Authorship attribution in the wild

Article *in* Language Resources and Evaluation · March 2011

DOI: 10.1007/s10579-009-9111-2 · Source: DBLP

3 authors:

Moshe Koppel
Bar Ilan University
125 PUBLICATIONS **3,445** CITATIONS
SEE PROFILE

Jonathan Schler
Bar Ilan University
39 PUBLICATIONS **2,268** CITATIONS
SEE PROFILE

Shlomo Argamon
Illinois Institute of Technology
117 PUBLICATIONS **4,204** CITATIONS
SEE PROFILE

Lang Resources & Evaluation (2011) 45:83-94
DOI 10.1007/s10579-009-9111-2

Authorship attribution in the wild

Moshe Koppel · Jonathan Schler · Shlomo Argamon

Published online: 13 January 2010

Abstract Most previous work on authorship attribution has focused on the case in which we need to attribute an anonymous document to one of a small set of candidate authors. In this paper, we consider authorship attribution as found in the wild: the set of known candidates is extremely large (possibly many thousands) and might not even include the actual author. Moreover, the known texts and the anonymous texts might be of limited length. We show that even in these difficult cases, we can use similarity-based methods along with multiple randomized feature sets to achieve high precision. Moreover, we show the precise relationship between attribution precision and four parameters: the size of the candidate set, the quantity of known-text by the candidates, the length of the anonymous text and a certain robustness score associated with a attribution.

Keywords Authorship attribution · Open candidate set · Randomized feature set

1 Introduction

Authorship attribution has been much studied in recent years and several recent articles (Juola 2008; Koppel et al. 2008; Stamatatos 2009) survey the plethora of methods that have been applied to the problem. A significant fact that examination of the literature reveals is that nearly all research in the field only considers the

M. Koppel (✉) · J. Schler
Bar-Ilan University, Ramat-Gan, Israel
e-mail: moishk@gmail.com

J. Schler
e-mail: schler@gmail.com

S. Argamon
Illinois Institute of Technology, Chicago, IL, USA
e-mail: argamon@iit.edu

simplest version of the problem, in which we are given a long anonymous text that must be attributed to one of a small, closed set of candidate authors for each of whom we have (more or less extensive) writing samples.

Unfortunately, this "vanilla" version of the authorship attribution problem does not often arise in the real world. The situations typically encountered when performing authorship attribution in the wild are significantly more difficult than the vanilla version in one or more of three key ways:

1. There may be thousands of known candidate authors.
2. The author of the anonymous text might be none of the known candidates.
3. The "known-text" for each candidate and/or the anonymous text might be very limited.

These difficulties have very rarely been addressed by the research community (several important exceptions will be discussed below). In this paper, we will present a novel attribution method that attacks all three of these difficulties at once. We will show that even under these conditions, we can achieve very high attribution precision, while paying a tolerable price in recall. Moreover, we will measure the effect of three key factors—number of candidates, size of known-text by candidates, and size of the anonymous text—on the reliability of attributions output by our method.

2 Previous work

Broadly speaking, methods for automated authorship attribution can be divided into two main paradigms (Stamatatos 2009). In the *similarity-based* paradigm, some metric is used to measure the distance between two documents and an anonymous document is attributed to that author to whose known writing (considered collectively as a single document) it is most similar (Burrows 2002; Hoover 2003; Argamon 2008; Abbasi and Chen 2008). In the *machine-learning* paradigm, the known writings of each candidate author (considered as a set of distinct training documents) are used to construct a classifier that can then be used to classify anonymous documents (cf. Abbasi and Chen 2008; Zhao and Zobel 2005; Zheng et al. 2006; Koppel et al. 2008).

Research in the similarity-based paradigm has focused on the choice of features for document representation, on methods for dimensionality reduction (such as PCA) of the feature space, and on the choice of distance metric. Research in the machine-learning paradigm has focused on choice of features for document representation and on choice of learning algorithms.

