Distractorless Authorship Verification

John Noecker Jr[1], and Michael Ryan[1]

[1]Evaluating Variations in Language Laboratory, Duquesne University, Pittsburgh, Pennsylvania, USA
{jnoecker,mryan}@jgaap.com

Abstract. Authorship verification is the task of, given a document and a candidate author, determining whether or not the document was written by the candidate author. Traditional approaches to authorship verification have revolved around a "candidate author vs. everything else" approach. Thus, perhaps the most important aspect of performing authorship verification on a document is the development of an appropriate distractor set to represent "everything not the candidate author". The validity of the results of such experiments hinges on the ability to develop an appropriately representative set of distractor documents. Here, we propose a method for performing authorship verification without the use of a distractor set. Using only training data from the candidate author, we are able to perform authorship verification with high confidence (greater than 90% accuracy rates across a large corpus).

Keywords. Authorship verification, authorship attribution, corpus linguistics, computational stylometry.

1 Introduction

In traditional authorship attribution, our task is to assign an author label (or a similar categorical label such as genre, publication date, etc.) to a work of unlabeled authorship. In the closed-set problem, we assign a label from a set of potential authors for whom we have some labeled training data. In the open-set problem, we also allow for the answer "none of the above". We build upon this here with the authorship verification task, which is essentially an open-class authorship attribution problem with only one author in the candidate pool. Thus for a given document D and candidate author A, we attempt to answer the question "Was D written by A?".

2 Background

Previous approaches to this problem [1-2] have involved the creation of a distractor set, which is normally controlled for genre, tone, length, etc. and performing an a traditional authorship attribution-style analysis to see whether the unlabeled document is more likely to be by the candidate author or one of the authors in the distractor set. This approach is not ideal because it relies heavily on the creation of an appropriate distractor set. That is, enough information was available about the candidate author and the training documents to choose a set of distractor authors that were appropriate for the task. Thus, although these methods performed well at the verification task, they do not lend themselves well to automation. Indeed, the entire result of this type of authorship verification hinges upon the documents chosen for the distractor set. This creates a sort of chicken-and-egg problem wherein it is necessary to know the answer in order to evaluate the suitability of the distractor set, yet it is necessary to know whether the distractor set is appropriate in order to evaluate the results.

We will attempt to remedy the errors introduced by the distractor set by eliminating the set entirely. Instead, we will consider only the document in question as well as a sample of writing known to belong to the candidate author. Thus, the validity of the verification task hinges only on obtaining a representative model of the candidate author's work, a requirement shared by traditional verification tasks, and does not involve any guesswork.

3 Materials and Methods

3.1 Distractorless Authorship Verification

Goal: Given a document D, and a candidate author A, determine the likelihood that D is written by A.

Method:

1. Compile a set of *training data*, which is known to be written by A.
2. Compile a *model* from the training data. This is normally accomplished by extracting linguistic or token-level features from the text and compiling a feature vector using any of various standard techniques from the authorship attribution field. We will label this feature vector $M = \langle m_1, m_2, \ldots, m_n \rangle$.
3. Extract a *feature set*, F, from D in the form of $F = \langle f_1, f_2, \ldots, f_n \rangle$, where f_i corresponds to m_i for all i.
4. Choose a "distance like" function, δ, such that if $\delta(x, y) > \delta(x, z)$, we can say that x is "closer to" or "more similar to" y than to z (in some meaningful way).
5. Choose a threshold, t, such that if $\delta(M, F) > t$, we accept the premise that M and F are written by the same author, A. This threshold is found empirically by ana-

lyzing the average δ between documents of the same author.

3.2 The Corpora

To evaluate the performance of our authorship verification algorithms, we used two publically-available corpora. We made use of the Ad-hoc Authorship Attribution Competition corpus (AAAC) [3] and the PAN 2011 Authorship Identification Training Corpus [4].

Ad-hoc Authorship Attribution Competition Corpus

The AAAC was an experiment in authorship attribution held as part of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities. The AAAC provides texts from a wide variety of different genres, languages and document lengths, assuring that the results will be useful for a wide variety of applications. The AAAC corpus consists of 264 training documents, 98 test documents and 63 authors, distributed across 13 different problems (labeled A-M). A description of each problem is found in Table 1.

PAN 2011 Authorship Identification Training Corpus

The PAN 2011 Authorship Identification Training Corpus consists of "real-world texts" (described by the originators as "often short and messy" [4]). These texts appear to come primarily from the Enron Email Dataset [5]. Each text contained authorship information within the text itself, which was removed during the preprocessing stage. They are included to ensure that the results obtained herein are applicable in real-world situations, and to avoid overtraining on the AAAC data. The PAN corpus contained 5,064 training documents and 1,251 test documents across 10 authors.

