A Survey of Modern Authorship Attribution Methods

Efstathios Stamatatos
Department of Information and Communication Systems Engineering, University of the Aegean, Karlovassi, Samos–83200, Greece. E-mail: stamatatos@aegean.gr

Authorship attribution supported by statistical or computational methods has a long history starting from the 19th century and is marked by the seminal study of Mosteller and Wallace (1964) on the authorship of the disputed "Federalist Papers." During the last decade, this scientific field has been developed substantially, taking advantage of research advances in areas such as machine learning, information retrieval, and natural language processing. The plethora of available electronic texts (e.g., e-mail messages, online forum messages, blogs, source code, etc.) indicates a wide variety of applications of this technology, provided it is able to handle short and noisy text from multiple candidate authors. In this article, a survey of recent advances of the automated approaches to attributing authorship is presented, examining their characteristics for both text representation and text classification. The focus of this survey is on computational requirements and settings rather than on linguistic or literary issues. We also discuss evaluation methodologies and criteria for authorship attribution studies and list open questions that will attract future work in this area.

Introduction

The main idea behind statistically or computationally supported authorship attribution is that by measuring some textual features, we can distinguish between texts written by different authors. The first attempts to quantify the writing style go back to the 19th century, with the pioneering study of Mendenhall (1887) on the plays of Shakespeare, followed by statistical studies in the first half of the 20th century by Yule (1938, 1944) and Zipf (1932). Later, the detailed study by Mosteller and Wallace (1964) on the authorship of "The Federalist Papers" (a series of 146 political essays written by John Jay, Alexander Hamilton, and James Madison, 12 of which claimed by both Hamilton and Madison) was undoubtedly the most influential work in authorship attribution. Their method was based on Bayesian statistical analysis of the frequencies of a small set of common words (e.g., "and," "to," etc.) and produced significant discrimination results between the candidate authors.

Essentially, the work of Mosteller and Wallace (1964) initiated *nontraditional authorship attribution* studies, as opposed to traditional human-expert-based methods. Since then and until the late 1990s, research in authorship attribution was dominated by attempts to define features for quantifying writing style, a line of research known as "stylometry" (Holmes, 1994, 1998). Hence, a great variety of measures, including sentence length, word length, word frequencies, character frequencies, and vocabulary richness functions, had been proposed. Rudman (1998) estimated that nearly 1,000 different measures had been proposed by the late 1990s. The authorship attribution methodologies proposed during that period were computer-assisted rather than computer-based, meaning that the aim was rarely at developing a fully automated system. In certain cases, there were methods that achieved impressive preliminary results and made many people think that the solution of this problem was too close. The most characteristic example is the CUSUM (or QSUM) technique (Morton & Michaelson, 1990) that gained publicity and was accepted in courts as expert evidence; however, the research community heavily criticized it and considered it generally unreliable (Holmes & Tweedie, 1995). Actually, the main problem of that early period was the lack of objective evaluation of the proposed methods. In most of the cases, the testing ground was literary works of unknown or disputed authorship (e.g., the Federalist Papers case), so the estimation of attribution accuracy was not even possible. The main methodological limitations of that period concerning the evaluation procedure were the following:

- The textual data were too long (usually including entire books) and probably not stylistically homogeneous.
- The number of candidate authors was too small (usually two or three).
- The evaluation corpora were not controlled for topic.
- The evaluation of the proposed methods was mainly intuitive (usually based on subjective visual inspection of scatterplots).
- The comparison of different methods was difficult due to lack of suitable benchmark data.

Received March 14, 2008; revised October 6, 2008; accepted October 6, 2008

 • Published online 16 December 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21001

Since the late 1990s, things have changed in authorship attribution studies. The vast amount of electronic texts available through Internet media (e-mail messages, blogs, online forums, etc.) have increased the need for efficiently handling this information. This fact had a significant impact in scientific areas such as information retrieval, machine learning, and natural language processing (NLP). The development of these areas influenced authorship attribution technology as described:

- Information retrieval research developed efficient techniques for representing and classifying large volumes of text.
- Powerful machine learning algorithms became available to handle multidimensional and sparse data, allowing more expressive representations. Moreover, standard evaluation methodologies have been established to compare different approaches on the same benchmark data.
- NLP research developed tools able to analyze text efficiently and provided new forms of measures for representing the style (e.g., syntax-based features).

More importantly, the plethora of available electronic texts revealed the potential of *authorship analysis* in various applications (Madigan, Genkin, Lewis, Argamon, Fradkin, & Ye, 2005) in diverse areas including intelligence (e.g., attribution of messages or proclamations to known terrorists, linking different messages by authorship) (Abbasi & Chen, 2005), criminal law (e.g., identifying writers of harassing messages, verifying the authenticity of suicide notes) and civil law (e.g., copyright disputes) (Chaski, 2005; Grant, 2007), and computer forensics (e.g., identifying the authors of source code of malicious software) (Frantzeskou, Stamatatos, Gritzalis, & Katsikas, 2006) in addition to the traditional application to literary research (e.g., attributing anonymous or disputed literary works to known authors) (Burrows, 2002; Hoover, 2004a). Hence, (roughly) the last decade can be viewed as a new era of authorship analysis technology, this time dominated by efforts to develop practical applications dealing with real-world texts (e.g., e-mail messages, blogs, online forum messages, source code, etc.) rather than solving disputed literary questions. Emphasis has now been given to the objective evaluation of the proposed methods as well as the comparison of different methods based on common benchmark corpora (Juola, 2004). In addition, factors playing a crucial role in the accuracy of the produced models are examined, such as the training text size (Hirst & Feiguina, 2007; Marton, Wu, & Hellerstein, 2005), the number of candidate authors (Koppel, Schler, Argamon, & Messeri, 2006), and the distribution of training texts over the candidate authors (Stamatatos, 2008).

In the typical authorship attribution problem, a text of unknown authorship is assigned to one candidate author, given a set of candidate authors for whom text samples of undisputed authorship are available. From a machine learning point of view, this can be viewed as a multiclass, single-label text-categorization task (Sebastiani, 2002). This task also is called *authorship* (or *author*) *identification* usually by researchers with a background in computer science. Several studies have focused exclusively on authorship attribution (Keselj, Peng, Cercone, & Thomas, 2003; Stamatatos, Fakotakis, & Kokkinakis, 2001; Zheng, Li, Chen, & Huang, 2006) while others have used it as just another testing ground for text-categorization methodologies (Khmelev & Teahan, 2003a; Marton et al., 2005; Peng, Shuurmans, & Wang, 2004; Zhang & Lee, 2006). Beyond this problem, several other authorship analysis tasks can be defined, including the following:

- Author verification (i.e., to decide whether a given text was written by a certain author) (Koppel & Schler, 2004).
- Plagiarism detection (i.e., finding similarities between two texts) (Meyer zu Eissen, Stein, & Kulig, 2007; Stein & Meyer zu Eissen, 2007).
- Author profiling or characterization (i.e., extracting information about the age, education, sex, etc., of the author of a given text) (Koppel, Argamon, & Shimoni, 2002).
- Detection of stylistic inconsistencies (as may happen in collaborative writing) (Collins, Kaufer, Vlachos, Butler, & Ishizaki, 2004; Graham, Hirst, & Marthi, 2005).

This article presents a survey of the research advances in this area during roughly the last decade (Earlier work was excellently reviewed by Holmes, 1994, 1998) emphasizing computational requirements and settings rather than linguistic or literary issues. First, a comprehensive review of the approaches to quantify the writing style is presented. Then, we focus on the authorship-identification problem (as described earlier). We propose the distinction of attribution methodologies according to how they handle the training texts, individually or cumulatively (per author), and examine their strengths and weaknesses across several factors. We then discuss the evaluation criteria of authorship attribution methods. The conclusions drawn by this survey are summarized, and future work directions in open research issues are indicated.

Stylometric Features

Previous studies on authorship attribution have proposed taxonomies of features to quantify the writing style, the so-called *style markers*, under different labels and criteria (Holmes, 1994; Stamatatos, Fakotakis, & Kokkinakis, 2000; Zheng et al., 2006). The current review of text representation features for stylistic purposes is mainly focused on the computational requirements for measuring them. First, lexical and character features consider a text as a mere sequence of word-tokens or characters, respectively. Note that although lexical features are more complex than character features, we start with them for the sake of tradition. Then, syntactic and semantic features require deeper linguistic analysis while application-specific features can be defined only in certain text domains or languages. The basic feature categories and the required tools and resources for their measurement are shown in Table 1. Moreover, various feature selection and extraction methods to form the most appropriate feature set for a particular corpus are discussed.

TABLE 1. Types of stylometric features together with computational tools and resources required for their measurement (Brackets indicate optional tools.).

Features		Required tools and resources
Lexical	Token-based (word length, sentence length, etc.)	Tokenizer, [Sentence splitter]
	Vocabulary richness	Tokenizer
	Word frequencies	Tokenizer, [Stemmer, Lemmatizer]
	Word *n*-grams	Tokenizer
	Errors	Tokenizer, Orthographic spell checker
Character	Character types (letters, digits, etc.)	Character dictionary
	Character *n*-grams (fixed length)	–
	Character *n*-grams (variable length)	Feature selector
	Compression methods	Text compression tool
Syntactic	Part-of-speech (POS)	Tokenizer, Sentence splitter, POS tagger
	Chunks	Tokenizer, Sentence splitter, [POS tagger], Text chunker
	Sentence and phrase structure	Tokenizer, Sentence splitter, POS tagger, Text chunker, Partial parser
	Rewrite rules frequencies	Tokenizer, Sentence splitter, POS tagger, Text chunker, Full parser
	Errors	Tokenizer, Sentence splitter, Syntactic spell checker
Semantic	Synonyms	Tokenizer, [POS tagger], Thesaurus
	Semantic dependencies	Tokenizer, Sentence splitter, POS tagger, Text Chunker, Partial parser, Semantic parser
	Functional	Tokenizer, Sentence splitter, POS tagger, Specialized dictionaries
Application-specific	Structural	HTML parser, Specialized parsers
	Content-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries
	Language-specific	Tokenizer, [Stemmer, Lemmatizer], Specialized dictionaries

Lexical Features

A simple and natural way to view a text is as a sequence of tokens grouped into sentences, with each token corresponding to a word, number, or punctuation mark. The very first attempts to attribute authorship were based on simple measures such as sentence length counts and word length counts (Mendenhall, 1887). A significant advantage of such features is that they can be applied to any language and any corpus with no additional requirements except the availability of a tokenizer (i.e., a tool to segment text into tokens). However, for certain natural languages (e.g., Chinese), this is not a trivial task. For use of sentential information, a tool that detects sentence boundaries also should be available. In certain text domains with heavy use of abbreviations or acronyms (e.g., e-mail messages), this procedure may introduce considerable noise in the measures.