Virtually all of this work has focused on problems with a small number of candidate authors. Recently, somewhat larger candidate sets have been considered by Madigan et al. (2005) (114 authors) and Luyckx and Daelemans (2008) (145 authors). Only Koppel et al. (2006) have considered candidate sets including thousands of authors. Both Koppel et al. (2006) and Luyckx and Daelemans (2008) observed that when there are very many candidate authors, similarity-based methods are more appropriate than machine-learning methods.

Similarly, almost all work in authorship attribution has focused on the case in which the candidate set is a closed set—the anonymous text is assumed to have been written by one of the known candidates. The more general case, in which the true author of an anonymous text might not be one of the known candidates, reduces to the binary *authorship verification* problem: determine if the given document was written by a specific author or not. The authorship verification problem has usually been considered in the context of plagiarism analysis (Clough 2000; Meyer zu Eissen et al. 2007). One general and effective method for authorship verification is unmasking, proposed by Koppel et al. (2007). The idea is that two texts are probably by different authors if the differences between them are robust to changes in the underlying feature set used to represent the documents. Koppel et al. (2007) used a machine-learning paradigm and measured differences using cross-validation accuracy. More generally, however, differences between documents can be more readily measured in the similarity-based paradigm. A document can be verified as having been written by a given author if the degree of similarity between the document and the author's known writing exceeds some threshold (van Halteren et al. 2005).

In this paper, we will consider the general authorship attribution problem, where candidate sets are simultaneously large and open. We will integrate the methods previously used to address the large candidate set problem and the open candidate set problem, as described above. The integrated method we propose is in fact simpler than both previous approaches and has the added advantage of being language-independent.

3 The corpus

We use a set of 10,000 blogs harvested in August 2004 from blogger.com. The corpus is balanced for gender within each of a number of age intervals. In addition, each individual blog is predominantly in English and contains sufficient text, as will be explained. For each blog, we choose 2000 words of known text and a *snippet*, consisting of the last 500 words of the blog, such that the posts from which the known text and the snippet are taken are disjoint. Our object will be to determine which—*if any*—of the authors of the known texts is the author of a given snippet.

Note that we will not necessarily use all the available data in each experiment. We will experiment using various subsets of the available text to determine the impact on attribution of the number of candidates, the quantity of known text for each candidate and the length of the anonymous snippet.

4 Naïve method

We begin by representing each text (both known texts and snippets) as a vector representing the respective frequencies of each *space-free character 4-gram*. For our purposes, a space-free character 4-gram is (a) a string of characters of length four that includes no spaces or (b) a string of four or fewer characters surrounded by spaces. (Note that the latter case corresponds roughly to words of length four or less,

but not exactly; any string of characters, including punctuation, numerals and sundry, is included.) In our corpus, there are just over 250,000 unique (but overlapping) space-free character 4-grams. (There would be considerably more such n-grams if we included those with spaces, but these are certainly adequate for our purposes.)

Character n-grams have been shown to be effective for authorship attribution (Keselj et al. 2003) and have the advantage of being measurable in any language without specialized background knowledge. We note that character n-gram statistics capture both aspects of document content and writing style. Although this distinction is often an important one in authorship studies, we do not dwell on it in this paper. For our purposes, we do not particularly care if attributions are based on style or content or both. We are content to show that our method works even using the most primitive and language-independent feature types imaginable.

Now, it is impractical to learn a single classifier for 10,000 classes; nor is it practical to learn 10,000 one-versus-all binary classifiers. Instead, we use a similarity-based method. Specifically, we use a common straightforward information retrieval method to assign an author to a given snippet. Using cosine similarity as a proximity measure, we simply return the author whose known writing (considered as a single vector of space-free character 4-gram frequencies) is most similar to the snippet vector (Salton and Buckley 1988). Testing this rather naïve method on 1,000 snippets selected at random from among the 10,000 authors, we find that 46% of the snippets are correctly assigned.

