, giving equal weight to A_X and X). Note that our objective here is not to maximize accuracy but rather to measure the degradation of accuracy; thus, it is enough to choose a simple feature set rather than the best possible one.

Using an SVM with linear kernel, we run the following unmasking scheme:

1. *Determine the accuracy results of a 10-fold cross-validation experiment for A_X against X* (If one of the sets, A_X or X, includes more chunks than the other, we randomly discard the surplus. Accuracy results are the average of five runs of 10-fold cross-validation in which we discard randomly for each run.)
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 each pair <A_X,X>. In Figure 4, we show such curves (using $n = 250$ and $k = 3$) for *An Ideal Husband* against each of 10 authors, including Oscar Wilde.

Meta-Learning: Identifying Same-Author Curves

We wish now to quantify the difference between *same-author* curves and *different-author* curves. To do so, we first represent each curve as a numerical vector in terms of its essential features. These features include, for $i = 0, \ldots, m$:

- accuracy after *i* elimination rounds,
- accuracy difference between Round *i* and $i + 1$,
- accuracy difference between Round *i* and $i + 2$,
- *i*th highest accuracy drop in one iteration, and
- *i*th highest accuracy drop in two iterations.

We sort these vectors into two subsets: those in which A_X and X are the by same author and those in which A_X and X are by different authors. We then apply a meta-learning scheme in which we use learners to determine what role to assign to various features of the curves (Note that although we have 20 *same-author* pairs, we really only have 13 distinct *same-author* curves since for authors with exactly two works in our corpus, the comparison of A_X with X is identical for each of the two books.)

To assess the accuracy of the method, we use the following cross-validation methodology. For each Book B in our corpus, we run a trial in which B is completely eliminated from consideration. We use unmasking to construct curves for all author/book pairs <A_X,X> (where B does not appear in A_X and is not X), and then we use a linear SVM to meta-learn to distinguish *same-author* curves from *different-author* curves. Then, for each Author A in the corpus, we use unmasking to construct a curve for the pair <A_B,B> and use the meta-learned model to determine if the curve is a *same-author* curve or a *different-author* curve.

DOI: 10.1002/asi

```
Given: anonymous book X, works of suspect author A,
(optionally) impostors {A1,…,An}

Step 1 - Impostors method(optional)
if impostors {A1,…,An} are given then
{
  For each impostor Ai
  {
    Build model Mi for classifying A vs. Ai
    Test each chunk of X with built model Mi
  }
  If for some Ai number of chunks assigned to Ai > number of chunks assigned to A
  then
    return different-author
}
Impostors_Method_END

Step 2 - Unmasking
Build degradation curve <A,X>
Represent degradation curve as feature vector (see text)
Test degradation curve vector (see text)
if test result positive
  return same-author
else
  return different-author
Unmasking_END

Method Build Degradation Curve:
  Use 10 fold cross validation for learning A against X
  For each fold
  {
    Do m iterations
    {
      Build a model for A against X
      Evaluate accuracy results
      Add accuracy number for iteration m to degradation curve <A,X> (as average for all iterations
                                                                      m on current fold)
      Remove k top contributing features (in each direction) from data
    }
  }
Method_END
```

FIG. 5. Overview of the authorship verification algorithm.

Using this testing protocol, we obtain the following results: All but one (*Pygmalion* by Shaw) of the 20 *same-author* pairs are correctly classified. In addition, 181 of 189 *different-author* pairs are correctly classified. Among the exceptions are the attributions of *The Professor* by Charlotte Bronte to each of her sisters. Thus, we obtain overall accuracy of 95.7%, with errors almost identically distributed between false positives and false negatives (Note that some of the eight misclassified *different-author* pairs result in a single book being attributed to two authors, which is obviously impossible. Nevertheless, since each of our author/book pairs is regarded as an independent experiment, we do not leverage this information.)