The *vocabulary richness* functions are attempts to quantify the diversity of the vocabulary of a text. Typical examples are the *type-token* ratio *V/N*, where *V* is the size of the vocabulary (unique tokens) and *N* is the total number of tokens of the text, and the number of *hapax legomena* (i.e., words occurring once) (de Vel, Anderson, Corney, & Mohay, 2001). Unfortunately, the vocabulary size depends heavily on text length (as the text length increases, the vocabulary also increases, quickly at the beginning and then more and more slowly). Various functions have been proposed to achieve stability over text length, including *K* (Yule, 1944), and *R* (Honore, 1979), with questionable results (Tweedie & Baayen, 1998). Hence, such measures are considered unreliable to be used alone.

The most straightforward approach to represent texts is by vectors of word frequencies. The vast majority of authorship attribution studies are (at least partially) based on lexical features to represent the style. This is also the traditional bag-of-words text representation followed by researchers in topic-based text classification (Sebastiani, 2002). That is, the text is considered as a set of words, each one having a frequency of occurrence disregarding contextual information. However, there is a significant difference in style-based text classification: The most common words (articles, prepositions, pronouns, etc.) are found to be among the best features to discriminate between authors (Argamon & Levitan, 2005; Burrows, 1987). Note that such words are usually excluded from the feature set of the topic-based text-classification methods since they do not carry any semantic information, and they are usually called "*function" words*. As a consequence, style-based text classification using lexical features requires much lower dimensionality in comparison to topic-based text classification. In other words, much less words are sufficient to perform authorship attribution (a few hundred words) in comparison to a thematic text categorization task (several thousand words). More importantly, function words are used in a largely unconscious manner by the authors, and they are topic-independent. Thus, they are able to capture pure stylistic choices of the authors across different topics.

The selection of the specific function words that will be used as features is usually based on arbitrary criteria and requires language-dependent expertise. Various sets of function words have been used for English, but limited information was provided about the way that they were selected: Abbasi and Chen (2005) reported a set of 150 function words; Argamon, Saric, and Stein (2003) used a set of 303 words; Zhao and Zobel (2005) used a set of 365 function

DOI: 10.1002/asi

words; 480 function words were proposed by Koppel and Schler (2003); another set of 675 words was reported by Argamon et al. (2007).

A simple and very successful method to define a lexical feature set for authorship attribution is to extract the most frequent words found in the available corpus (comprising all the texts of the candidate authors). Then, a decision has to be made about the amount of the frequent words that will be used as features. In the earlier studies, sets of at most 100 frequent words were considered adequate to represent the style of an author (Burrows, 1987, 1992). Another factor that affects the feature-set size is the classification algorithm that will be used since many algorithms overfit the training data when the dimensionality of the problem increases. However, the availability of powerful machine learning algorithms able to deal with thousands of features, such as support vector machines (SVM; Joachims, 1998), enabled researchers to increase the feature-set size of this method. Koppel, Schler, and Bonchek-Dokow (2007) used the 250 most frequent words while Stamatatos (2006a) extracted the 1,000 most frequent words. On a larger scale, Madigan et al. (2005) used all the words that appear at least twice in the corpus. Note that the first dozens of most frequent words of a corpus are usually dominated by closed-class words (articles, prepositions, etc.) After a few hundred words, open-class words (nouns, adjectives, verbs) are the majority. Hence, when the dimensionality of this representation method increases, some content-specific words also may be included in the feature set.

Despite the availability of a tokenizer, word-based features may require additional tools for their extraction. This would involve simple routines such as conversion to lowercase to more complex tools such as stemmers (Sanderson & Guenter, 2006), lemmatizers (Gamon, 2004; Tambouratzis et al., 2004), or detectors of common homographic forms (Burrows, 2002). Another procedure used by van Halteren (2007) is to transform words into an abstract form. For example, the Dutch word "waarmaken" is transformed to "#L#6+/L/ken," where the first "L" indicates low frequency, "6+" indicates the length of the token, the second "L" a lowercase token, and "ken" are its last three characters.

The bag-of-words approach provides a simple and efficient solution, but disregards word-order (i.e., contextual) information. For example, the phrases "take on," "the second take" and "take a bath" would just provide three occurrences of the word "take." To take advantage of contextual information, *word n-grams* (*n* contiguous words also known as word collocations) have been proposed as textual features (Coyotl-Morales, Villaseñor-Pineda, Montes-y-Gómez, & Rosso, 2006; Peng et al., 2004; Sanderson & Guenter, 2006). However, the classification accuracy achieved by word *n*-grams is not always better than individual word features (Coyotl-Morales et al., 2006; Sanderson & Guenter, 2006). The dimensionality of the problem following this approach increases considerably with *n* to account for all the possible combinations between words. Moreover, the representation produced by this approach is very sparse since most of the word combinations are not encountered in a given (especially short) text, making it very difficult to be handled effectively by a classification algorithm. Another problem with word *n*-grams is that it is quite possible to capture content-specific information rather than stylistic information (Gamon, 2004).

From another point of view, Koppel and Schler (2003) proposed various writing error measures to capture the idiosyncrasies of an author's style. To that end, they defined a set of spelling errors (e.g., letter omissions and insertions) and formatting errors (e.g., "all caps" words) and proposed a methodology to extract such measures automatically using a spell checker. Interestingly, human experts mainly use similar observations to attribute authorship; however, the availability of accurate spell checkers is still problematic for many natural languages.

Character Features

According to this family of measures, a text is viewed as a mere sequence of characters. That way, various character-level measures can be defined, including alphabetic characters count, digit characters count, uppercase and lowercase characters count, letter frequencies, punctuation marks count, and so on. (de Vel et al., 2001; Zheng et al., 2006). This type of information is easily available for any natural language and corpus, and it has been proven to be quite useful to quantify the writing style (Grieve, 2007).

A more elaborate, although still computationally simplistic, approach is to extract frequencies of *n*-grams on the character level. For instance, the character 4-grams of the beginning of this paragraph would be:[1] |A_mo|, |_mor|, |more|, |ore_|, |re_e|, and so on. This approach is able to capture nuances of style, including lexical information (e.g., |_in_|, |text|), hints of contextual information (e.g., |in_t|), use of punctuation and capitalization, and so on. Another advantage of this representation is its ability to be tolerant to noise. In cases where the texts in question are noisy, containing grammatical errors or making strange use of punctuation, as it usually happens in e-mail or in online forum messages, the character *n*-gram representation is not affected dramatically. For example, the words "simplistic" and "simpilstc" would produce many common character trigrams. On the other hand, these two words would be considered different in a lexically based representation. Note that in style-based text categorization, such errors could be considered personal traits of the author (Koppel & Schler, 2003). This information also is captured by character *n*-grams (e.g., in the uncommon trigrams |stc| and |tc_|). Finally, for oriental languages where the tokenization procedure is quite hard, character *n*-grams offer a suitable solution (Matsuura & Kanada, 2000). As can be seen in Table 1, the computational requirements of character *n*-gram features are minimal.

Note that as with words, the most frequent character *n*-grams are the most important features for stylistic purposes. The procedure of extracting the most frequent *n*-grams is

[1]The characters "|"and "_" are used to denote *n*-gram boundaries and a single space character, respectively.

DOI: 10.1002/asi

language-independent and requires no special tools; however, the dimensionality of this representation is considerably increased in comparison to the word-based approach (Stamatatos, 2006a, 2006b). This happens because character *n*-grams capture redundant information (e.g., |and_|, |_and|), and many character *n*-grams are needed to represent a single long word.

The application of this approach to authorship attribution has proven quite successful. Kjell (1994) first used character bigrams and trigrams to discriminate the Federalist Papers. Forsyth and Holmes (1996) found that bigrams and character *n*-grams of variable length performed better than lexical features in several text-classification tasks including authorship attribution. Peng, Shuurmans, Keselj, and Wang (2003), Keselj et al. (2003), and Stamatatos (2006b) reported very good results using character *n*-gram information. Moreover, one of the best performing algorithms in an authorship attribution competition organized in 2004 also was based on a character *n*-gram representation (Juola, 2004, 2006). Likewise, a recent comparison of different lexical and character features on the same evaluation corpora (Grieve, 2007) showed that character *n*-grams were the most effective measures (outperformed in the specific experiments only by a combination of frequent words and punctuation marks).

An important issue of the character *n*-gram approach is the definition of *n*; that is, how long the strings should be. A large *n* would better capture lexical and contextual information, but it also would better capture thematic information. Furthermore, a large *n* would increase substantially the dimensionality of the representation (producing hundreds of thousands of features). On the other hand, a small *n* (i.e., 2 or 3) would be able to represent subword (syllable-like) information, but it would not be adequate for representing the contextual information. Note that the selection of the best *n* value is a language-dependent procedure since certain natural languages (e.g., Greek, German) tend to have long words in comparison to English. Therefore, probably a larger *n* value would be more appropriate for such languages in comparison to the optimal *n* value for English. The problem of defining a fixed value for *n* can be avoided by the extraction of *n*-grams of variable length (Forsyth & Holmes, 1996; Houvardas & Stamatatos, 2006). Sanderson and Guenter (2006) described the use of several sequence kernels based on character *n*-grams of variable length, and the best results for short English texts were achieved when examining sequences of up to 4-grams. Moreover, various Markov models of variable order have been proposed for handling character-level information (Khmelev & Teahan, 2003a; Marton et al., 2005). Finally, Zhang and Lee (2006) constructed a suffix tree representing all possible character *n*-grams of variable length and then extracted groups of character *n*-grams as features.

A quite particular case of using character information is the *compression-based approaches* (Benedetto, Caglioti, & Loreto, 2002; Khmelev & Teahan, 2003a; Marton et al., 2005). The main idea is to use the compression model acquired from one text to compress another text, usually based on off-the-shelf compression programs. If the two texts are written by the same author, the resulting bit-wise size of the compressed file will be relatively low. Such methods do not require a concrete representation of text, and the classification algorithm incorporates the quantification of textual properties. However, the compression models that describe the characteristics of the texts are usually based on repetitions of character sequences, and as a result, they can capture subword and contextual information. In that sense, they can be considered as character-based methods.

Syntactic Features

A more elaborate text-representation method is to employ syntactic information. The idea is that authors tend to unconsciously use similar syntactic patterns. Therefore, syntactic information is considered more a reliable authorial fingerprint in comparison to lexical information. Moreover, the success of function words in representing style indicates the usefulness of syntactic information since they are usually encountered in certain syntactic structures. On the other hand, this type of information requires robust and accurate NLP tools able to perform syntactic analysis of texts. This fact means that the syntactic measure extraction is a language-dependent procedure since it relies on the availability of a parser able to analyze a particular natural language with relatively high accuracy. Moreover, such features will produce noisy datasets due to unavoidable errors made by the parser.