While this is perhaps surprisingly high, a precision of 46% is inadequate for most applications. To remedy this problem, we adopt the approach of (Koppel et al. 2006) which permits a response of *Don't Know* in cases where attribution is uncertain. The objective is to obtain high precision for those cases where an answer is given, while trying to offer an answer as often as possible. Our specific method for doing so will differ from that of (Koppel et al. 2006) in several ways. Unlike that work, our simpler method includes a natural parameter for recall-precision tradeoff, does not require a training corpus for learning meta-models, and does not use language-dependent lexical features.

5 Improved method

The key to our new approach is an insight initially confirmed by Koppel et al. (2007). The known text of a snippet's actual author is likely to be the text most similar to the snippet even as we vary the feature set that we use to represent the texts. Another author's text might happen to be the most similar for one or a few specific feature sets, but it is highly unlikely to be consistently so over many different feature sets.

This observation suggests using the following algorithm:

Given: snippet of length L1; known-texts of length L2 for each of C candidates

1. **Repeat** *k1* times
   a. Randomly choose some fraction *k2* of the full feature set
   b. Find top match using cosine similarity

2. **For each** candidate author A,
   a. Score(A) = proportion of times A is top match

Output: arg $\max_A$ Score(A) **if** max Score(A) > σ^*; **else** *Don't Know*

The idea is to check if a given author proves to be most similar to the test snippet for many different randomly selected feature sets of fixed size. The number of different feature sets used (*k1*) and the fraction of all possible features in each such set (*k2*) are parameters that need to be selected. The threshold σ^* that serves as the minimal score an author needs to be deemed the actual author is the parameter that we vary for recall-precision tradeoff.

We note that our method is similar in many respects to classifier ensemble methods in which different classifiers are learned using different subsets of features (Bryll et al. 2003).

6 Results

Preliminary experiments show that the greater the number of iterations, *k1*, the better the results, but that the added value of additional iterations begins to vanish as *k1* approaches 100. Thus, for all our experiments, we set *k1* = 100. Also, except where otherwise stated, we assume that the actual author is in the candidate set. This actually entails no loss of generality, as we will see.

6.1 Feature set size: *k2*

For our first experiment, we set the snippet length (L1) to 500, the known-text length for each candidate (L2) to 2,000 and the number of candidates to 10,000 and we test the impact of the choice of the fraction of available features used in the feature set, *k2*. Testing on 1,000 snippets, we construct recall-precision curves for various values of *k2* (Fig. 2). We find that larger feature sets yield greater accuracy. Using 40% (=100,000) of the 250,000 available features per iteration, at $\sigma^* = .90$, we achieve 87.9% precision with 28.2% recall (Fig. 1).

6.2 Number of candidate authors: C

We now consider how the number of candidate authors affects precision and recall. In Fig. 2, we show recall-precision curves (*k2* = 40%) for various numbers of candidate authors, using the same L1 and L2 settings as above. Note that, as expected, accuracy increases as the number of candidate authors diminishes. We mark on each curve the point $\sigma^* = .90$. For example, for 1,000 candidates, at $\sigma^* = .90$, we achieve 93.2% precision at 39.3% recall.

6.3 Open candidate sets

We have assumed thus far that the author of the snippet is among the candidate authors. We now consider the possibility that none of the candidate authors is the

