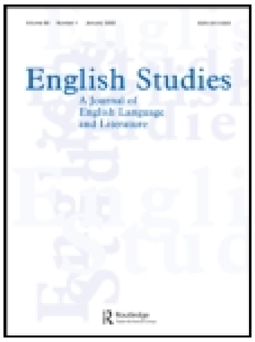

English Studies

Routledge
Taylor & Francis Group

ISSN: 0013-838X (Print) 1744-4217 (Online) Journal homepage: https://www.tandfonline.com/loi/nest20

The “Fundamental Problem” of Authorship Attribution

Moshe Koppel , Jonathan Schler , Shlomo Argamon & Yaron Winter

To cite this article: Moshe Koppel , Jonathan Schler , Shlomo Argamon & Yaron Winter (2012) The “Fundamental Problem” of Authorship Attribution, English Studies, 93:3, 284-291, DOI: 10.1080/0013838X.2012.668794

To link to this article: https://doi.org/10.1080/0013838X.2012.668794

Published online: 22 May 2012.

English Studies
Vol. 93, No. 3, May 2012, 284–291

Routledge
Taylor & Francis Group

The ''Fundamental Problem'' of Authorship Attribution

Moshe Koppel, Jonathan Schler, Shlomo Argamon and Yaron Winter

We introduce the ''fundamental problem'' of authorship attribution: determining if two, possibly short, documents were written by a single author. A solution to this problem can serve as a building block for solving almost any conceivable authorship attribution problem. Our preliminary work on this problem is based on earlier work in authorship attribution with large open candidate sets.

1. Introduction

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

Unfortunately, in the real world, we often encounter situations in which our list of candidates might be very large and in which there is no guarantee that the true author of an anonymous text is even among the candidates. Furthermore, the amount of writing we have by each candidate might be very limited, and the anonymous text itself might be short.

We therefore consider what we call the ''fundamental problem'' of authorship attribution: given two (possibly short) documents, determine if they were written by a single author or not. Plainly, if we can solve the fundamental problem, we can solve any of the standard authorship attribution problems, whether in the idealized form often considered or in the more difficult form typically encountered in real life.

Moshe Koppel is affiliated with the Department of Computer Science, Bar-Ilan University, Israel; Jonathan Schler is affiliated with the Department of Computer Science, Bar-Ilan University, Israel; Shlomo Argamon is affiliated with the Department of Computer Science, Illinois Institute of Technology, USA; Yaron Winter is affiliated with the Department of Computer Science, Bar-Ilan University, Israel. Email: koppel@cs.biu.ac.il

[1] Juola; Koppel, Schler, and Argamon, ''Computational Methods''; Stamatatos.

ISSN 0013-838X (print)/ISSN 1744-4217 (online) © 2012 Taylor & Francis
http://dx.doi.org/10.1080/0013838X.2012.668794

The main idea is as follows. Given two input texts, X and Y, we generate a set of impostors and then check if Y can be singled out from among the impostors as the most likely author of X: that is, we use artificially generated impostors to reduce the fundamental problem to that of authorship attribution on large open candidate sets. We will describe our solution to this latter problem in some detail, drawing on our recent work.[2] We will then briefly sketch how that work can be adapted to solve the fundamental problem.

2. Related Work

We have previously suggested a method known as "unmasking" for determining if two book-length texts were written by the same author.[3] Unfortunately, unmasking does not work for short documents.[4] 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. This principle proves to be valuable in the method reported here as well.

As a rule, automated techniques for authorship attribution can be divided into two main types.[5] In *similarity-based* methods, some metric is used to measure the distance between two documents, and an anonymous document is attributed to that author whose known writing (considered collectively as a single document) is most similar.[6] In *machine-learning* methods, the known writings of each candidate author (considered as a set of distinct training documents) are used to construct a classifier that can then be used to classify anonymous documents.[7]

Research in the similarity-based paradigm has focused on the choice of features for document representation, on methods for dimensionality reduction (such as Principal Components Analysis—PCA) of the feature space, and on the choice of distance metric. Research in the machine-learning paradigm has focused on the choice of features for document representation and on the 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 David Madigan et al. (114 authors) and Kim Luyckx and Walter Daelemans (145 authors).[8] Moshe Koppel et al. have considered candidate sets including thousands of authors.[9] Both Koppel et al. and Luyckx and Daelemans observed that when there are very many candidate authors, similarity-based methods are more appropriate than machine-learning methods.[10]

[2] Koppel, Schler, and Bonchek-Dokow.