Baayen, van Halteren, and Tweedie (1996) were the first to use syntactic information measures for authorship attribution. Based on a syntactically annotated English corpus, comprising a semiautomatically produced full parse tree of each sentence, they were able to extract rewrite rule frequencies. Each rewrite rule expresses a part of syntactic analysis; for instance, the following rewrite rule:

$$\mathrm{A : PP \rightarrow P : PREP + PC : NP}$$

means that an adverbial prepositional phrase is constituted by a preposition followed by a noun phrase as a prepositional complement. That detailed information describes both what the syntactic class of each word is and how the words are combined to form phrases or other structures. Experimental results have shown that this type of measure performs better than do vocabulary richness and lexical measures. On the other hand, it required a sophisticated and accurate fully automated parser able to provide a detailed syntactic analysis of English sentences. Similarly, Gamon (2004) used the output of a syntactic parser to measure rewrite rule frequencies as described earlier. Although the proposed syntactic features alone performed worse than did lexical features, the combination of the two improved the results.

Another attempt to exploit syntactic information was proposed by Stamatatos et al. (2000, 2001). They used an NLP tool able to detect sentence and chunk (i.e., phrases) boundaries in unrestricted Modern Greek text. For example, the first sentence of this paragraph would be analyzed as following:

NP[*Another attempt*] VP[*to exploit*] NP[*syntactic information*] VP[*was proposed*] PP[*by Stamatatos et al. (2000)*].

where NP, VP, and PP stand for noun phrase, verb phrase, and prepositional phrase, respectively. This type of information is simpler than that used by Baayen et al. (1996) since there is no structural analysis within the phrases or the combination of phrases into higher structures, but it could be extracted automatically with relatively high accuracy. The extracted measures referred to noun phrase counts, verb phrase counts, length of noun phrases, length of verb phrases, and so on. More interesting, another type of relevant information also was used which Stamatatos et al. (2000, 2001) called *analysis-level* measures. This type of information is relevant to the particular architecture of that specific NLP tool. In more detail, that particular tool analyzed the text in several steps. The first steps analyzed simple cases while the last steps attempted to combine the outcome of the first steps to produce more complex results. The analysis-level measures proposed for that tool had to do with the percentage of text each step achieved to analyze. Essentially this is as a type of indirect syntactic information, and it is tool-specific in addition to language-specific. However, it is a practical solution for extracting syntactic measures from unrestricted text given the availability of a suitable NLP tool.

In a similar framework, tools that perform partial parsing can be used to provide syntactic features of varying complexity (Hirst & Feiguina, 2007; Luyckx & Daelemans, 2005; Uzuner & Katz, 2005). Partial parsing is between text chunking and full parsing, and can handle unrestricted text with relatively high accuracy. Hirst and Feiguina (2007) transformed the output of a partial parser into an ordered stream of syntactic labels. For instance, the analysis of the phrase "a simple example" would produce the following stream of labels:

NX DT JJ NN

in words, a noun phrase consisting of a determiner, an adjective, and a noun. Then, they extracted measures of bigram frequencies from that stream to represent contextual syntactic information, and found this information useful to discriminate the authors of very short texts (i.e., ~200 words long).

An even simpler approach is to use just a part-of-speech (POS) tagger, a tool that assigns a tag of morpho-syntactic information to each word-token based on contextual information. Usually, POS taggers perform quite accurately in unrestricted text, and several researchers have used POS tag frequencies or POS tag *n*-gram frequencies to represent style (Argamon-Engelson, Koppel, & Avneri, 1998; Diederich, Kindermann, Leopold, & Paass, 2003; Gamon, 2004; Koppel & Schler, 2003; Kukushkina, Polikarpov, & Khmelev, 2001; Zhao & Zobel, 2007). However, POS tag information provides only a hint of the structural analysis of sentences since it is not clear how the words are combined to form phrases or how the phrases are combined into higher level structures.

Perhaps the most extensive use of syntactic information was described by van Halteren (2007). He applied a morpho-syntactic tagger and a syntactic analyzer for Dutch to a corpus of student essays and extracted unigrams, bigrams, and trigrams of morpho-syntactic tags as well as various *n*-gram measures from the application of rewrite rules. As a result, a huge set of about 900,000 features was constructed to quantify syntactic information!

Another interesting use of syntactic information was proposed by Koppel and Schler (2003) based on syntactic errors such as sentence fragments, run-on sentences, mismatched tense, and so on. To detect such information, they used a commercial spell checker. As with orthographic errors, this type of information is similar to that used by human experts when they attempt to attribute authorship. Unfortunately, the spell checkers are not very accurate, and Koppel and Schler reported they had to modify the output of that tool to improve the error detection results.

Finally, Karlgren and Eriksson (2007) described a preliminary model based on two syntactic features: adverbial expressions and occurrence of clauses within sentences. However, the quantification of these features is not the traditional relative frequency of occurrence within the text. They used sequence patterns aiming to describe the use of these features in consecutive sentences of the text. Essentially, this is an attempt to represent the distributional properties of the features in the text, a promising technique that can capture important stylistic properties of the author.

Semantic Features

It should be clear by now that the more detailed the text analysis required for extracting stylometric features, the less accurate (and the more noisy) the produced measures. NLP tools can be applied successfully to low-level tasks such as sentence splitting, POS tagging, text chunking, and partial parsing, so relevant features would be measured accurately and so the noise in the corresponding datasets remains low. On the other hand, more complicated tasks such as full syntactic parsing, semantic analysis, or pragmatic analysis cannot yet be handled adequately by current NLP technology for unrestricted text. As a result, very few attempts have been made to exploit high-level features for stylometric purposes.

Gamon (2004) used a tool able to produce semantic dependency graphs, but did not provide information about the accuracy of this tool. Two kinds of information were then extracted: binary semantic features and semantic modification relations. The former concerned number and person of nouns, tense and aspect of verbs, and so on. The latter described the syntactic and semantic relations between a node of the graph and its daughters (e.g., a nominal node with a nominal modifier indicating location). Reported results have shown that semantic information when combined with lexical and syntactic information improved the classification accuracy.

McCarthy, Lewis, Dufty, and McNamara (2006) described another approach to extract semantic measures. Based on WordNet (Fellbaum, 1998), they estimated information about synonyms and hypernyms of the words as well as the identification of causal verbs. Moreover, they applied *latent*

DOI: 10.1002/asi

semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) to lexical features to automatically detect semantic similarities between words. However, there was no detailed description of the features, and the evaluation procedure did not clarify the contribution of semantic information in the classification model.

Perhaps the most important method of exploiting semantic information so far was described by Argamon et al. (2007). Inspired by the theory of *Systemic Functional Grammar* (SFG; Halliday, 1994), they defined a set of *functional features* that associate certain words or phrases with semantic information. In more detail, in SFG, the "CONJUNCTION" scheme denotes how a given clause expands on some aspect of its preceding context. Types of expansion could be "ELABORATION" (exemplification or refocusing), "EXTENSION" (adding new information), or "ENHANCEMENT" (qualification). Certain words or phrases indicate certain modalities of the "CONJUNCTION" scheme. For example, the word "specifically" is used to identify a "CLARIFICATION" of an "ELABORATION" of a "CONJUNCTION" while the phrase "in other words" is used to identify an "APPOSITION" of an "ELABORATION" of a "CONJUNCTION." To detect such semantic information, they used a lexicon of words and phrases produced semiautomatically based on online thesauruses including WordNet. Each entry in the lexicon associated a word or phrase with a set of syntactic constraints (in the form of allowed POS tags) and a set of semantic properties. The set of functional measures, then, contained measures showing, for instance, how many "CONJUNCTIONs" were expanded to "ELABORATIONs" or how many "ELABORATIONs" were elaborated to "CLARIFICATIONs," and so on. However, no information was provided on the accuracy of those measures. Experiments of authorship identification on a corpus of English novels of the 19th century showed that the functional features can improve the classification results when combined with traditional function-word features.

Application-Specific Features

The previously described lexical, character, syntactic, or semantic features are application-independent since they can be extracted from any textual data given the availability of the appropriate NLP tools and resources required for their measurement. Beyond that, one can define application-specific measures to better represent the nuances of style in a given text domain. This section reviews the most important of these measures.

The application of the authorship attribution technology in domains such as e-mail messages and online-forum messages revealed the possibility to define *structural* measures to quantify the authorial style. Structural measures include the use of greetings and farewells in the messages, types of signatures, use of indentation, paragraph length, and so on (de Vel et al., 2001; Li, Zheng, & Chen, 2006; Teng, Lai, Ma, & Li, 2004; Zheng et al., 2006). Moreover, provided the texts in question are in HTML format, measures related to HTML tag distribution (de Vel et al., 2001), font-color counts, and font-size counts (Abbasi & Chen, 2005) also can be defined. Apparently, such features can be defined only in given text genres. Moreover, they are particularly important in very short texts where the stylistic properties of the textual content cannot be adequately represented using application-independent methods; however, accurate tools are required for their extraction. Zheng et al. (2006) reported that they had difficulties in accurately measuring their structural features.

In general, the style factor of a text is considered orthogonal to its topic. As a result, stylometric features attempt to avoid content-specific information to be more reliable in cross-topic texts. However, in cases in which all the available texts for all the candidate authors are on the same thematic area, carefully selected, content-based information may reveal some authorial choices. To better capture the properties of an author's style within a particular text domain, *content-specific keywords* can be used. In more detail, given that the texts in question deal with certain topics and are of the same genre, one can define certain words frequently used within that topic or that genre. For example, in the framework of the analysis of online messages from the newsgroup *misc.forsale.computers*, Zheng et al. (2006) defined content-specific keywords such as "deal," "sale," or "obo" (i.e., or best offer). The difference of these measures and the function words discussed earlier is that they carry semantic information and are characteristic of particular topics and genres. It remains unclear how to select such features for a given text domain.

Other types of application-specific features can be defined only for certain natural languages. For example, Tambouratzis et al. (2004) attempted to take advantage of the diglossia phenomenon in Modern Greek and proposed a set of verbal endings which are usually found in "Katharevousa" and "Dimotiki;" that is, roughly the formal and informal variations of Modern Greek, respectively. Although such measures have to be defined manually, they can be very effective when dealing with certain text genres.

Feature Selection and Extraction

The feature sets used in authorship attribution studies often combine many types of features. In addition, some feature types such as lexical and character features can considerably increase the dimensionality of the feature set. In such cases, *feature selection* algorithms can be applied to reduce the dimensionality of the representation (Forman, 2003). That way, the classification algorithm is helped to avoid overfitting on the training data.

In general, the features selected by these methods are examined individually on the basis of discriminating the authors of a given corpus (Forman, 2003); however, certain features that seem irrelevant when examined independently may be useful in combination with other variables. In this case, the performance of certain classification algorithms that can handle high-dimensional feature sets (e.g., SVM) might be diminished by reducing the dimensionality (Brank,

DOI: 10.1002/asi

Grobelnik, Milic-Frayling, & Mladenic, 2002). To avoid this problem, feature subset selection algorithms examine the discriminatory power of feature subsets (Kohavi & John, 1997). For example, Li et al. (2006) described the use of a genetic algorithm to reduce an initial set of 270 features to an optimal subset for the specific training corpus comprising 134 features. As a result, the classification performance improved from 97.85% (when the full set was used) to 99.01% (when the optimal set was used).