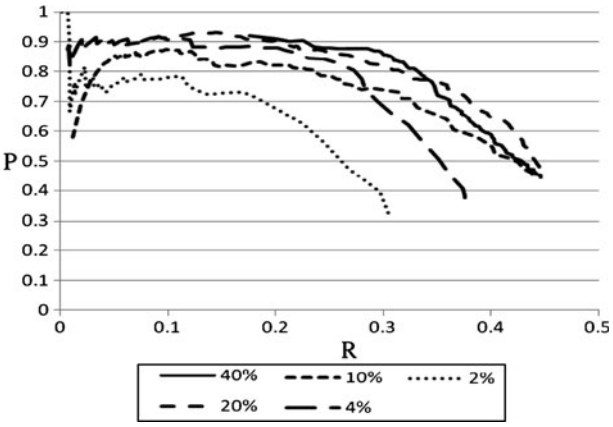

Fig. 1 Recall-precision for various feature set sizes

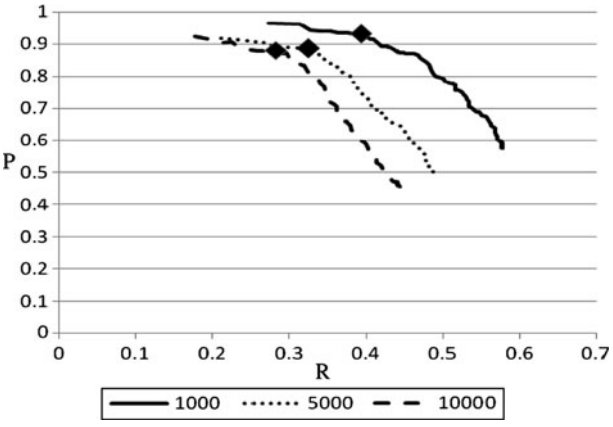

Fig. 2 Recall-precision for various candidates set sizes

actual author of the snippet. That is, we now wish to consider the open set attribution problem. What we would hope to find is that in such cases the method does not attribute the snippet to any of the candidates.

In fact, testing on 1000 snippets that belong to none of the candidates, we find that at $\sigma^* = .90$, very few are mistakenly attributed to one of the candidate authors: 2.5% for 10,000 candidates, 3.5% for 5000 and 5.5% for 1000. Perhaps counter-intuitively, for snippets by authors not among the candidates, having fewer candidates actually makes the problem *more* difficult since the fewer competing candidates there are, the more likely it is that there is some consistently most similar (but inevitably wrong) candidate.

6.4 Snippet length: L1

Next, we consider the effect of snippet size. In Fig. 4, we show recall-precision curves ($k2 = 40\%$) for 10,000 candidate authors as snippet size is reduced. We see

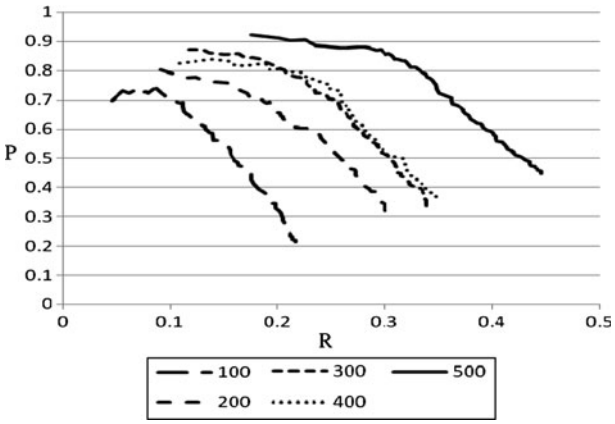

Fig. 3 Recall-precision for various snippet sizes

that, although results degrade rapidly with decreasing snippet size, even for as few as 100 words, we get precision of 71% at 10% recall (Fig. 3).

6.5 Known-text length: L2

Finally, we consider the effect of known-text size. For simplicity, we use the same amount of known-text for each candidate author. In Fig. 4, we show recall-precision curves ($k2 = 40\%$) for 10,000 candidate authors and snippet size of 500 as known-text size is reduced. We see that increased size of known-texts improves results, but that known-text size of 2,000 offers only a marginal improvement over known-text size of 1,500.