[3] Ibid.

[4] Sanderson and Guenter.

[5] Stamatatos.

[6] Burrows, "Delta." Hoover, "Multivariate Analysis"; Abbasi and Chen; Argamon.

[7] Zhao and Zobel; Zheng et al.; Abbasi and Chen; Koppel, Schler, and Argamon, "Computational Methods."

[8] Madigan et al.; Luyckx and Daelemans.

[9] Koppel et al.; Koppel, Schler, and Argamon, "Authorship Attribution in the Wild."

[10] Koppel et al.; Luyckx and Daelemans.

For this reason, our approach here is in the similarity-based paradigm, so we devote a few words to its main principles. The basic idea is that the most reasonable attribution can be found by viewing each document as a point in some multidimensional space, and assigning a questioned document to the author whose documents as a whole are ‘‘closest’’, according to some appropriate distance measure. There are a number of reasonable such distance measures that have been applied to the problem. A popular such method is Burrows’s Delta,[11] which, with extensions, has been used for a variety of attribution problems;[12] it amounts to an approximate probabilistic ranking based on a multidimensional Laplacian distribution over word frequencies.[13] Many other similarity functions have been applied to different feature sets as well.[14]

Just as most work in authorship attribution has focused on tasks with small numbers of candidate authors, almost all work has also focused on the case in which the candidate set is *closed*, that is, 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 in the simplest case 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.[15]

The general authorship attribution problem, where candidate sets are simultaneously large and open was recently considered by Moshe Koppel, Jonathan Schler, and Shlomo Argamon.[16] In what follows we present the main results of that paper, as they relate to the problem at hand.

3. Handling Large Open Candidate Sets

We use a set of ten thousand 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 two thousand words of known text and a *snippet*, consisting of the last five hundred 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.

We will experiment using various subsets of the available text to determine the impact on the attribution of the number of candidates.

[11] Burrows, ‘‘Delta.’’

[12] Burrows, ‘‘Englishing of Juvenal’’; Hoover, ‘‘Delta Prime’’; Hoover, ‘‘Testing Burrows’s Delta.’’

[13] Stein and Argamon; Argamon.

[14] Craig; Chaski; Stamatatos, Fakotakis, and Kokkinakis; Keselj, Cercone, and Thomas; van Halteren et al.

[15] Clough; Meyer zu Eisen, Stein, and Kulig.

[16] Koppel, Schler, and Argamon, ‘‘Authorship Attribution in the Wild.’’

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. In our corpus, there are just over 250,000 unique (but overlapping) space-free character 4-grams. We select the 100,000 such features most frequent in the corpus as our feature universe. Character n-grams have been shown to be effective for authorship attribution[17] and have the advantage of being measurable in any language without specialized background knowledge.

It is not practical to learn a single classifier for ten thousand classes, nor is it practical to learn ten thousand 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.[18] Testing this rather naïve method on one thousand snippets selected at random from among the ten thousand authors, we find that 46 per cent of the snippets are correctly assigned.

While this is perhaps surprisingly high, a precision of 46 per cent is inadequate for most applications. To remedy this problem, we adopt the approach of Koppel et al.,[19] which permits a response of *Don't Know* in cases where attribution is uncertain. The objective is to obtain high precision for those cases where an answer is given, while trying to offer an answer as often as possible.