However, the best features may strongly correlate with one of the authors due to content-specific rather than to stylistic choices (e.g., imagine we have two authors for whom there are articles about politics for the one and articles about sports for the other). In other words, the features identified by a feature selection algorithm may be too corpus-dependent with questionable general use. On the other hand, in the seminal work of Mosteller and Wallace (1964), the features were carefully selected based on their universal properties to avoid dependency on a specific training corpus.

The most important criterion for selecting features in authorship attribution tasks is their frequency. In general, the more frequent a feature, the more stylistic variation it captures. Forsyth and Holmes (1996) were the first to compare (character n-gram) feature sets selected by frequency with feature sets selected by distinctiveness; they found the latter more accurate. However, they restricted the size of the extracted feature sets to a relatively very low level (96 features). Houvardas and Stamatatos (2006) proposed an approach for extracting character n-grams of variable length using frequency information only. The comparison of this method with *information gain*, a well-known feature selection algorithm individually examining the discriminatory power of features (Forman, 2003), showed that the frequency-based feature set was more accurate for feature sets comprising up to 4,000 features. Similarly, Koppel, Akiva, and Dagan (2006) presented experiments comparing frequency-based feature selection with *odds ratio*, another typical feature selection algorithm using discrimination information (Forman, 2003). More importantly, the frequency information they used was not extracted from the training corpus. Again, the frequency-based feature subsets performed better than did those produced by odds ratio. When the frequency information was combined with odds ratio, the results were further improved.

Koppel, Akiva, and Dagan (2006) also proposed an additional important criterion for feature selection in authorship attribution, the *instability* of features. Given a number of variations of the same text, all with the same meaning, the features that remain practically unchanged in all texts are considered stable. In other words, stability may be viewed as the availability of "synonyms" for certain language characteristics. For example, words such as "and" and "the" are very stable since there are no alternatives for them. On the other hand, words such as "benefit" or "over" are relatively unstable since they can be replaced by "gain" and "above," respectively, in certain situations. Therefore, unstable features are more likely to indicate stylistic choices of the author. To produce the required variations of the same text, Koppel, Akiva, and Dagan (2006) used several machine translation programs to generate translations from English to another language and then back to English. Although the quality of the produced texts was obviously low, this procedure was fully automated. Let $\{d_1, d_2, \ldots, d_n\}$ be a set of texts and $\{d_i^j, d_i^2, \ldots, d_i^m\}$ a set of variations of the ith text, all with roughly the same meaning. For a stylometric feature c, let c_i^j be the value of feature c in the jth variation of the ith text and $k_i = \sum_j c_i^j$. Then, the instability of c is defined by:

$$IN_c = 1 - \frac{\sum_i \left[k_i \log k_i - \sum_j c_i^j \log c_i^j \right]}{\sum_i k_i * \log m}$$

Experiments showed that features selected by the instability criterion alone were not as effective as features selected by frequency; however, when the frequency and the instability criteria were combined, the results were much better.

Another approach to reduce dimensionality is via *feature extraction* (Sebastiani, 2002). Here, a new set of "synthetic" features is produced by combining the initial set of features. The most traditional feature-extraction technique in authorship attribution studies is the *principal components analysis*, which provides linear combinations of the initial features. The two most important principal components can, then, be used to represent the texts in a two-dimensional space (Binongo, 2003; Burrows, 1987, 1992). However, the reduction of the dimensionality to a single feature (or a couple of features) has the consequence of losing too much variation information. Therefore, such simple features are generally unreliable to be used alone. Another, more elaborate feature-extraction method was described by Zhang and Lee (2006). They first built a suffix tree representing all the possible character n-grams of the texts and then extracted groups of character n-grams according to frequency and redundancy criteria. The resulting *key-substring-groups*, each one accumulating many character n-grams, were the new features. The application of this method to authorship attribution and other text-classification tasks provided promising results.

Attribution Methods

In every authorship-identification problem, there is a set of candidate authors, a set of text samples of known authorship covering all the candidate authors (training corpus), and a set of text samples of unknown authorship (test corpus); each one of them should be attributed to a candidate author. In this survey, we distinguish the authorship attribution approaches according to whether they treat each training text individually or cumulatively (per author). In more detail, some approaches concatenate all the available training texts per author in one big file and extract a cumulative representation of that author's style (usually called the author's *profile*) from this concatenated text. That is, the differences between texts written by the same author are disregarded. We examine such

DOI: 10.1002/asi

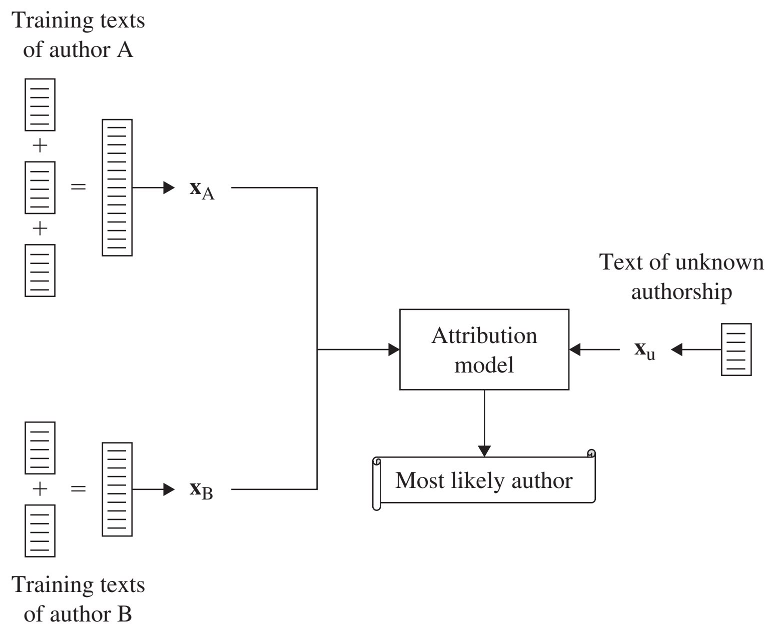

FIG. 1. Typical architecture of profile-based approaches.

profile-based approaches[2] first since early work in authorship attribution has followed this practice (Mosteller & Wallace, 1964). On the other hand, another family of approaches requires multiple training text samples per author to develop an accurate attribution model. That is, each training text is individually represented as a separate instance of authorial style. Such *instance-based approaches* [3] are described in the next section, followed by hybrid approaches attempting to combine characteristics of profile-based and instance-based methods. We then compare these two basic approaches and discuss their strengths and weaknesses across several factors.

Note that in this review, the distinction between profile-based and instance-based approaches is considered the most basic property of the attribution methods since it largely determines the philosophy of each method (e.g., a classification model of generative or discriminative nature). Moreover, it shows the kind of writing style that each method attempts to handle: a general style for each author or a separate style of each individual document.

Profile-Based Approaches

One way to handle the available training texts per author is to concatenate them in one single text file. This large file is used to extract the properties of the author's style. An unseen text is, then, compared with each author file, and the most likely author is estimated based on a distance measure. It should be stressed that there is no separate representation of each text sample but only one representation of a large file per author. As a result, the differences between the training texts by the same author are disregarded. Moreover, the stylometric measures extracted from the concatenated file may be quite different in comparison to each of the original training texts. A typical architecture of a profile-based approach is depicted in Figure 1. Note that x denotes a vector of text representation features. Hence, x_A is the profile of Author A, and x_u is the profile of the unseen text.

The profile-based approaches have a very simple training process. Actually, the training phase just comprises the extraction of profiles for the candidate authors. Then, the attribution model is usually based on a distance function that computes the differences of the profile of an unseen text and the profile of each author. Let $PR(x)$ be the profile of text x and $d[PR(x),PR(y)]$ the distance between the profile of text x and the profile of text y. Then, the most likely author of an unseen text x is given by:

$$author(x) = \underset{a \in \mathrm{A}}{\arg\min} d(PR(x),\ PR(x_a))$$

where A is the set of candidate authors and x_a is the concatenation of all training texts for author a. In the following, we first describe how this approach can be realized by using probabilistic and compression models, and then the CNG method and its variants are discussed.

Probabilistic models. One of the earliest approaches to author identification that is still used in many modern studies employs the use of probabilistic models (Clement & Sharp, 2003; Madigan et al., 2005; Mosteller & Wallace, 1964; Peng et al., 2004; Sanderson & Guenter, 2006; Zhao & Zobel, 2005). Such methods attempt to maximize the probability $P(x|a)$ for a text x to belong to a candidate author a. Then, the attribution model seeks the author that maximizes the

[2] Note that this term should not be confused with *author profiling* methods (e.g., extracting information about the author gender, age, etc.) (Koppel et al., 2002).

[3] Note that this term should not be confused with *instance-based learning* methods (Mitchell, 1997).

DOI: 10.1002/asi

following similarity metric:

$$author(x) = \arg\max_{a \in A} \log_2 \frac{P(x|a)}{P(x|\bar{a})}$$

where the conditional probabilities are estimated by the concatenation x_a of all available training texts of author a and the concatenation of all the rest of the texts, respectively. Variants of such probabilistic classifiers (e.g., naïve Bayes) have been studied in detail in the framework of topic-based text categorization (Sebastiani, 2002). An extension of the naïve Bayes algorithm augmented with statistical language models was proposed by Peng et al. (2004) and achieved high performance in authorship attribution experiments. In comparison to standard naïve Bayes classifiers, the approach of Peng et al. allows local Markov chain dependencies in the observed variables to capture contextual information. Moreover, sophisticated smoothing techniques from statistical language modeling can be applied to this method (The best results for authorship attribution were obtained using absolute smoothing.) More interesting, this method can be applied to both character and word sequences. Actually, Peng et al. achieved their best results for authorship attribution using word-level models for a specific corpus; however, this was not confirmed in other corpora as well.

Compression models. The most successful of the compression-based approaches follows the profile-based methodology (Khmelev & Teahan, 2003a; Kukushkina et al., 2001; Marton et al., 2005). Such methods do not produce a concrete vector representation of the author's profile. Therefore, we can consider $PR(x) = x$. Initially, all the available texts for the ith author are first concatenated to form a large file x_a, and a compression algorithm is called to produce a compressed file $C(x_a)$. Then, the unseen text x is added to each text x_a, and the compression algorithm is called again for each $C(x_a + x)$. The difference in bit-wise size of the compressed files $d(x, x_a) = C(x_a + x) - C(x_a)$ indicates the similarity of the unseen text with each candidate author. Essentially, this difference calculates the cross-entropy between the two texts. Several off-the-shelf compression algorithms have been tested with this approach, including RAR, LZW, GZIP, BZIP2, 7ZIP, and so on; in most of the cases, RAR was found to be the most accurate (Khmelev & Teahan, 2003a; Kukushkina et al., 2001; Marton et al., 2005).