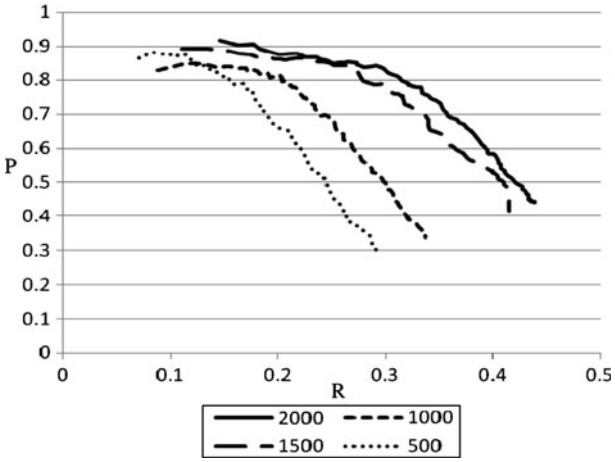

Fig. 4 Recall-precision for varying known-text size

7 Assessing attribution confidence

The above results show that the expected precision of an attribution of a snippet to one of a set of candidate authors depends on at least four factors: the number of candidate authors (C), the size of the snippet (L1), the size of the known-text (L2) and the score σ. We wish now to assess the importance of each of these four factors. We also wish to use this information to augment the output of our algorithm: in addition to presenting the most likely author of the snippet, we wish to estimate the probability that that author is in fact the correct one. In many applications, especially forensic ones, such confidence measures are crucial for usefulness of the results (e.g., to ensure admissibility in court). (Note that in our experiments below, we assume that the known-texts for all candidates are of uniform length. This is a convenience. When working with known-texts of varying lengths, our estimate of the probability that our attribution is correct will lie somewhere between the confidence value obtained using the shortest known-text and that obtained using the longest known-text.)

We consider a wide variety of combinations of the values C (ranging from 100 to 10,000), L1 (ranging from 20 to 500) and L2 (ranging from 500 to 2,000). For each of 1,000 snippets, we record the score σ achieved for each combination of such values. (We range over such values systematically by beginning with maximal candidate sets and known-text and snippets and iteratively (and independently) truncating known-texts and snippets and eliminating candidates.) For each combination of parameter values and each score σ, we compute the *coverage* H (the percentage of snippets for which the score σ is obtained) and the *precision* P (the percentage of those cases for which the author achieving that score is in fact the actual author).[1] In Table 1, we show these results for selected combinations of such values, for the cases in which the snippet author is among the candidates. We do the same for the case in which the snippet author is not among the candidates (table not shown).

Our goal now is to predict coverage and precision based on the problem parameters (L1, L2, C and σ), using regression. Since coverage for any given score is usually quite small, and thus there is considerable data sparseness, we smooth estimates by substituting for H and P for given σ, H and P for the interval $[\sigma - 5, \sigma + 5]$. (At the extremes, we use the largest possible interval symmetric around σ.)

Formally, our independent parameters are log(L1), log(L2), log(C) and σ, each scaled to lie in the interval [0,1]. We applied both ordinary linear regression and logistic regression to predict the dependent variables precision and coverage. Results for predicting precision using logistic regression are given in Table 2a. We see that prediction is fairly good ($r^2 = 0.77$) and correlation of P with each of the four independent parameters is significant at $p < .0001$. Results for predicting coverage using logistic regression are given in Table 2b. In this case $r^2 = 0.75$ and correlation of H with each of the three parameters other than log(C) are significant

[1] For purposes of clarity, we note the following: Recall, as discussed in Sect. 6, is simply $H*P$, the product of coverage and precision. Furthermore, results shown in Sect. 6 refer to those at score σ *or above*, while results shown in this section refer to those at a given score or in a given score interval.

Table 1 Precision (*P*) and coverage (*H*) percentages for selected combinations of parameter values