Problem	Language	Description	Number of Authors	Training Docs.	Test Docs.
A	English	Student Essays	13	38	13
B	English	Student Essays	13	38	13
C	English	Novels	4	17	9
D	English	Plays	3	12	4
E	English	Plays	3	12	4
F	Middle English	Letters	3	60	10
G	English	Novels	2	6	4
H	English	Speech Transcripts	3	3	3
I	French	Novels	2	5	4
J	French	Cross-Genre	2	5	2
K	Serbian-Slavonic	Cross-Genre	3	14	4
L	Latin	Poetry	4	6	4
M	Dutch	Student Essays	8	48	24

Table 1. AAAC Breakdown

3.3 Preprocessing

Preprocessing was performed on the text based on current best practices from traditional authorship attribution. Both corpora were preprocessed to standardize whitespace and character case. Any sequence of whitespace characters in the documents converted to a single space, and all characters were converted to lower case. As previously mentioned, for the PAN corpus, we also removed the author tags from each document.

3.4 Features

We used a variety of features, also chosen for their known performance in traditional authorship attribution, in our approach. Current research [6] shows that character n-grams are strong performers in traditional authorship attribution tasks. As such, we have focused on these features, examining results for character n-grams for n from 1 to 20. For completeness, we also provide results for word n-grams for n from 1 to 10. Here, a word is defined as any series of non-whitespace characters separated by whitespace. The n-grams are generated using a sliding window of size *n* and slide 1. We have limited ourselves to these simple features as they can be calculated very rapidly and without risk of error (such as that introduced by imperfect part-of-speech taggers), and thus lend themselves well to rapid, confident analysis.

3.5 Author Model

In order to perform the distractorless authorship verification, it is necessary to accurately model the writing style of the candidate author. We accomplished this using the centroid of the feature vectors for each training document. The centroid was calculated by using the average *relative* frequency of each event across the training documents, to adjust for variations in training document length.

3.6 Analysis Method

For this study, we have limited ourselves to the use of a normalized dot-product (*Cosine Distance*) analysis method. This method was shown in [7] to be among the best performing and simplest methods for authorship attribution. The advantage to using the Cosine Distance, particularly in conjunction with the simple features described above, is that it is possible to perform this verification extremely quickly, even on very large data sets. In order

to answer the verification task, we need only consider the dot product of the unknown document with the candidate author model. So, let:

$$\delta(M,F) = \frac{M \cdot F}{\|M\|\|F\|} = \frac{\sum_{i=1}^{n} m_i f_i}{\sum_{i=1}^{n} m_i^2 \sum_{i=1}^{n} f_i^2}$$

4 Results

4.1 AAAC Corpus Results

Characters

For the AAAC Corpus with character n-grams, we achieved our best results using Character 12-grams. Our highest accuracy was 87.44% and our highest F-Score was 47.12%. These results were achieved by setting a threshold, t, of t=0.099387 and t = 0.001597 respectively. For character n-grams of various values of n, our best accuracies varied from about 86% to about 88%, while our best F-Scores varied from 37% to 47%, as shown in Figures 1 and 2.

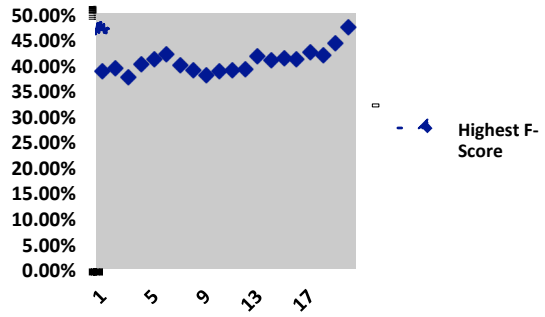

Fig. 1. AAAC Character n-grams by Highest F-Score

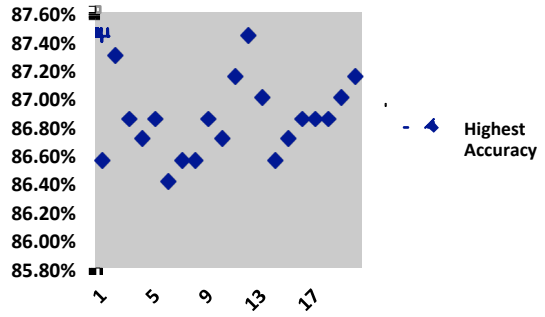

Fig. 2. AAAC Character n-grams by Highest Accuracy

Words

For the AAAC Corpus with word n-grams, we achieved our best results using Word 4-grams. Our highest accuracy was 88.04% and our highest F-Score was 44.58%. These results were achieved by setting a threshold, t, of t=0.006923 and t = 0.000029 respectively. For word n-grams of various values of n, our best accuracies varied from about 86% to about 88%, while our best F-Scores varied from 37% to 45%, as shown in Figures 3 and 4.