Note that the algorithm includes three parameters: n, the size of the initial feature set; k, the number of eliminated features from each extreme in each iteration; and m, the number of iterations we consider. The results reported previously are based on experiments using $n = 250$, $k = 3$, and $m = 10$. We chose $n = 250$ because experimentation indicated that this was a reasonable rough boundary between common words and words tightly tied to a particular work. Koppel et al. (2007) showed that results are somewhat robust with regard to choice of k and m (In fact, some parameter choices turn out to be better than those shown here.), but the recall results for *same-author* degrade considerably as the size of the initial feature set increases. Moreover, parameter settings that proved successful on the English literature corpus considered here also proved successful on a corpus of Hebrew legal writings, thus demonstrating some degree of robustness over variation in language and genre.

Koppel et al. (2007) further showed that unmasking can be augmented by exploiting known negative examples, using the “impostors” method described earlier; the augmented method correctly classes all 189 *different-author* pairs and 18 of 20 *same-author* pairs. Finally, one limitation is that unmasking requires a large amount of training text (Sanderson & Guenter, 2006) ; preliminary tests suggest that the minimum would be in the area of 5,000 to 10,000 words.

In Figure 5, we summarize the entire algorithm (including the optional augmentation using negative examples).

Conclusions

We have surveyed the variety of feature types and categorization methods that have been proposed in the past for authorship attribution. These methods ranged from early attempts to find individual statistical markers that could serve as authorial fingerprints, through multivariate methods of varying degrees of sophistication, and ultimately to text categorization methods rooted in machine learning. We conclude that two of the most sophisticated machine learning methods, SVM and Bayesian regression, used in conjunction with word classes derived from SFL or with character n-grams, offer easily scalable, efficient, and effective solutions to the ordinary authorship attribution problem, assuming proper methodological controls for text genre and the like.

Since many realistic authorship problems do not fit the standard attribution paradigm, we consider also three variations that are likely to arise in practice. For the *profiling* problem, where no individual candidates are known, we find that we can identify, with varying degrees of accuracy, an author's gender, age, native language, and personality type. For the *needle-in-a-haystack* problem, where there are possibly many thousands of candidate authors, we find that information-retrieval methods can be used to identify the correct author—of even very short texts—with high accuracy for some considerable fraction of cases. These cases can be isolated using meta-learning methods that take into account the degree to which a single author is more likely than any of the other candidates to be the actual author. Finally, for the *verification* problem, where we need to determine if a given author wrote a given text, we find that our unmasking technique is highly effective at identifying actual authors, although it is limited to cases in which the attested text is sufficiently long.

Acknowledgments

The research reported here was partially supported by a grant from the Binational Science Foundation. The authors gratefully acknowledge the contributions of their co-authors on papers that served as a partial basis for this paper: James Pennebaker, Elisheva Bonchek-Dokow and Eran Messeri. The authors also acknowledge the assistance of Gilad Weiss in preparing the table in the Appendix. Finally, the authors are grateful to the three anonymous reviewers whose comments substantially improved the paper.

References

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group Web forum messages. IEEE Intelligent Systems, 20(5), 67–75.

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection. ACM Transactions on Information Systems, 26(2), 1–29.

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and probabilistic foundations. Literary and Linguistic Computing, 23(2), 131–147.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. (2003). Gender, genre, and writing style in formal written texts. Text, 23(3), 321–346.

Argamon, S., Koppel, M., Pennebaker, J., & Schler, J. (in press). Automatically profiling the author of an anonymous text. Communications of the ACM, in press.

Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In Proceedings of the ACH/ALLC Conference, Victoria, BC, Canada.

Argamon, S., Whitelaw, C., Chase, P., Hota, S., Garg, N., & Levitan, S. (2007). Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology, 58(6), 802–821.

Argamon-Engelson, S., Koppel, M., & Avneri, G. (1998). Style-based text categorization: What newspaper am I reading? In Proceedings of the AAAI Workshop on Learning for Text Categorization (pp. 1–4). Menlo Park, CA: AAAI Press.

Baayen, H., van Halteran, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. Journees internationales d'Analyse statistique des Donnees Textuelles, 6.

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

Benedetto, D., Caglioti, E., & Loreto, V. (2002). Language trees and zipping. Physical Review Letters, 88(4), 487–490.

Binongo, J.N.G. (2003). Who wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution. Chance, 16(2), 9–17.