L2	L1	C	$\sigma < .80$		$\sigma \in [.80,.84]$		$\sigma \in [.85,.89]$		$\sigma \in [.90,.94]$		$\sigma \in [.95,1.00]$	
			P	*H*	*P*	*H*	*P*	*H*	*P*	*H*	*P*	*H*
1,000	167	640	14	84	39	2	70	2	64	3	78	9
		2,560	15	84	50	1	55	2	75	1	84	11
		10,240	13	84	64	1	43	2	68	2	82	10
	333	640	21	78	70	2	61	2	79	2	83	15
		2,560	15	80	52	3	77	1	82	3	82	13
		10,240	16	79	52	3	44	2	52	3	88	14
	500	640	17	73	64	2	57	2	79	4	84	19
		2,560	20	74	50	3	70	3	63	3	86	16
		10,240	18	74	62	3	64	1	77	3	84	18
1,500	167	640	17	81	33	3	70	2	74	3	84	11
		2,560	19	82	47	2	82	2	73	3	76	11
		10,240	16	82	57	2	82	2	64	2	84	11
	333	640	22	75	62	3	45	2	85	4	93	16
		2,560	22	75	71	2	85	3	77	3	82	17
		10,240	22	76	53	2	83	2	85	4	92	17
	500	640	22	72	62	3	61	2	70	4	85	19
		2,560	25	71	61	2	67	2	76	3	89	21
		10,240	23	70	67	3	89	3	81	3	87	21
2,000	167	640	20	80	74	2	61	2	75	2	85	13
		2,560	18	14	67	2	76	2	73	2	84	12
		10,240	19	80	58	3	70	2	85	3	90	13
	333	640	23	71	67	2	60	3	94	4	89	21
		2560	25	71	71	3	66	3	71	4	87	20
		10,240	23	70	69	3	63	2	69	4	89	22
	500	640	22	70	73	3	80	3	82	4	92	21
		2,560	24	67	62	3	72	3	88	4	92	24
		10,240	24	67	74	3	70	3	81	4	88	23

at $p < .0001$ (though there is no significant correlation with log(C)). Scatterplots showing precision and coverage vs. score are shown in Fig. 5.

In the case of *H*, we also observe that running separate regressions for the case $80 \leq \sigma < 90$ and $\sigma \geq 90$ yields considerably improved results, with respective r^2 values of 0.77 and 0.85. No such improvement holds for *P*. Results on all the above using linear regression are essentially the same for *P*, but not quite as good for *H*.

We now run identical experiments for the case in which snippets do not belong to any candidate authors and use the same methods to estimate for each combination of parameter values <C, L1, L2, σ> the value of *E*, namely, the probability that some candidate author will achieve a score of σ if none of the candidates are the true author.

Table 2 Logistic regression results for precision and coverage (as described in the text), showing the coefficients for the four parameters and the constant in the logistic function, together with the r^2 value and number of data points (N) for each regression

	Precision	Precision ($80 \leq \sigma < 90$)	Precision ($\sigma \geq 90$)
(a)			
Intercept	−11.3	−10.0	−11.2
log(C)	−0.104	−0.175	−0.0350
log(L1)	3.16	3.23	3.10
log(L2)	2.87	2.27	3.41
σ	7.56	6.67	6.99
r^2	0.77	0.65	0.71
N	2,359	1,125	1,234
	Coverage	Coverage ($80 \leq \sigma < 90$)	Coverage ($\sigma \geq 90$)
(b)			
Intercept	−12.9	−6.76	−17.9
log(C)	−0.0233	−0.0416	−0.0149
log(L1)	3.21	1.92	4.42
log(L2)	2.93	2.20	3.60
σ	5.06	0.152	8.57
r^2	0.75	0.77	0.85
N	2,359	1,125	1,234

Results are shown for regression on the full data (for $\sigma \geq 80$), as well as for regression on the subsets of the data with $80 \leq \sigma < 90$ and $\sigma \geq 90$, respectively

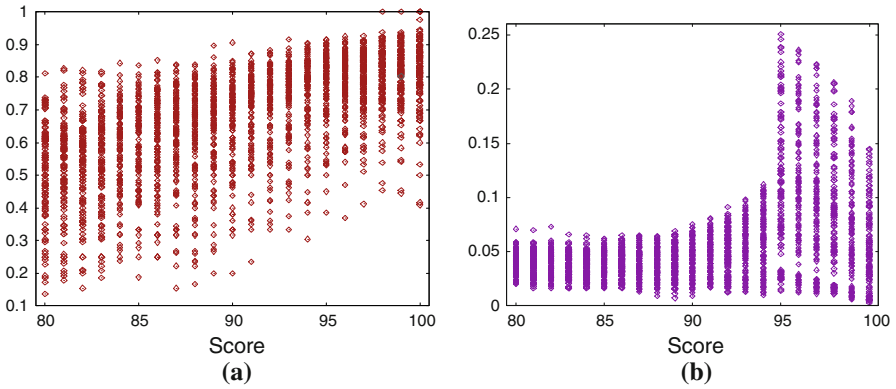

Fig. 5 Scatterplots showing precision and coverage vs. score. **a** Precision. **b** Coverage