Note that the *prediction by partial matching* (PPM) algorithm (Teahan & Harper, 2003) used by RAR to compress text files works practically the same as the method of Peng et al. (2004); however, there is a significant difference with the previously described probabilistic method. In particular, in the method of Khmelev and Teahan (2003a), the models describing x_a were adaptive with respect to x; that is, the compression algorithm was applied to the text $x_a + x$, so the compression model was modified as it processed the unseen text. In the method of Peng et al. (2004), the models describing x_a were static; that is, the n-gram Markov models were extracted from text x_a and then applied to unseen text x, and no modification of the models was allowed in the latter phase. For that reason, the application of the probabilistic method to the classification of an unseen text is faster in comparison to this compression-based approach. Another advantage of the language modeling approach is that it can be applied to both character and word sequences while the PPM compression models are applied only to character sequences.

Common n*-grams and variants.* A profile-based method of particular interest, the *common n-grams* (CNG) approach, was described by Keselj et al. (2003). This method used a concrete representation of the author's profile. In particular, the profile $PR(x)$ of a text x was composed by the L most frequent character n-grams of that text. The following distance is, then, used to estimate the similarity between two texts, x and y:

$$d(PR(x), PR(y)) = \sum_{g \in P(x) \cup P(y)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2$$

where g is a character n-gram while $f_x(g)$ and $f_y(g)$ are the relative frequencies of occurrence of that n-gram in texts x and y, respectively. In words, this measure computes the dissimilarity between two profiles by calculating the relative difference between their common n-grams. All the n-grams of the two profiles that are not common contribute a constant value to the distance. The CNG method has two important parameters that should be tuned: the profile size L and the character n-gram length n; that is, how many and how long strings constitute the profile. Keselj et al. (2003) reported their best results for $1{,}000 \leq L \leq 5{,}000$ and $3 \leq n \leq 5$. This basic approach has been applied successfully to various authorship-identification experiments including the authorship attribution competition organized in 2004 (Juola, 2004, 2006).

An important problem in authorship attribution tasks arises when the distribution of the training corpus over the candidate authors is uneven. For example, it is not unusual, especially in forensic applications, to have multiple training texts for some candidate authors and very few training texts for other authors. Moreover, the length of these samples may not allow their segmentation into multiple parts to enrich the training instances of certain authors. In machine learning terms, this constitutes the *class imbalance* problem. The majority of the authorship attribution approaches studies present experiments based on balanced training sets (i.e., an equal amount of training text samples for each candidate author), so it is not possible to estimate their accuracy under class-imbalance conditions. Only a few studies have taken this factor into account (Marton et al., 2005; Stamatatos, 2007).

The CNG distance function performs well when the training corpus is relatively balanced, but it fails in imbalanced cases where at least one author's profile is shorter than L (Stamatatos, 2007). For example, if we use $L = 4{,}000$ and $n = 3$, and the available training texts of a certain candidate author are too short, then the total amount of 3-grams that can be extracted from that authors' texts may be less than 4,000. The distance function favors that author because the union of the profile of the unseen text and the profile of that author will

DOI: 10.1002/asi

result in significant less n-grams, so the distance between the unseen text and that author would be estimated as quite low in comparison to the other authors. To overcome that problem, Frantzeskou et al. (2006) proposed a different and simpler distance, called *simplified profile intersection* (SPI), which simply counts the amount of common n-grams of the two profiles and disregarding the rest. The application of this measure to author identification of source code provided better results than did the original CNG distance. Note that in contrast to CNG distance, SPI is a similarity measure, meaning that the most likely author is the author with the highest SPI value. A problem of that distance can arise when all the candidate authors except one have very short texts. Then, SPI metric will favor the author with long texts since many more common n-grams will be detected in their texts and an unseen text.

Another variation of the CNG dissimilarity function was proposed by Stamatatos (2007):

$$d(PR(x), PR(y), PR(N)) = \sum_{g \in P(x)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2 \cdot \left(\frac{2(f_x(g) - f_N(g))}{f_x(g) + f_N(g)} \right)^2$$

where N is the *corpus norm* (i.e., the concatenation of all available texts of all the candidate authors) and $f_N(g)$ is the relative frequency of occurrence of the n-gram g in the corpus norm. Note that this function is not symmetric as the original CNG function. In particular, the first argument $PR(x)$ is the profile of the unseen text, and the second argument is an author profile. That way, only the n-grams of the unseen-text's profile contribute to the calculated sum. As a result, the problems described earlier with imbalanced corpora are significantly reduced since the distance between the unseen text and the candidate authors is always based on the same amount of terms. Moreover, each term is multiplied by the relative distance of the specific n-gram frequency from the corpus norm. Hence, the more an n-gram deviates from its "normal" frequency, the more it contributes to the distance. On the other hand, if the frequency of an n-gram is found exactly the same as its "normal" frequency, it does not contribute at all at the distance value (i.e., the norm factor is 0). Experiments reported by Stamatatos (2007) showed that this distance function can better handle cases where limited and imbalanced corpora were available for training. Furthermore, it was quite stable with respect to the parameter L; however, in cases where enough training texts were available, the original CNG method produced better results.

Instance-Based Approaches

The majority of the modern authorship-identification approaches considers each training text sample as a unit that contributes separately to the attribution model. In other words, each text sample of known authorship is an instance of the problem in question. A typical architecture of such an instance-based approach is shown in Figure 2. In detail, each text sample of the training corpus is represented by a vector of attributes (x) following methods described earlier, and a classification algorithm is trained using the set of instances of known authorship (training set) to develop an attribution model. Then, this model will be able to estimate the true author of an unseen text.

Note that such classification algorithms require multiple training instances per class for extracting a reliable model. Therefore, according to instance-based approaches, in case we have only one, but a quite long, training text for a particular candidate author (e.g., an entire book), this should be segmented into multiple parts, probably of equal length. From another point of view, when there are multiple training

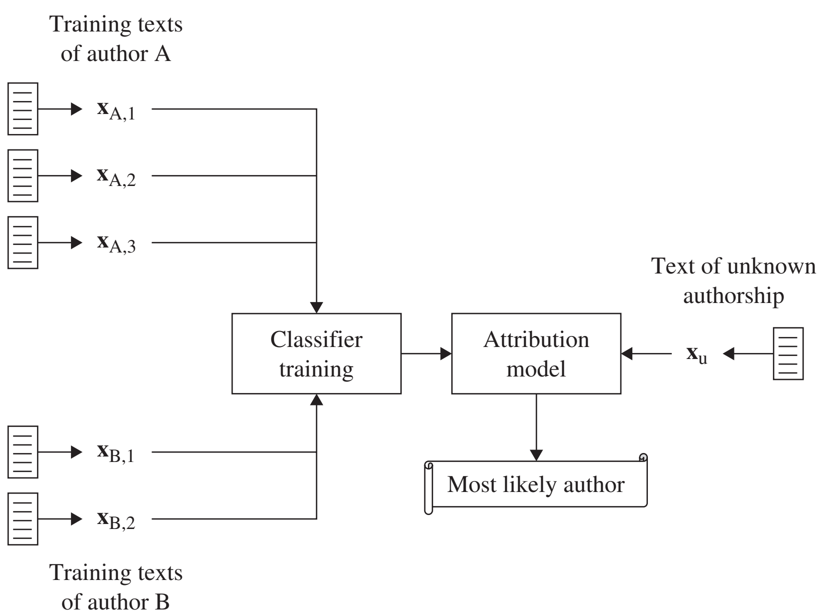

FIG. 2. Typical architecture of instance-based approaches.

DOI: 10.1002/asi

text samples of variable length per author, the training text instance length should be normalized. To that end, the training texts per author are segmented to equal-sized samples (Sanderson & Guenter, 2006). In all these cases, the text samples should be long enough so that the text representation features can adequately represent their style. Various lengths of text samples have been reported in the literature. Sanderson and Guenter (2006) produced chunks of 500 characters. Koppel et al. (2007) segmented the training texts into chunks of about 500 words. Hirst and Feiguina (2007) conducted experiments with text blocks of varying length (i.e., 200, 500, and 1000 words) and reported significantly reduced accuracy as the text-block length decreases. Therefore, the choice of the training instance text sample is not a trivial process and directly affects the performance of the attribution model.

In what follows, we first describe the vector space models that comprise the majority of the instance-based approaches. Then, various similarity-based and meta-learning models are discussed.

Vector space models. Given that the training texts are represented in a multivariate form, we can consider each text as a vector in a multivariate space. Then, a variety of powerful statistical and machine learning algorithms can be used to build a classification model, including discriminant analysis (Chaski, 2005; Stamatatos et al., 2000; Tambouratzis et al., 2004), SVM (de Vel et al., 2001; Diederich et al, 2003; Li et al., 2006; Sanderson & Guenter, 2006; Teng et al., 2004), decision trees (Uzuner & Katz, 2005; Zhao & Zobel, 2005; Zheng et al., 2006), neural networks (Khosmood & Levinson, 2006; Matthews & Merriam, 1993; Merriam & Matthews, 1994; Tweedie, Singh, & Holmes, 1996; Zheng et al., 2006), genetic algorithms (Holmes & Forsyth, 1995), memory-based learners (Luyckx & Daelemans, 2005), classifier ensemble methods (Stamatatos, 2006a), and so on.

Such algorithms have been studied thoroughly in the framework of (mostly topic-based) text-categorization research (Sebastiani, 2002). Therefore, we will not discuss them further. Note, though, that some of these algorithms can effectively handle high-dimensional, noisy, and sparse data, allowing more expressive representations of texts. For example, an SVM model is able to avoid overfitting problems even when several thousands of features are used and is considered one of the best solutions of current technology (Li et al., 2006; Stamatatos, 2008).

The effectiveness of vector space models is usually diminished by the presence of the class-imbalance problem. Recently, Stamatatos (2008) proposed an approach to deal with this problem in the framework of vector space instance-based approaches. In more detail, the training set can be rebalanced by segmenting the text samples of a particular author according to the size of their class (i.e., the length of all texts of that author). That way, many short text samples can be produced for minority authors (i.e., the authors for whom only a few training texts were available) while less, but longer, texts can be produced for the majority authors (i.e., the authors for whom multiple training texts were available). Moreover, text resampling (i.e., using some text parts more than once) could be used to increase the training set of the minority authors.