Binongo, J.N.G., & Smith, M.W.A. (1999). The application of principal component analysis to stylometry. Literacy and Linguistic Computing, 14, 445–466.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In Proceedings of the 3rd Conference on Applied Natural Language Processing (pp. 152–155). East Stroudsburg, PA: ACL.

Brinegar, C.S. (1963). Mark Twain and the Quintus Curtius Snodgrass Letters: A statistical test of authorship. Journal of the American Statistical Association, 58, 85–96.

Burger, J., & Henderson, J. (2006). An exploration of features for predicting blogger age. In the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (pp. 47–54). NY: ACM Press.

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. (1989). "An ocean where each kind . . . :" Statistical analysis and some major determinants of literary style. Computers and the Humanities, 23(4), 309–321.

Burrows, J.F. (1992a). Computers and the study of literature. In C. Butler (Ed.), Computers and written text: Applied language studies (pp. 167–204). Oxford, England: Blackwell.

Burrows, J.F. (1992b). Not unless you ask nicely: The interpretative nexus between analysis and information. Literary and Linguistic Computing, 7(2), 91–109.

Burrows, J.F. (2002a). Delta: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17, 267–287.

Burrows, J.F. (2002b). The Englishing of Juvenal: Computational stylistics and translated texts. Style, 36, 677–699.

Burrows, J. (2007). All the way through: Testing for authorship in different frequency strata. Literary and Linguistic Computing, 21, 27–47.

Chambers, J.K., Trudgill, P., & Schilling-Estes, N. (2004). The handbook of language variation and change. Oxford, England: Blackwell.

Chaski, C. (2001). Empirical evaluations of language-based author identification techniques. Forensic Linguistics, 81, 1–65.

Chaski, C. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. International Journal of Digital Evidence, 4(1), 1–13.

Chaski, C. (2007, July). Multilingual forensic author identification through n-gram analysis. Paper presented at the 8th Biennial Conference on Forensic Linguistics/Language and Law, Seattle, WA.

Chung, C.K., & Pennebaker, J.W. (2007). The psychological function of function words. In K. Fiedler (Ed.), Social communication: Frontiers of social psychology (pp. 343–359). New York: Psychology Press.

Clement, R., & Sharp, D. (2003). Ngram and Bayesian classification of documents. Literary and Linguistic Computing, 18, 423–447.

Clough, P. (2000). Plagiarism in natural and programming languages: An overview of current tools and technologies. University of Sheffield, UK, Research Memoranda No. CS-00–05, Department of Computer Science.

Corney, M., de Vel, O., Anderson, A., & Mohay, G. (2002). Gender-preferential text mining of e-mail discourse. In Proceedings of the 18th annual Computer Security Applications Conference, pp. 282–289. Piscataway, NJ: IEEE Press.

Craig, H. (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? Literacy and Linguistic Computing, 14, 103–113.

Damashek, M. (1995). Gauging similarity with n-grams: Language independent categorization of text. Science, 267(5199), 843–848.

de Vel, O., Anderson, A., Corney, M., & Mohay, G.M. (2001). Mining e-mail content for author identification forensics. SIGMOD Record, 30(4), 55–64.

de Vel, O., Corney, M., Anderson, A., & Mohay, G. (2002). E-mail authorship attribution for computer forensics. In D. Barbará & S. Jajodia (Eds.), Applications of data mining in computer security. Dordrecht, The Netherlands: Kluwer.

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship attribution with support vector machines. Applied Intelligence, 19(1), 109–123.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. Proceedings of ACM-CIKM 1998 (pp. 148–155). NY: ACM Press.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3(1), 1289–1305.

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

Foster, D. (2000). Author unknown: On the trail of anonymous. New York: Holt.

Fucks, W. (1952). On the mathematical analysis of style. Biometrica, 39, 122–129.

Gamon, M. (2004). Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In Proceedings of the 20th International Conference of Computational Linguistics (pp. 611–617), Geneva.

Genkin, A., Lewis, D., & Madigan, D. (2006). Large-scale Bayesian logistic regression for text categorization. Technometrics. 49(3), 291–301.