The key to our new approach is the same as the underlying principle of "unmasking".[20] 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
```

[17] Keselj, Cercone, and Thomas.

[18] Salton and Buckley.

[19] Koppel et al.

[20] Koppel, Schler, and Bonchek-Dokow.

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^* serves as the minimal score an author needs to be deemed the actual author. We vary this parameter to obtain a trade-off between recall and precision.

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.[21]

4. Results

We set the number of iterations (*k1*) to 100, the snippet length (L1) to 500, the known-text length for each candidate (L2) to 2,000, and the fraction of available features used in the feature set (*k2*) to 40 per cent. We consider how the number of candidate authors affects precision and recall. In Figure 1, we show recall-precision curves for various numbers of candidate authors. 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 per cent precision at 39.3 per cent recall.

We now consider the possibility that none of the candidate authors is the actual author of the snippet. 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 1,000 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 per cent for 10,000 candidates, 3.5 per cent for 5,000 and 5.5 per cent for 1,000. 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.

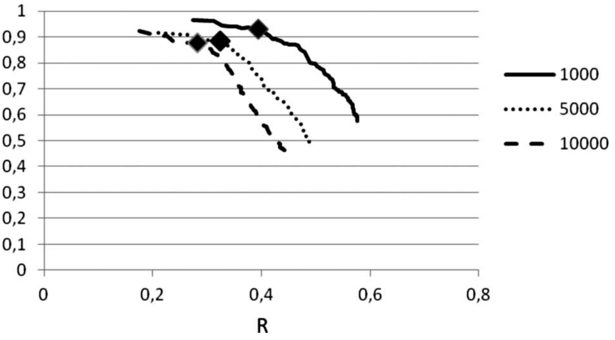

Figure 1 Recall-Precision for Various Candidates Set Sizes.

[21] Bryll, Gutierrez-Osuna, and Quek.

5. The Fundamental Problem of Authorship Attribution

The above method can serve as the basis for solving what we call the "fundamental problem" of authorship attribution: given two (possibly short) documents, determine if the two documents were written by the same author or two different authors. Plainly, if we can solve this problem, we can solve the standard attribution problems considered above, as well as almost any authorship attribution problem we can conceive of.

Our approach to solving the fundamental problem is as follows. Given two texts, X and Y, we generate a set of impostors (Y_1, …, Y_n) and then use the above method to determine if X was written by Y or any of the impostors or by none of them. If and only if we obtain that X was written by Y with a sufficiently high score, we say that the two documents are by a single author. (Clearly, we can additionally, or alternatively, generate impostors X_1, …, X_n and compare them to Y.)

The questions that we need to explore in order to adapt the above method to our problem are the following:

- How many impostors should be used?
- How should the impostors be chosen?
- What score should we require in order to conclude that two documents are by a single author?

We report here for the first time preliminary answers to these questions based on recently concluded experiments.

Using a development set consisting of 500 pairs of blog posts written by a single author and 500 pairs written by two different authors, we restricted our potential impostor universe to other blog posts. (This restriction is simply a matter of convenience; we don't claim it is optimal.) Given a pair <X, Y>, we found that the best way to choose actual impostors Y1, …, Yn from this universe, is as follows. Identify the 250 most similar blog posts to Y. (This ensures that the impostors at least roughly resemble Y.) Then randomly choose twenty-five blog posts from among these to serve as impostors.

Using this protocol on an independent test set, we assign <X, Y> to a single author if and only if Y is selected from among the set {Y, Y_1, …, Y_n} as most similar to X in at least eleven trials out of one hundred. We found that we obtain precision of 90 per cent and recall of 83 per cent.

We believe that these results suggest that the fundamental problem of authorship is a solvable problem.

6. Conclusions

We leverage our work on authorship attribution with large open candidate sets to propose a method for solving what we call the "fundamental problem" of authorship attribution: determining if two, possibly short, documents were written by the same author. The key is to properly choose an impostor set. Our method for choosing the set is to identify some relevant universe, then choose a large set of potential impostors that are sufficiently similar to the input documents, and finally randomly choose twenty-five impostors from among that set. The identification of a "relevant universe" requires further study, but in preliminary experiments we have found that it is sufficient to simply generate the universe by using results of Google searches on queries randomly selected from the input documents.