Similarity-based models. The main idea of similarity-based models is the calculation of pairwise similarity measures between the unseen text and all the training texts, and then the estimation of the most likely author based on a nearest-neighbor algorithm. The most notable approach of this category was proposed by Burrows (2002) under the name "Delta." First, this method calculates the z distributions of a set of function words (originally, the 150 most frequent words). Then, for each document, the deviation of each word frequency from the norm is calculated in terms of z score, roughly indicating whether it is used more (positive z score) or less (negative z score) times than the average. Finally, the Delta measure indicating the difference between a set of (training) texts written by the same author and an unknown text is the mean of the absolute differences between the z scores for the entire function word set in the training texts and the corresponding z scores of the unknown text. The smaller the Delta measure, the greater stylistic similarity between the unknown text and the candidate author. This method was mainly evaluated on literary texts (English poems and novels), producing remarkable results (Burrows, 2002; Hoover, 2004a). It has been demonstrated that it is a very effective attribution method for texts of at least 1,500 words. For shorter texts, the accuracy drops according to length. However, even for quite-short texts, the correct author was usually included in the first five positions of the ranked authors, which provides a means for reducing the set of candidate authors.

A theoretical understanding of the operation of Delta has been described by Argamon (2008). In more detail, he showed that Delta can be viewed as an axis-weighted form of nearest-neighbor classification, where the unknown text is assigned to the nearest category instead of the nearest training text. It also was shown that the distance ranking of candidate authors produced by Delta is equivalent to probability ranking under the assumption that word frequencies follow a Laplace distribution. This view indicates many extensions and generalizations of Delta (e.g., using Gaussian distributions of word frequencies in place of Laplace distributions, etc.). A detailed study of variations of Burrows' (2002) Delta was presented by Hoover (2004a). He found that by using larger sets of frequent words ($>$500), the accuracy of the method was increasing. The performance also was improved when the personal pronouns and words for which a single text supplied most of their occurrences were eliminated. Some variations of the Delta score itself also were examined, but no significant improvement over the original method was achieved (Hoover, 2004b).

Another similarity-based approach utilizing text-compression models to estimate the difference between texts has been described by Benedetto et al. (2002). The training phase of this method merely comprises the compression of each training text in separate files using an off-the-shelf

DOI: 10.1002/asi

algorithm (GZIP). For estimating the author of an unseen text, this text is concatenated to each training text file, and then each resulting file is compressed by the same algorithm. Let $C(x)$ be the bitwise size of the compression of file x while $x + y$ is the concatenation of text files x and y. Then, the difference $C(x + y) - C(x)$ indicates the similarity of a training text x with the unseen text y. Finally, a 1-nearest-neighbor decision estimates the most likely author.

This method was strongly criticized by several researchers (Goodman, 2002; Khmelev & Teahan, 2003b), indicating many weaknesses. First, it is too slow since it has to call the compression algorithm so many times (as many as the training texts). Note that in the corresponding profile-based approach of Khmelev and Teahan (2003a), the compression algorithm is called as many times as the candidate authors. Hence, the running time will be significantly lower for the profile-based, compression-based method. Moreover, various authorship-identification experiments have shown that the compression-based approach following the profile-based technique usually outperforms the corresponding instance-based method (Marton et al., 2005). An important factor that contributes to this direction is that 1-nearest-neighbor approach is sensitive to noise; however, this problem could be faced by using the k-nearest-neighbors and a majority vote or a weighted vote scheme. Finally, GZIP is a dictionary-based compression algorithm and uses a sliding window of 32,000 to build the dictionary. This means that if a training text is long enough, the beginning of that document will be ignored when GZIP attempts to compress the concatenation of that file with the unseen text. Comparative experiments on various corpora have shown that the RAR compression algorithm outperforms GZIP in most of the cases (Marton et al., 2005).

An alternative distance measure for the compression-based approach was proposed by Cilibrasi and Vitanyi (2005). Based on the notion of the Kolmogorov complexity, they defined the *normalized compression distance* (NCD) between two texts x and y as follows:

$$NCD(x, y) = \frac{C(x + y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

Cilibrasi and Vitanyi used this distance metric and the BZIP2 compression algorithm to cluster literary works in Russian by four different authors and reported excellent results. They even attempted to cluster the corresponding English translations of those texts with relatively good results.

Meta-learning models. In addition to the general-purpose classification algorithms described earlier, one can design more complex algorithms specifically designed for authorship attribution. To this end, an existing classification algorithm may serve as a tool in a meta-learning scheme. The most interesting approach of this kind is the *unmasking* method proposed by Koppel et al. (2007) originally for author verification. The main difference with the typical instance-based approach shown in Figure 2 is that in the unmasking method, the training phase does not exist. For each unseen text, an SVM classifier is built to discriminate it from the training texts of each candidate author. Thus, for n candidate authors, Koppel et al. (2007) built n classifiers for each unseen text. Then, in an iterative procedure, they removed a predefined amount of the most important features for each classifier and measured the drop in accuracy. At the beginning, all the classifiers had more or less the same very high accuracy. After a few iterations, the accuracy of the classifier that discriminates between the unseen text and the true author would be too low while the accuracy of the other classifiers would remain relatively high. This happens because the differences between the unseen text and the other authors are manifold, so by removing a few features, the accuracy is not affected dramatically. Koppel et al. (2007) proposed a simple meta-learning method to learn to automatically discriminate the true author and reported very good results. This method seems more appropriate when the unknown texts are long enough since each unknown text has to be segmented in multiple parts to train the SVM classifiers. This was confirmed by Sanderson and Guenter (2006), who examined the unmasking method in long texts (entire books) with high accuracy results, but in short texts of newspaper articles, the results were not encouraging.

Hybrid Approaches

A method that borrows some elements from both profile-based and instance-based approaches was described by van Halteren (2007). In more detail, all the training text samples were represented separately, as it happens with the instance-based approaches. However, the representation vectors for the texts of each author were feature-wisely averaged and produced a single profile vector for each author, as happens with the profile-based approaches. The distance of the profile of an unseen text from the profile of each author was then calculated by a weighted feature-wise function. Three weighting parameters had to be tuned empirically: one for the difference between the feature values of the unseen text profile and the author profile, one for the feature importance for the unseen text, and one for the feature importance for the particular author. A similar hybrid approach also was used by Grieve (2007).

Comparison

Table 2 shows the results of comparing profile-based and instance-based approaches across several factors. As already noted, the main difference is the representation of training texts. The former produces one cumulative representation for all training texts per author while the latter produces individual representations for each training text. In certain cases, this is an important advantage of profile-based methods. First, when only short texts are available for training (e.g., e-mail messages, online forum messages), their concatenation may produce a more reliable representation in

TABLE 2. Comparison of profile-based and instance-based approaches.

	Profile-based approaches	Instance-based approaches
Text representation	One cumulative representation for all the training texts per author	Each training text is represented individually. Text segmentation may be required.
Stylometric features	Difficult to combine different features. Some (text-level) features are not suitable.	Different features can be combined easily.
Classification	Generative (e.g., Bayesian) models, Similarity-based methods	Discriminative models, Powerful machine learning algorithms (e.g., SVM), similarity-based methods
Training time cost	Low	Relatively high (low for compression-based methods)
Running time cost	Low (relatively high for compression-based methods)	Low (very high for compression-based methods)
Class imbalance	Depends on the length of training texts	Depends mainly on the amount of training texts

comparison to individual representations of short texts. Furthermore, when only one long text (or a few long texts) is available for one author, instance-based approaches require its segmentation to multiple parts.

On the other hand, instance-based approaches take advantage of powerful machine learning algorithms able to handle high-dimensional, noisy, and sparse data (e.g., SVM). Moreover, it is easy to combine different kinds of stylometric features in an expressive representation. This is more difficult in profile-based approaches that are based on generative (e.g., Bayesian) models or similarity-based methods, and they usually can handle homogeneous feature sets (e.g., function words, character *n*-grams, etc.) An exception was described by van Halteren (2007), although this is not a pure profile-based method. In addition, several stylometric features defined on the text level (e.g., use of greetings and signatures) cannot be easily used by profile-based approaches since the profile attempts to represent the general properties of the author's style rather than the properties of a typical text sample by that author.

Another main difference is the existence of the training phase in the instance-based approaches, with the exception of compression-based models (Benedetto et al., 2002). The training phase of profile-based approaches is relatively simple, comprising just the extraction of measures from training texts. In both cases, the running time cost is low again, with the exception of compression-based methods. The running time cost of instance-based compression methods is analogous to the number of training texts while the running time cost for the corresponding profile-based approaches is analogous to the number of the candidate authors (Marton et al., 2005).

In instance-based approaches, class imbalance depends on the amount of training texts per author. In addition, the text length of training texts may produce class-imbalance conditions when long texts are segmented into many parts. On the other hand, the class-imbalance problem in profile-based approaches depends only on text length. Hence, we may have two candidate authors with exactly the same amount of training text samples; however, the first author's texts are short while the other author's texts are long. This means that the concatenation of the training texts per author will produce two files that differ significantly in text length.

Evaluation

The seminal study of Mosteller and Wallace (1964) was about the disputed authorship of the Federalist Papers. This case offered a well-defined set of candidate authors, sets of known authorship for all the candidate authors, and a set of texts of disputed authorship. Moreover, all the texts were of the same genre and about the same thematic area. Hence, it was considered the ideal testing ground for early authorship attribution studies as well as the first fully automated approaches (Holmes & Forsyth, 1995; Tweedie et al., 1996). It also has been used in some modern studies (Marton et al., 2005; Teahan & Harper, 2003). Although appealing, this case has a number of important weaknesses. More specifically, the set of candidate authors is too small, the texts are relatively long, and the disputed texts may be the result of collaborative writing of the candidate authors (Collins et al., 2004).

A significant part of modern authorship attribution studies have applied the proposed techniques to literary works of undisputed authorship, including American and English literature (Argamon et al., 2007; Koppel et al., 2007; McCarthy et al., 2006; Uzuner & Katz, 2005; Zhao & Zobel, 2007), Russian literature (Cilibrasi & Vitanyi, 2005; Kukushkina et al., 2001), Italian literature (Benedetto et al., 2002), and so on. A case of particular difficulty concerns the separation of works of the Bronte sisters, Charlotte and Anna, since they share the same characteristics (Burrows, 1992; Hirst & Feiguina, 2007; Koppel, Akiva, & Dagan, 2006). The main problem when using literary works for evaluating author-identification methods is the text length of training and test texts (usually entire books). Certain methods can work effectively in long texts, but not so well on short or very short texts (Hirst & Feiguina, 2007; Sanderson & Guenter, 2006). To this end, poems provide a more reliable testing ground (Burrows, 2002).