Graham, N., Hirst, G., & Marthi, B. (2005). Segmenting documents by stylistic character. Natural Language Engineering, 11(4), 397–415.

Granger, S., Dagneaux, E., & Meunier, F. (2002). The international corpus of learner English. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 22(3), 251–270.

Halliday, M.A.K. & Matthiessen, C.M.I.M. (2003). An introduction to functional grammar. London: Hodder Arnold.

Hirst, G., & Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing, 22(4), 405–417.

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

Holmes, D. (2003). Stylometry and the Civil War. Chance, 16(2), 18–25.

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

Holmes, D.I., Gordon, L., & Wilson, C. (2001). A widow and her soldier: Stylometry and the American Civil War. Literary and Linguistic Computing, 16(4), 403–420.

Holmes, D.I., Robertson, M., & Paez, R. (2001). Stephen Crane and the New York Tribune: A case study in traditional and non-traditional authorship attribution. Computers and the Humanities, 35(3), 315–331.

Honore, A. (1979). Some simple measures of richness of vocabulary. Association for Literary and Linguistic Computing Bulletin, 7(2), 172–177.

Hoorn, J., Frank, S., Kowalczyk, W., & van der Ham, F. (1999). Neural network identification of poets using letter sequences. Literary and Linguistic Computing, 14(3), 311–338.

Hoover, D.L. (2002). Frequent word sequences and statistical stylistics. Literary and Linguistic Computing, 17, 157–180.

Hoover, D.L. (2003a). Frequent collocations and authorial style. Literary and Linguistic Computing, 18, 261–286.

Hoover, D.L. (2003b). Multivariate analysis and the study of style variation. Literary and Linguistic Computing, 18, 341–360.

Hoover, D.L. (2003c). Another perspective on vocabulary richness. Computers and the Humanities, 37, 151–178.

Hoover, D. (2004a). Testing Burrows's Delta. Literary and Linguistic Computing, 19(4), 453–475.

Hoover, D. (2004b). Delta prime? Literary and Linguistic Computing, 19(4), 477–495.

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. In Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (pp. 77–86). NY: Springer.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning (ECML-98) (pp. 137–142). NY: Springer.

John, O.P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In O.P. John & L.A. Pervin (Eds.), Handbook of personality: Theory and research (pp. 66–100). New York: Guilford Press.

Juola, P. (1998). Cross-entropy and linguistic typology. In Proceedings of New Methods in Language Processing 3. Sydney, Australia.

Juola, P. (2008). Author attribution, Foundations and Trends in Information Retrieval, 1(3), 233–334.

Juola, P., & Baayen, H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. Literary and Linguistic Computing, 20(Suppl. 1), 59–67.

Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In Proceedings of the 15th Conference on Computational Linguistics (pp. 1071–1075), Kyoto, Japan.

Keselj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In Proceedings of PACLING 2003 (pp. 255–264), Halifax, Canada.

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of genre. In Proceedings of the 35th annual meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics (pp. 32–38). East Stroudsburg, PA: ACL.

Khmelev, D.V. (2001). Disputed authorship resolution through using relative empirical entropy for Markov Chains of letters in human language text. Journal of Quantitative Linguistics, 7(3), 201–207.

Khmelev, D.V., & Teahan, W.J. (2003). A repetition based measure for verification of text collections and for text categorization. In Proceedings of the 26th SIGIR Conference (pp. 104–110). NY: ACM Press.

Khmelev, D.V., & Tweedie, F.J. (2002). Using Markov chains for identification of writers. Literary and Linguistic Computing, 16(4), 299–307.

Kjell, B. (1994a). Authorship attribution of text samples using neural networks and Bayesian classifiers. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (p. 1660), San Antonio, TX.

Kjell, B. (1994b). Authorship determination using letter pair frequencies with neural network classifiers. Literary and Linguistic Computing, 9(2), 119–124.

Kjell, B., Woods, W.A., & Frieder, O. (1995). Information retrieval using letter tuples with neural network and nearest neighbor classifiers. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2, 1222–1225, Vancouver, BC.