Finally, we can combine the above results, to assign a probability to some combination <C, L1, L2, σ> . In order to do so, we need to introduce one more parameter. Denote by p the prior probability that the actual author of a snippet is in the candidate set. Note that in almost all authorship attribution research, p is simply

assumed to be 1; that is, it is taken as given that the correct author is in the candidate set. Since we do not make that assumption here, we think of p as a parameter the value of which must be provided by the user. (Of course, in the absence of any information, some reasonable default value for p, perhaps ½, can be chosen.)

Now consider a given snippet attribution problem with values of C, L1 and L2 and user-provided p. Then if the best candidate receives the score σ, the probability that this author is the actual author of the snippet can be estimated as $\frac{p \cdot H \cdot P}{p \cdot H + (1-p) \cdot E}$.

8 Conclusions

We have found that a naïve similarity-based method can be used to solve even the most difficult authorship attribution problems, provided that results are filtered through a robustness test based on randomized variation of feature sets. Thus, for example, the method can attribute a 500-word snippet to one of 1,000 authors with coverage of 42.2% and precision of 93.2%. Snippets that are not written by any of the candidates are rarely falsely attributed, though interestingly, the fewer candidates the greater the probability of such a false attribution. We note that passable results can be achieved even for snippets as short as 100 words.

Furthermore, we have found that the four parameters, snippet size, known-text size, number of candidates and score, account for most of the variability in coverage and precision, so that for any given attribution we can assign a fairly accurate estimate of the likelihood that the attribution is correct.

We conclude by briefly surveying the state of authorship attribution in the wild. The case of small closed candidate sets is well handled by standard text categorization methods. The case of large (open or closed) candidate sets is reasonably well handled by the method offered in this paper. The case of small open candidate sets is handled by unmasking (Koppel et al. 2007), provided that the anonymous text is very large. The remaining case with no satisfactory solution is that of a small open candidate set and limited anonymous text. The method pursued here cannot be directly applied in such cases since we have found that for small candidate sets, there is the danger that an anonymous text not written by any of the candidates might be attributed to one of them. One promising direction that we leave for future work is to artificially expand the candidate set in some plausible manner and then to apply our method.

References

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection. *ACM Transactions on Information Systems, 26*(2), 7.

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

Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition, 36*(6), 1291–1302.

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

Clough, P. (2000). Plagiarism in natural and programming languages: An overview of current tools and technologies, Research Memoranda: CS-00-05, Department of Computer Science, University of Sheffield, UK.

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

Juola, P. (2008). Author attribution, foundations and trends in information. *Retrieval, 1*(3), 233–334.

Keselj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-Gram-Based Author Profiles for Authorship Attribution. In *Proceeding of PACLING'03* (pp. 255–264). Halifax, Canada.

Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th ACM SIGIR Conference on Research and Development on Information Retrieval*. Seattle, Washington.

Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *JMLR, 8*, 1261–1276.

Koppel, M., Schler, J., & Argamon, S. (2008). Computational methods in authorship attribution. *JASIST, 60*(1), 9–26.

Luyckx, K., & Daelemans, W. (2008). Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)* (pp. 513–520). Manchester, UK.

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005). Author identification on the large scale. In *Proceedings of the Meeting of the Classification Society of North America, 2005*.

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

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

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *JASIST, 60*(3), 538–556.

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

Zhao, Y., & Zobel, J. (2005). Effective authorship attribution using function word. In *Proceedings of the 2nd AIRS Asian information retrieval symposium* (pp. 174–190). Berlin: Springer.

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