Beyond literature, several evaluation corpora for authorship attribution studies have been built, covering certain text domains such as online newspaper articles (Diederich et al.,

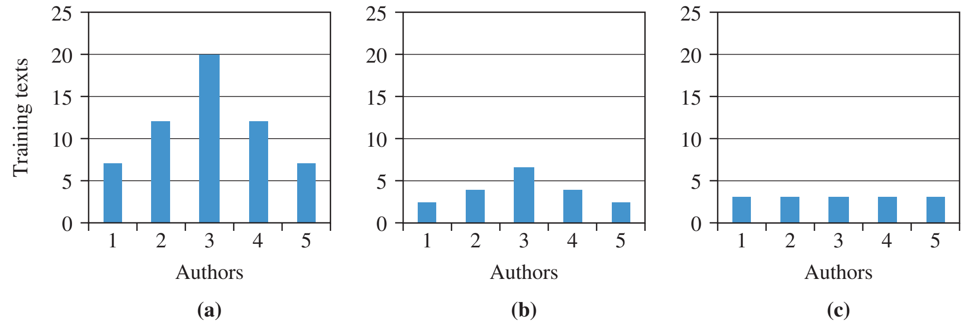

FIG. 3. Different distributions of training and test texts over 5 candidate authors: (a) an imbalanced distribution of training texts, (b) an imbalanced distribution of test texts imitating the distribution of training texts, and (c) a balanced distribution of test texts.

2003; Luyckx & Daelemans, 2005; Sanderson & Guenter, 2006; Stamatatos et al., 2000), e-mail messages (de Vel et al., 2001; Koppel & Schler, 2003), online forum messages (Abbasi & Chen, 2005; Argamon et al., 2003; Zheng et al., 2006), newswire stories (Khmelev & Teahan, 2003a; Zhao & Zobel, 2005), blogs (Koppel, Schler, Argamon, & Messeri, 2006), and so on. Alternatively, corpora built for other purposes also have been used in the framework of authorship attribution studies, including parts of the Reuters-21578 corpus (Marton et al., 2005; Teahan & Harper, 2003), the Reuters Corpus Volume 1 (Khmelev & Teahan, 2003a; Madigan et al., 2005; Stamatatos, 2007), and the TREC corpus (Zhao & Zobel, 2005) that were initially built for evaluating thematic text-categorization tasks. Such corpora offer the possibility to test methods on cases with many candidate authors and relatively short texts.

Following the practice of other text-categorization tasks, some of these corpora have been used as a benchmark to compare different methods on exactly the same training and test corpus (Sebastiani, 2002). One such corpus comprising Modern Greek newspaper articles was introduced by Stamatatos et al. (2000, 2001) and was used later by Keselj et al. (2003), Peng et al., (2003), Peng et al. (2004), Zhang and Lee (2006), and Stamatatos (2006a, 2006b). Moreover, in the framework of an ad hoc authorship attribution competition organized in 2004, various corpora have been collected,[4] covering several natural languages (English, French, Latin, Dutch, and Serbian-Slavonic) and difficulty levels (Juola, 2004, 2006).

Any good evaluation corpus for authorship attribution should be controlled for genre and topic. That way, authorship would be the most important discriminatory factor between the texts. While genre can be easily controlled, the topic factor reveals difficulties. Ideally, all the texts of the training corpus should be on exactly the same topic for all the candidate authors. A few such corpora have been reported. Chaski (2001) described a writing-sample database comprising texts of 92 people on 10 common subjects (e.g., a letter of apology to your best friend, a letter to your insurance company, etc.). Clement and Sharp (2003) reported a corpus of movie reviews comprising 5 authors who review the same five movies. Another corpus comprising various genres was described by Baayen, van Halteren, Neijt, and Tweedie (2002), which also was used by Juola and Baayen (2005) and van Halteren (2007). It consisted of 72 texts by 8 students of Dutch literature on specific topics covering three genres. In more detail, each student was asked to write three argumentative nonfiction texts on specific topics (e.g. the unification of Europe), three descriptive nonfiction texts (e.g., about soccer), and three fiction texts (e.g., a murder story in the university). Other factors that should be controlled in the ideal evaluation corpus include age, education level, nationality, and so on to reduce the likelihood that the stylistic choices of a given author are characteristic of a broad group of people rather than strictly personal. In addition, all the texts per author should be written in the same period to avoid style changes over time (Can & Patton, 2004).

A thorough evaluation of an authorship attribution method would require the examination of its performance under various conditions. The most important evaluation parameters are the following:

- Training corpus size, in terms of both the amount and length of training texts.
- Test corpus size (in terms of text length of the unseen texts).
- Number of candidate authors.
- Distribution of the training corpus over the authors (balanced or imbalanced).

In the case of an imbalanced training corpus, an application-dependent methodology should be followed to form the most appropriate test corpus. One option is the distribution of test corpus over the candidate authors to imitate the corresponding distribution of the training corpus (Khmelev & Teahan, 2003a; Madigan et al., 2005). Examples of such training and test corpora are shown in Figures 3a and 3b, respectively. Consequently, a model that learns to guess the author of an unseen text taking into account the amount of available training texts per author would achieve good performance on such a test corpus. This practice is usually followed in the evaluation of topic-based, text-categorization methods. However, in the framework of authorship attribution, it seems suitable for applications that aim at only filtering texts according to authorial information. Another option would be the balanced distribution of the test corpus over the candidate authors (Stamatatos, 2007, 2008). Examples of such training and test corpora are shown in Figures 3a and 3c,

[4]http://www.mathcs.duq.edu/~juola/authorship_contest.html

DOI: 10.1002/asi

respectively. As a result, a model that learns to guess the author of an unseen text taking into account the amount of available training texts per author will achieve low performance on that balanced test corpus. This approach seems appropriate for the majority of authorship attribution applications, including intelligence, criminal law, or forensics, where the availability of texts of known authorship should not increase the likelihood of certain candidate authors. That is, in most cases, it just happens to have many (or few) texts of known authorship for some authors. Note also that an imbalanced training corpus is the most likely real-world scenario in a given authorship-identification application.

Another important factor in the evaluation of an authorship attribution approach is its ability to handle more than one natural language. Recall that many features used to represent the stylistic properties are language-dependent. In general, methods using character features can be transferred easily to other languages. A few studies have presented experiments in multiple natural languages. Peng et al. (2003) evaluated their method in three languages—English, Greek, and Chinese—while Keselj et al. (2003) used English and Greek corpora. In addition, Abassi and Chen (2005) and Stamatatos (2008) used English and Arabic corpora while Li et al. (2006) evaluated their approach in English and Chinese texts. Beyond language, an attribution method should be tested on a variety of text genres (e.g., newspaper articles, blogs, literature, etc.) to reveal its ability to handle unrestricted text or just certain text domains.

Discussion

Rudman (1998, p. 351) criticized the state of authorship attribution studies, saying that

> Non-traditional authorship attribution studies—those employing the computer, statistics, and stylistics—have had enough time to pass through any "shake-down" phase and enter one marked by solid, scientific, and steadily progressing studies. But after 30 years and 300 publications, they have not.

It is a fact that much of redundancy and methodological irregularities still remain in this field partly due to its interdisciplinary nature. However, during the last decade, significant steps have been taken towards the right direction. From a marginal scientific area dealing only with famous cases of disputed or unknown authorship of literary works, authorship attribution now provides robust methods able to handle real-world texts with relatively high-accuracy results. Fully automated approaches can give reliable solutions in a number of applications of the Internet era (e.g., analysis of e-mail messages, blogs, online forum messages, etc.) To this end, this area has taken advantage of recent advances in information retrieval, machine learning, and NLP.

Authorship attribution can be viewed as a typical text-categorization task, and actually, several researchers develop general text-categorization techniques and evaluate them on authorship attribution together with other tasks such as topic identification, language identification, genre detection, and so on (Benedetto et al., 2002; Teahan & Harper, 2003; Marton et al., 2005; Peng et al., 2004; Zhang & Lee, 2006). However, there are some important characteristics that distinguish authorship attribution from other text-categorization tasks. First, in style-based text categorization, the most significant features are the most frequent ones (Houvardas & Stamatatos, 2006; Koppel, Akiva, & Dagan, 2006) while in topic-based text categorization, the best features should be selected based on their discriminatory power (Forman, 2003). Second, in authorship attribution tasks, especially in forensic applications, there is extremely limited training text material while in most text-categorization problems (e.g., topic identification, genre detection), there is plenty of both labeled and unlabeled (that can be manually labeled) data. Hence, it is crucial for the attribution methods to be robust with a limited amount of short texts. Moreover, in most cases, the distribution of training texts over the candidate authors is imbalanced. In such cases, the evaluation of authorship attribution methods should not follow the practice of other text-categorization tasks; that is, the test corpus follows the distribution of training corpus (discussed earlier). On the contrary, the test corpus should be balanced. This is the most appropriate evaluation method for most of the authorship attribution applications (e.g., intelligence, criminal law, forensics, etc.). Note that this does not necessarily stand for other style-based, text-categorization tasks such as genre detection.

Several crucial questions remain open for the authorship attribution problem. Perhaps the most important issue is the text length: How long should a text be so that we can adequately capture its stylistic properties? Various studies have reported promising results dealing with short texts (<1,000 words) (Hirst & Feguina, 2007; Sanderson & Guenter, 2006); however, it is not yet possible to define such a text-length threshold. Moreover, it is not yet clear whether other factors (beyond text length) also affect this process. For example, let *a* and *b* be two texts of 100 words and 1,000 words, respectively. A given authorship attribution tool can easily identify the author of *a*, but not the author of *b*. What are the properties of *a* that make it an easy case, and what makes *b* so difficult, albeit much longer, than *a*? On the other hand, what are the minimum requirements in training text we need to be able to identify the author of a given text?

Another important question is how to discriminate between the three basic factors: authorship, genre, and topic. Are there specific stylometric features that can capture only stylistic, and specifically authorial, information? Several features described earlier have claimed to capture only stylistic information (e.g., function words); however, the application of stylometric features to topic-identification tasks has revealed the potential of these features to indicate content information as well (Clement & Sharp, 2003; Mikros & Argiri, 2007). It seems that low-level features such as character *n*-grams are very successful for representing texts for stylistic purposes (Grieve, 2007; Keselj et al., 2003; Peng et al., 2003; Stamatatos, 2006b). Recall that the compression-based techniques operate also on the character level; however, these features unavoidably capture thematic

DOI: 10.1002/asi

information as well. Is it the combination of stylistic and thematic information that makes them so powerful discriminators?

More elaborate features, capturing syntactic or semantic information, are not yet able to represent adequately the stylistic choices of texts. Hence, they can be used only as a complement in other, more powerful features coming from the lexical or the character level. Perhaps the noise introduced by the NLP tools in the process of their extraction is the crucial factor for their failure. It remains to be seen whether NLP technology can provide even more accurate and reliable tools to be used for stylometric purposes. Moreover, distributional features (Karlgren & Eriksson, 2007) should be thoroughly examined since they can represent detailed sequential patterns of authorial style rather than mere frequencies of occurrence.