Koppel, M., Akiva, N., & Dagan, I. (2006). Feature instability as a criterion for selecting potential style markers. Journal of the American Society for Information Science and Technology, 57(11), 1519–1525.

Koppel, M., Argamon, S., & Shimoni, A. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401–412.

Koppel, M., Mughaz, D., & Akiva, N. (2006). New methods for attribution of Rabbinic literature. Hebrew Linguistics: A Journal for Hebrew Descriptive, Computational and Applied Linguistics, 57, 5–18.

Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of the IJCAI 2003 Workshop on Computational Approaches to Style Analysis and Synthesis (pp. 69–72), Acapulco, Mexico.

Koppel, M., & Schler, J. (2004). Authorship verification as a one class classification problem. In Proceedings of ECML (p. 62), Banff, Canada.

Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship attribution with thousands of candidate authors. In Proceedings of the 29th ACM SIGIR Conference on Research & Development on Information Retrieval (pp. 659–660). NY: ACM Press.

Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research, 8, 1261–1276.

Koppel, M., Schler, J., & Zigdon, K. (2005). Determining an author's native language by mining a text for errors. In Proceedings of KDD 2005 (pp. 624–628), Chicago, IL.

Kukushkina, O.V., Polikarpov, A.A., & Khmelev, D.V. (2001). Using literal and grammatical statistics for authorship attribution. Problems of Information Transmission. 37(2), 172–184.

Ledger, G., & Merriam, T. (1994). Shakespeare, Fletcher, and the two noble kinsmen. Literacy and Linguistic Computing, 9, 235–248.

Lewis, D.D., & Ringuette, M. (1994). Comparison of two learning algorithms for text categorization. In Proceedings of the 3rd annual Symposium on Document Analysis and Information Retrieval (SDAIR '94) (pp. 81–93).

Li, J., Zheng, R., & Chen, H. (2006). From fingerprint to writeprint. Communications of the ACM, 49(4), 76–82.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear threshold algorithm. Machine Learning, 2(4), 285–318.

Love, H. (2002). Attributing authorship: An introduction. Cambridge University Press.

Lowe, D., & Matthews, R. (1995). Shakespeare vs. Fletcher: A stylometric analysis by Radial Basis Functions. Computers and the Humanities, 29, 449–461.

Madigan, D., Genkin, A., Lewis, D.D., Argamon, S., Fradkin, D., & Ye, L. (2006). Author identification on the large scale. In Proceedings of the Classification Society of North America.

Manevitz, L.M., & Yousef, M. (2001). One-class svms for document classification. Journal of Machine Learning Research, 2, 139–154.

Martindale, C., & McKenzie, D. (1995). On the utility of content analysis in author attribution: The "Federalist." Computers and the Humanities, 29, 259–270.

Marton, Y., Wu, N., & Hellerstein, L. (2005). On compression-based text classification. In Proceedings of the 27th European Conference on IR Research (pp. 300–314). NY: Springer.

Mascol, C. (1888a). Curves of pauline and pseudo-pauline style i. Unitarian Review, 30, 452–460.

Mascol, C. (1888b). Curves of pauline and pseudo-pauline style ii. Unitarian Review, 30, 539–546.

Matthews, R., & Merriam, T. (1993). Neural computation in stylometry: An application to the works of Shakespeare and Fletcher. Literary and Linguistic Computing, 8(4), 203–209.

Matthiessen, C. (1992). Lexicogrammatical cartography. Tokyo: International Languages Sciences.

McEnery, A., & Oakes, M. (2000). Authorship studies/textual statistics. In R. Dale, H. Moisl, & H. Somers (Eds.), Handbook of natural language processing (pp. 545–562). New York: Marcel Dekker.

Mealand, D.L. (1995). Correspondence analysis of Luke. Literacy and Linguistic Computing, 10, 171–182.

Mendenhall, T.C. (1887). The characteristic curves of composition. Science, 9, 237–249.

Merriam, T. (1996). Marlowe's hand in Edward III revisited. Literary and Linguistic Computing, 11(1), 19–22.

Merriam, T., & Matthews, R. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. Literary and Linguistic Computing, 9, 1–6.