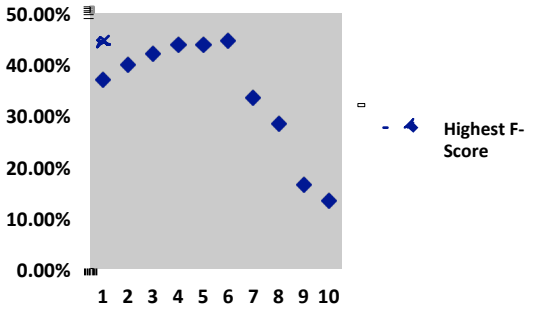

Fig. 3. AAAC Word n-grams by Highest F-Score

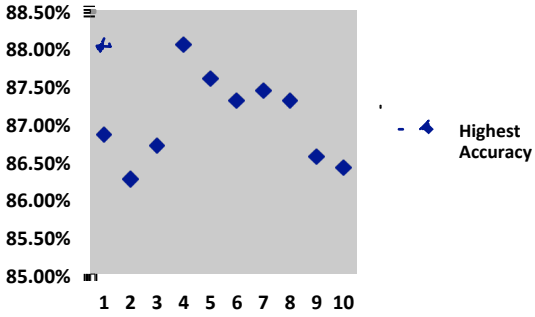

Fig. 4. AAAC Word n-grams by Highest Accuracy

4.2 PAN Corpus Results

Characters

For the PAN Corpus with character n-grams, we achieved our best results using Character 7-grams. Our highest accuracy was 92.23% and our highest F-Score was 51.35%. These results were achieved by setting a threshold, *t*, of *t = 0.1643* and *t = 0.1707* respectively. For character n-grams of various values of n, our best accuracies varied from about 90% to about 92%, while our best F-Scores varied from 20% to 51%, as shown in Figures 5 and 6.

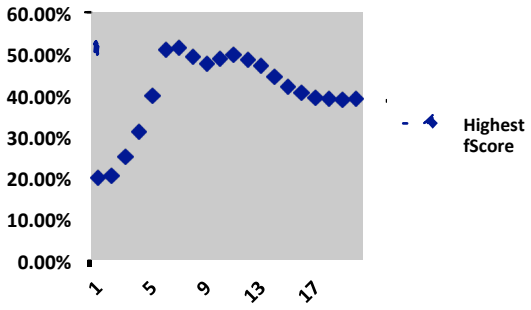

Fig. 5. PAN Character n-grams by Highest F-Score

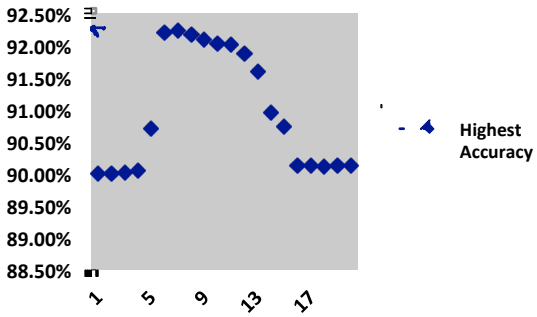

Fig. 6. PAN Character n-grams by Highest Accuracy

Words

For the PAN Corpus with word n-grams, we achieved our best results using Word 2-grams. Our highest accuracy was 91.53% and our highest F-Score was 43.08%. These results were achieved by setting a threshold, *t*, of *t = 0.1518* and *t = 0.1078* respectively. For word n-grams of various values of n, our best accuracies varied from about 90% to about 92%, while our best F-Scores varied from 10% to 43%, as shown in Figures 7 and 8.