By varying the threshold for identifying an input pair as being by the same author, we control the trade-off between precision and recall. We have found that even when we insist on precision of 90 per cent (as we must, for example, in forensic applications), we obtain recall above 80 per cent. These results seem to be a very promising basis for future investigations.

References

Abbasi, A., and H. Chen. "Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection." *ACM Transactions on Information Systems* 26, no. 2 (2008): 7.

Argamon, Shlomo. "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations." *Literary and Linguistic Computing* 23, no. 2 (2008): 131–47.

Bryll, Robert, Ricardo Gutierrez-Osuna, and Francis Quek. "Attribute Bagging: Improving Accuracy of Classifier Ensembles by Using Random Feature Subsets." *Pattern Recognition* 36, no. 6 (2003): 1291–302.

———. "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17, no. 3 (2002): 267–87.

———. "The Englishing of Juvenal: Computational Stylistics and Translated Texts." *Style* 36, no. 4 (2002): 677–99.

Chaski, Carole E. "Empirical Evaluations of Language-based Author Identification Techniques." *Forensic Linguistics* 81, no. 1 (2001): 1–65.

Clough, Paul. *Plagiarism in Natural and Programming Languages: An Overview of Current Tools and Technologies.* Sheffield: University of Sheffield, Department of Computer Science. Available from http://ir.shef.ac.uk/cloughie/papers/plagiarism2000.pdf (accessed 20 September 2011).

Craig, Hugh. "Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything about Them?" *Literary and Linguist Computing* 14, no. 1 (1999): 103–13.

Hoover, David L. "Delta Prime?" *Literary and Linguistic Computing* 19, no. 4 (2004): 477–95.

———. "Multivariate Analysis and the Study of Style Variation." *Literary and Linguistic Computing* 18, no. 4 (2003): 341–60.

———. "Testing Burrows's Delta." *Literary and Linguistic Computing* 19, no. 4 (2004): 453–75.

Juola, Patrick. "Author Attribution." *Foundations and Trends in Information Retrieval* 1, no. 3 (2008): 233–334.

Keselj, V. Fuchun Peng, Nick Cercone, and Calvin Thomas. "N-gram-based Author Profiles for Authorship Attribution." *Computational Linguistics* 3 (2003): 255–64.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. "Authorship Attribution in the Wild." *Language Resources and Evaluation* 45, no. 1 (2011): 83–94.

———. "Computational Methods in Authorship Attribution." *Journal of the American Society for Information Science and Technology* 60, no. 1 (2009): 9–26.

Koppel, Moshe, Jonathan Schler, Shlomo Argamon, and Eran Messeri. "Authorship Attribution with Thousands of Candidate Authors." In *Proceedings of the 29th ACM SIGIR Conference on Research and Development in Information Retrieval*, 659–60. 2006. New York: ACM.

Koppel, Moshe, Jonathan Schler, and Elisheva Bonchek-Dokow. "Measuring Differentiability: Unmasking Pseudonymous Authors." *Journal of Machine Learning Research* 8 (2007): 1261–76.

Luyckx, Kim, and Walter Daelemans. "Authorship Attribution and Verification with Many Authors and Limited Data." In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, 513–20. Menlo Park, CA: ACL.

Madigan, David, Alexander Genkin, David D. Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. "Author Identification on the Large Scale." In *Proceedings of the Meeting of the Classification Society of North America*, 1–20. 2005. St. Louis, MI: CSNA.

Meyer zu Eissen, Sven, Benno Stein, and Marion Kulig. "Plagiarism Detection without Reference Collections." In *Advances in Data Analysis,* edited by Reinhold Decker and Hans Joachim Lenz, 359–66. New York: Springer, 2007.

Salton, Gerard, and Christopher Buckley. "Term-weighting Approaches in Automatic Text Retrieval." *Information Processing and Management* 24, no. 5 (1988): 513–23.

Sanderson, Conrad, and Simon Guenter. "Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation." In *Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 482–91. Sydney: Association for Computational Linguistics, 2006.

Stamatatos, Efstathios. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60, no. 3 (2009): 538–56.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis. "Computer-based Authorship Attribution without Lexical Measures." *Computers and the Humanities* 35, no. 2 (2001): 193–214.

Stein, Sterling, and Shlomo Argamon. "A Mathematical Explanation of Burrows's Delta." In *Proceedings of the Digital Humanities Conference*. London: ALLC, 2006.

Van Halteren, Hans, Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. "New Machine Learning Methods Demonstrate the Existence of a Human Stylome." *Journal of Quantitative Linguistics* 12, no. 1 (2005): 65–77.

Zhao, Ying, and Justin Zobel. "Effective Authorship Attribution Using Function Word." *Lecture Notes in Computer Science* 3689 (2005): 174–89.

Zheng, Rong, Jiexun Li, Hsinchun Chen, and Zan Huang. "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, no. 3 (2006): 378–93.