Meyer zu Eissen, S., Stein, B., & Kulig, M. (2007). Plagiarism detection without reference collections. In R. Decker & H.J. Lenz (Eds.), Advances in data analysis (pp. 359–366). NY: Springer.

Morton, A.Q. (1965). The authorship of Greek prose. Journal of the Royal Statistical Society (A), 128, 169–233.

Morton, A.Q. (1978). Literary detection. New York: Scribners.

Mosteller, F., & Wallace, D.L. (1964). Inference and disputed authorship: The Federalist. Reading, MA: Addison-Wesley.

Novak, J., Raghavan, P., & Tomkins, A. (2004). Anti-aliasing on the web. In Proceedings of the 13th International World Wide Web Conference (pp. 30–39).

O'Donnell, B. (1966). Stephen Crane's The O'Ruddy: A problem in authorship discrimination. In J. Leed (Ed.), The computer and literary style (pp. 107–15). Kent, OH: Kent State University Press.

Pavelec, D., Justino, E., & Oliveira, L.S. (2007). Author identification using stylometric features. Inteligencia Artificial, 11(36), 59–65.

Peng, F., Schuurmans, D., & Wang, S. (2004). Augmenting Naive Bayes text classifier with statistical language models. Information Retrieval, 7(3–4), 317–345.

Pennebaker, J.W., & King, L.A. (1999). Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77(6), 1296–1312.

Pennebaker, J.W., Mehl, M.R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology, 54, 547–577.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report No. MST TR 98(14), Microsoft Research.

Quinlan, J.R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.

Rudman, J. (1997). The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities, 31(4), 351–365.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management: An International Journal, 24(5), 513–523.

Sanderson, C., & Guenter, S. (2006). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. Proceedings of the International Conference on Empirical Methods in Natural Language Processing (pp. 482–491).

Schler, J. (2007). Authorship attribution in the absence of a closed candidate set. Unpublished doctoral dissertation, Bar-Ilan University, Ramat-Gan, Israel, Department of Computer Science.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of age and gender on blogging. In Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (pp. 199–205). Menlo Park, CA: AAAI Press.

Schülkopf, B., Platt, J., Shawe-Taylor, J., Smola, A.J., & Williamson, R.C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13, 1443–1471.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47.

Sichel, H.S. (1975). On a distribution law for word frequencies. Journal of the American Statistical Association, 70, 542–547.

Sichel, H.S. (1986). Word frequency distributions and type-token characteristics. Mathematical Scientist, 11, 45–72.

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. Information Processing and Management, 44(2), 790–799.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. Computational Linguistics, 26(4), 471–495.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. Computers and the Humanities, 35, 193–214.

Stein, S., & Argamon, S. (2006). A mathematical explanation of Burrows's Delta. In Proceedings of the Digital Humanities Conference (pp. 207–209). Paris, France.

Tax, D.M.J. (2001). One-class classification. Unpublished doctoral dissertation, Technische Universiteit Delft.

Tweedie, F.J., & Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities, 32, 323–352.

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

Uzuner, O., & Katz, B. (2005). A comparative study of language models for book and author recognition. Springer Lecture Notes in Computer Science, 3651, 969–980.

van Halteren, H. (2004, July). Linguistic profiling for authorship recognition and verification. In Proceedings of the 42nd Conference of the ACL (pp. 199–206). East Stroudsburg, PA: ACL.

van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M., & Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. Journal of Quantitative Linguistics, 12(1), 65–77.

Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. Journal of Research in Personality, 40(5), 472–481.

Waugh, S., Adams, A., & Tweedie, F.J. (2000). Computational stylistics using Artificial Neural Networks. Literary and Linguistic Computing, 15(2), 187–198.

Whitelaw, C., Herke-Couchman, M., & Patrick, J. (2004, March). Identifying interpersonal distance using systemic features. AAAI Spring Symposium on Exploring Attitude and Affect in Text, Stanford, CA.

Witten, I.H., & Frank, E. (2000). Data mining: Practical machine learning tools with Java implementations. San Francisco: Kaufmann.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1(1–2), 67–88.