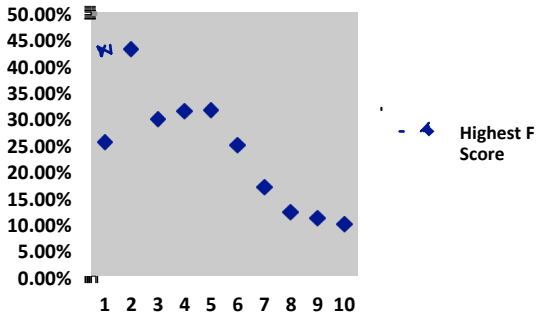

Fig. 7. PAN Word n-grams by Highest F Score

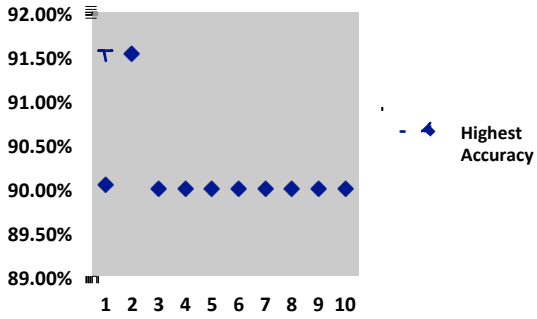

Fig. 8. PAN Word n-grams by Highest Accuracy

5 Discussion and Conclusion

These results are as expected, and show that distractorless authorship verification is a viable technique with high accuracy even on extremely difficult problems. The lower accuracy rates on the AAAC Corpus are expected, as this corpus contains many times more authors than the PAN corpus with fewer training documents per author. Despite this, we were able to achieve accuracy rates of up to 88% on this more difficult corpus, and 92% on the "easier" PAN corpus.

It should be noted that the results of this technique are extremely tunable. For instance, we can tune for any combination of accuracy, precision, recall and F-Score, as desired. That is, this distractorless authorship verification technique allows us to tune the Type I vs. Type II error rates depending on the application of the technology. For instance, in a forensic context we may want to err on the side of saying "no", and thus prefer to reduce false positives, while in a less stringent application we may wish to improve the overall accuracy at a cost of possibly having more false positives.

Given the performance on both the PAN corpus, which consisted mainly of real-life e-mail messages, and the AAAC corpus, which contained a wide variety of documents, the distractorless authorship verification technique shows promise for a wide range of genres and document lengths, and appears to work across a variety of languages, all without much tuning. Indeed, the only difficulty appears to be in finding an appropriate threshold, *t*, for given candidate author. This is possible by analyzing the average δ between documents by the candidate author. It can also be approximated from a large corpus, as was done here, although these results make it clear that there will be some of the same problems in controlling the corpus for genre, document length, etc. that were present in forming a distractor set for traditional authorship verification.

Future work will focus on improving these results, mainly through the addition of confidence ratings for the verification answers. That is, by allowing the system to decline to answer some percentage of the verification

questions asked (or, effectively, to answer "I don't know"), we have seen some improvement in these experimental results. Although full results for this study are not yet available, we have seen that by dropping the 20% "most difficult" verification problems, we see an increase in accuracy to approximately 96% (from 92% on the strict binary problem). Whether or not this is truly an increase in accuracy depends upon the intended application of the technology, but we believe this future work will provide interesting results in application-specific tradeoffs, as described above.

Overall, we believe we have shown this distractorless authorship verification to be a useful tool on the stylometrist's workbench. Although no tool is itself a panacea, we are also planning efforts to combine this technique with a mixture-of-experts style voting system, effectively using multiple distance functions and feature sets on the same problem to increase confidence in our answer. Finally, we hope to both more fully explore the process of determining the appropriate threshold without the need for extraneous texts, and to find an acceptable "default" threshold for cases where there is little training data. We have found enough similarity in the highest accuracy and highest F-Score thresholds that we believe a threshold point may exist that, despite being slightly less accurate than application-specific thresholds, will perform adequately across a wide range of documents.

6 References

1. Koppel Moshe and Jonathan Schler, "Authorship Verification as a One-Class Classification Problem", ICML '04 Proceedings of the Twenty-First International Conference on Machine Learning., 2004, New York, NY
2. Koppel Moshe and Jonathan Schler, "Text Categorization for Authorship Verification", 8th Symposium on Artificial Intelligence, 2004.
3. Juola, Patrick. (2004). "Ad-Hoc Authorship Attribution Competition" ALLC/ACH 2004 Conference Abstracts. Gothenberg: University of Gothenberg.
4. PAN 2011 Lab: Uncovering Plagiarism, Authorship, and Social Software Misuse. Authorship Identification Training Corpus. Amsterdam, 2011.
5. Enron Email Dataset. http://www.cs.cmu.edu/~enron/
6. Juola, Patrick and Michael Ryan,. Authorship Attribution, Similarity, and Noncommutative Divergence Measures. In Selected Papers from the Chicago DHCS Colloquium. 2008. Chicago, IL,: Chicago Colloquium on Digital Humanities and Computer Science.
7. Noecker Jr, John and Patrick Juola,. Cosine Distance Nearest-Neighbor Classification for Authorship Attribution. In Proceedings from Digital Humanities 2009. College Park, Md.: Digital Humanities 2009