The accuracy of current authorship attribution technology depends mainly on the number of candidate authors, the size of texts, and the amount of training texts. However, this technology is not yet reliable enough to meet the court standards in forensic cases. An important obstacle is that it is not yet possible to explain the differences between the authors' style. It is possible to estimate the significance of certain (usually character or lexical) features for specific authors. But what we need is a higher level abstract description of the authorial style. Moreover, in the framework of forensic applications, the open-set classification setting is the most suitable (i.e., the true author is not necessarily included in the set of candidate authors). Most of the authorship attribution studies have considered the closed-set case (i.e., the true author should be one of the candidate authors). Additionally, in the open-set case, apart from measuring the accuracy of the decisions of the attribution model, special attention must be paid to the confidence of those decisions (i.e., how sure it is that the selected author is the true author of the text). Another line of research that has not been adequately examined so far is the development of robust attribution techniques that can be trained on texts from one genre and applied to texts of another genre by the same authors. This is quite useful, especially for forensic applications. For instance, it is possible to have blog postings for training, a harassing e-mail message for test or business letters for training, and a suicide note for test (Juola, 2007).

A significant advance of the authorship attribution technology during the last years was the adoption of objective evaluation criteria and the comparison of different methodologies using the same benchmark corpora, following the practice of thematic text categorization. A crucial issue is to increase the available benchmark corpora so that they cover many natural languages and text domains. It also is very important for the evaluation corpora to offer control over genre, topic, and demographic criteria. To that end, it would be extremely useful to establish periodic events including competitions of authorship attribution methods (Juola, 2004). Such competitions should comprise multiple tasks that cover a variety of problems in the style of Text Retrieval Conferences.[5] This is the fastest way to develop authorship attribution research and provide commercial applications.

[5]http://trec.nist.gov/

Acknowledgement

The author wishes to thank the anonymous JASIST reviewers for their valuable and insightful comments.

References

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

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

Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing.

Argamon, S., Saric, M., & Stein, S. (2003). Style mining of electronic messages for multiple authorship discrimination: First results. In Proceedings of the 9th ACM SIGKDD (pp. 475–480). New York: ACM Press.

Argamon, S., Whitelaw, C., Chase, P., Hota, S.R., Garg, N., & Levitan, S. (2007). Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology, 58(6), 802–822.

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

Baayen, R., van Halteren, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. In Proceedings of JADT 2002: Sixth International Conference on Textual Data Statistical Analysis (pp. 29–37).

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

Benedetto, D., Caglioti, E., & Loreto, V. (2002). Language trees and zipping. Physical Review Letters, 88(4), 048702.

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

Brank, J., Grobelnik, M., Milic-Frayling, N., & Mladenic, D. (2002). Interaction of feature selection methods and linear classification models. In Proceedings of the ICML-02 Workshop on Text Learning.

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

Burrows, J.F. (2002). "Delta:" A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3), 267–287.

Can, F., & Patton, J.M. (2004). Change of writing style with time. Computers and the Humanities, 38, 61–82.

Chaski, C.E. (2001). Empirical evaluations of language-based author identification techniques. Forensic Linguistics, 8(1), 1–65.

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

Cilibrasi, R., & Vitanyi, P.M.B. (2005). Clustering by compression. IEEE Transactions on Information Theory, 51(4), 1523–1545.

Clement, R., & Sharp, D. (2003). Ngram and Bayesian classification of documents for topic and authorship. Literary and Linguistic Computing, 18(4), 423–447.

Collins, J., Kaufer, D., Vlachos, P., Butler, B., & Ishizaki, S. (2004). Detecting collaborations in text: Comparing the authors' rhetorical language choices in the Federalist Papers. Computers and the Humanities, 38, 15–36.

Coyotl-Morales, R.M., Villaseñor-Pineda, L., Montes-y-Gòmez, M., & Rosso, P. (2006). Authorship attribution using word sequences. In Proceedings of the 11th Iberoamerican Congress on Pattern Recognition (pp. 844–853). Berlin, Germany: Springer.

Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

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

Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship attribution with support vector machines. Applied Intelligence, 19(1/2), 109–123.

Fellbaum, C. (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

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

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

Frantzeskou, G., Stamatatos, E., Gritzalis, S., & Katsikas, S. (2006). Effective identification of source code authors using byte-level information. In Proceedings of the 28th International Conference on Software Engineering (pp. 893–896). New York: ACM Press.

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

Goodman, J. (2002). Extended comment on language trees and zipping. Retrieved December 2, 2008, from http://arxiv.org/abs/cond-mat/0202383

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

Grant, T.D. (2007). Quantifying evidence for forensic authorship analysis. International Journal of Speech Language and the Law, 14(1), 1–25.

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

Halliday, M.A.K. (1994). Introduction to functional grammar (2nd ed.). London: 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.I. (1994). Authorship attribution. Computers and the Humanities, 28, 87–106.

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

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

Holmes, D.I., & Tweedie, F.J. (1995). Forensic stylometry: A review of the cusum controversy. In Revue Informatique et Statistique dans les Sciences Humaines (pp. 19–47). Liege, Belgium: University of Liege.

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

Hoover, D. (2004a). Testing Burrows' 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). Berlin, Germany: 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 (pp. 137–142). Berlin, Germany: Springer.

Juola, P. (2004). Ad-hoc authorship attribution competition. In Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (pp. 175–176).

Juola, P. (2006). Authorship attribution for electronic documents. In M. Olivier & S. Shenoi (Eds.), Advances in digital forensics II (pp. 119–130). Boston: Springer.

Juola, P. (2007). Future trends in authorship attribution. In P. Craiger & S. Shenoi (Eds.), Advances in digital forensics III (pp. 119–132). Boston: Springer.

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

Karlgren, J., & Eriksson, G. (2007). Authors, genre, and linguistic convention. In Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection (pp. 23–28).

Keselj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In Proceedings of the Pacific Association for Computational Linguistics (pp. 255–264).

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

Khmelev, D.V., & Teahan, W.J. (2003b). Comment: "Language trees and zipping." Physical Review Letters, 90, 089803.

Khosmood, F., & Levinson, R. (2006). Toward unification of source attribution processes and techniques. In Proceedings of the 5th International Conference on Machine Learning and Cybernetics (pp. 4551–4556). Washington, DC: IEEE.

Kjell, B. (1994). Discrimination of authorship using visualization. Information Processing and Management, 30(1), 141–150.

Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1–2), 273–324.

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.R. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401–412.

Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis (pp. 69–72).

Koppel, M., & Schler, J. (2004). Authorship verification as a one-class classification problem. In Proceedings of the 21st International Conference on Machine Learning (p. 62). New York: ACM Press.

Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship attribution with thousands of candidate authors. In Proceedings of the 29th ACM SIGIR (pp. 659–660). New York: ACM Press.

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

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.

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

Luyckx, K., & Daelemans, W. (2005). Shallow text analysis and machine learning for authorship attribution. In Proceedings of the 15th meeting of Computational Linguistics in the Netherlands (pp. 149–160). Utrecht, Netherlands: LOT.

Madigan, D., Genkin, A., Lewis, D., Argamon, S., Fradkin, D., & Ye, L. (2005). Author identification on the large scale. In Proceedings of CSNA-05.

Marton, Y., Wu, N., & Hellerstein, L. (2005). On compression-based text classification. In Proceedings of the European Conference on Information Retrieval (pp. 300–314). Berlin, Germany: Springer.

Matsuura, T., & Kanada, Y. (2000). Extraction of authors' characteristics from Japanese modern sentences via n-gram distribution. In Proceedings of the 3rd International Conference on Discovery Science (pp. 315–319). Berlin, Germany: Springer.

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.

DOI: 10.1002/asi

McCarthy, P.M., Lewis, G.A., Dufty, D.F., & McNamara, D.S. (2006). Analyzing writing styles with coh-metrix. In Proceedings of the Florida Artificial Intelligence Research Society International Conference (pp. 764–769).

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

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), 1–6.

Meyer zu Eissen, S., Stein, B., & Kulig, M. (2007). Plagiarism detection without reference collections. Advances in data analysis (pp. 359–366). Berlin, Germany: Springer.

Mikros, G., & Argiri, E. (2007). Investigating topic influence in authorship attribution. In Proceedings of the International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (pp. 29–35).

Mitchell, T. (1997). Machine learning. New York: McGraw-Hill.

Morton, A.Q., & Michaelson, S. (1990). The qsum plot. Technical Report CSR-3-90, University of Edinburgh, UK.

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

Peng, F., Shuurmans, D., Keselj, V., & Wang, S. (2003). Language independent authorship attribution using character level language models. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (pp. 267–274). Morristown, NJ: Association for Computational Linguistics.

Peng, F., Shuurmans, D., & Wang, S. (2004). Augmenting naive Bayes classifiers with statistical language models. Information Retrieval Journal, 7(1), 317–345.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities, 31, 351–365.

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

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

Stamatatos, E. (2006a). Authorship attribution based on feature set subspacing ensembles. International Journal on Artificial Intelligence Tools, 15(5), 823–838.

Stamatatos, E. (2006b). Ensemble-based author identification using character n-grams. In Proceedings of the 3rd International Workshop on Text-Based Information Retrieval (pp. 41–46).

Stamatatos, E. (2007). Author identification using imbalanced and limited training texts. In Proceedings of the 4th International Workshop on Text-Based Information Retrieval (pp. 237–241).

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(2), 193–214.

Stein, B., & Meyer zu Eissen, S. (2007). Intrinsic plagiarism analysis with meta learning. In Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection (pp. 45–50).

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G., & Tambouratzis, D. (2004). Discriminating the registers and styles in the Modern Greek language—Part 2: Extending the feature vector to optimize author discrimination. Literary and Linguistic Computing, 19(2), 221–242.

Teahan, W., & Harper, D. (2003). Using compression-based language models for text categorization. In W.B. Croft & J. Lafferty (Eds.), Language modeling and information retrieval (pp. 141–165). Berlin, Germany: Springer.

Teng, G., Lai, M., Ma, J., & Li, Y. (2004). E-mail authorship mining based on SVM for computer forensics. In Proceedings of the International Conference on Machine Learning and Cybernetics, 2 (pp. 1204–1207). Washington, DC: IEEE.

Tweedie, F., & Baayen, R. (1998). How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities, 32(5), 323–352.

Tweedie, F., Singh, S., & Holmes, D. (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. In Proceedings of the 2nd International Joint Conference on Natural Language Processing (pp. 969–980). Berlin, Germany: Springer.

van Halteren, H. (2007). Author verification by linguistic profiling: An exploration of the parameter space. ACM Transactions on Speech and Language Processing, 4(1), 1–17.

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 University Press.

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

Zhao, Y., & Zobel, J. (2005). Effective and scalable authorship attribution using function words. In Proceedings of the 2nd Asia Information Retrieval Symposium. Berlin, Germany: Springer.

Zhao, Y., & Zobel, J. (2007). Searching with style: Authorship attribution in classic literature. In Proceedings of the 30th Australasian Computer Science Conference (pp. 59–68). New York: ACM Press.

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 of Information Science and Technology, 57(3), 378–393.

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

DOI: 10.1002/asi