Yule, G.U. (1938). On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship. Biometrika, 30, 363–390.

Yule, G.U. (1944). The statistical study of literary vocabulary. Cambridge, England: Cambridge University Press.

Zhang, D., & Lee, W.S. (2006). Extracting key-substring-group features for text classification. In Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (pp. 474–483). NY: ACM Press.

Zhao, Y., & Zobel, J. (2005). Effective authorship attribution using function word. In Proceedings of the 2nd AIRS Asian Information Retrieval Symposium (pp. 174–190). NY: Springer.

Zhao, Y., & Zobel, J. (2007). Searching with style: Authorship attribution in classic literature. In Proceedings of the 30th Australasian Conference on Computer Science (Vol. 62, 59–68), Ballarat, Australia.

Zhao, Y., Zobel, J., & Vines, P. (2006). Using relative entropy for authorship attribution. In Proceedings of the 3rd AIRS Asian Information Retrieval Symposium (pp. 92–105). NY: Springer.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology, 57(3), 378–393.

Zigdon, K. (2005). Automatically determining an author's native language. Unpublished master's thesis, Bar-Ilan University, Ramat-Gan, Israel, Department of Computer Science.

Zipf, G.K. (1932). Selected studies of the principle of relative frequency in language. Cambridge, MA: Harvard University Press.

Appendix

History of studies on authorship attribution problems. For each, we identify the corpus on which methods were tested, the feature types used and the categorization method used.

Paper reference	Year	Corpus	Feature types	Classification method
Mendenhall	1887	Bacon/Marlowe/Shakespeare	Sentence length, word length	Distance
Mascol	1888a, 1888b	Pauline Epistles	FW(10s), punctuation	Distance
Yule	1938	de Gerson	Sentence length	Distance
Yule	1944	de Gerson	Vocabulary richness (K-measure)	Distance
Fucks	1952	English and German authors	Word length	Distance
Brinegar	1963	QCS letters	Word length	Distance
Mosteller & Wallace	1964	Federalist Papers	FW(10s)	NB
Morton	1965	Ancient Greek Prose	Sentence length	Distance
Burrows	1987	Austen/S.Fielding/H.Fielding	FW(10s)	MVA + PCA
Burrows	1992a	Brontes	FW(10s)	MVA + PCA
Matthews & Merriam	1993	Shakespeare/Fletcher	FW(1s)	NN
Kjell	1994a, 1994b	Federalist Papers	Character n-grams	NN, NB
Merriam & Matthews	1994	Shakespeare/Marlowe	FW(1s)	NN
Ledger & Merriam	1994	Shakespeare/Fletcher	Character n-grams	MVA
Holmes & Forsyth	1995	Federalist Papers	FW(10s), vocabulary richness	MVA, genetic algorithm
Kjell et al.	1995	WSJ	Character n-grams	NN, k-NN
Lowe & Matthews	1995	Fletcher/Shakespeare	FW(1s)	RBF-NN
Martindale & McKenzie	1995	Federalist Papers	Words	MVA + LDA, NN
Mealand	1995	Book of Luke	FW(10s), POS	MVA
Baayen et al.	1996	Federalist Papers	Syntax	NN
Merriam	1996	Shakespeare	FW(1s)	MVA + PCA
Tweedie et al.	1996	Federalist Papers	FW(1s)	NN
Argamon-Engelson et al.	1998	Newspapers & magazines	FW(100s), POS n-grams	ID3, Ripper
Tweedie & Baayen	1998	English prose	FW(10s), vocabulary richness	Distance, MVA + PCA
Binongo & Smith	1999	Shakespeare	FW(10s)	MVA + PCA

(*Continued*)

DOI: 10.1002/asi

Appendix. *(Continued)*

Paper reference	Year	Corpus	Feature types	Classification method
Craig	1999	Middleton	Words	Distance
Hoorn et al.	1999	Dutch poets	Character n-grams	NN, NB, k-NN
Stamatatos et al.	2000	Greek newspapers	Syntactic chunks	Distance
Waugh et al.	2000	Renaissance plays, Federalist Papers	Words	NN
Kukushkina et al.	2001	Russian texts	Character n-grams, POS n-grams	Distance (Markov)
Chaski	2001	Four women	Syntax, punctuation, various	Distance
de Vel et al.	2001	E-mail	FW(10s), complexity, various	SVM
Holmes, Gordon, & Wilson	2001	Pickett letters	FW(10s)	MVA + PCA
Holmes, Robertson, & Paez	2001	Crane articles (purported)	FW(10s)	MVA + PCA
Stamatatos et al.	2001	Greek newspapers	Syntactic chunks	Distance (LDA)
Baayen et al.	2002	Dutch texts	FW(10s), syntax	MVA + PCA
Benedetto et al.	2002	Italian texts	Character n-grams	Distance (compression)
Burrows	2002a, 2002b	Restoration-era poets	FW(10s)	MVA + PCA
Hoover	2002	Novels and articles	Words, word n-grams	MVA
Khmelev & Tweedie	2002	Federalist Papers, various	Character n-grams	Distance (Markov)
Binongo	2003	Oz books	FW(10s)	MVA + PCA
Clement & Sharp	2003	Movie reviews	Character n-grams	NB
Diederich et al.	2003	German newspapers	Words	SVM
Hoover	2003a	Novels and articles	Words, word n-grams	MVA
Hoover	2003b	Orwell/Golding/Wilde	Words, word n-grams	MVA
Hoover	2003c	Novels	Vocabulary richness	MVA
Keselj et al.	2003	English novels, Greek newspapers	Character n-grams	MVA
Khmelev & Teahan	2003	Russian texts	Character n-grams	Distance (Markov)
Koppel & Schler	2003	E-mail	FW(100s), POS n-grams, idiosyncrasies	SVM, J4.8
Argamon et al.	2003	BNC	FW(100s), POS n-grams	Winnow
Hoover	2004a	American novels	Words	MVA + PCA
Hoover	2004b	Novels and articles	Words	MVA + PCA
Peng et al.	2004	Greek newspapers	Character n-grams, word n-grams	NB
van Halteren	2004	Dutch texts	Word n-grams, syntax	MVA
Abbasi & Chen	2005	Arabic forum posts	Characters, words, vocabulary richness, various	SVM, J4.8
Chaski	2005	10 anonymous authors	Character n-grams, word n-grams, POS n-grams, various	Distance (LDA)
Juola & Baayen	2005	Dutch texts	FW(10s)	Distance (cross-entropy)
Zhao & Zobel	2005	Newswire stories	FW(100s)	NB, J4.8, k-NN
Koppel et al.	2005	Learner English	FW(100s), POS n-grams, idiosyncrasies	SVM
Koppel, Mughaz, & Akiva	2006	Brontes, BNC	FW(100s), POS n-grams	Balanced Winnow
Zhao et al.	2006	AP stories, English novels	FW(100s), POS, punctuation	SVM, distance
Madigan et al.	2006	Federalist Papers	Characters, FW(100s), words, various	Bayesian regression
Zheng et al., & Li et al.	2006	English and Chinese newsgroups	Characters, FW(100s), syntax, vocabulary richness, various	NN, J4.8, SVM
Argamon et al.	2007	Novels and articles	FW(100s), syntax, SFL	SVM
Burrows	2007	Restoration poets	Words	MVA + zeta
Hirst & Feiguina	2007	Brontes	Syntax	SVM
Pavelec et al.	2007	Portuguese newspapers	Conjunction types	SVM
Zhao & Zobel	2007	Shakespeare, Marlowe, various	FW(100s), POS, POS n-grams	Distance (infogain)
Abbasi &Chen	2008	E-mail, online comments, chats	Characters, FW(100s), syntax, vocabulary richness, various	SVM, PCA, other
Argamon et al.	2008	Blogs, student essays, learner English	Words, SFL	Bayesian regression
Stamatatos	2008	English and Arabic news	Character n-grams	SVM

NB = Naïve Bayes; NN = neural nets; k-NN = k nearest neighbors; MVA = multivariate analysis; PCA = principle component analysis; LDA = linear discriminant analysis.