Authorship Attribution with Convolutional Neural Networks and POS-Eliding

Julian Hitschler and **Esther van den Berg** and **Ines Rehbein**
Computational Linguistics
University of Heidelberg
69120 Heidelberg, Germany
{hitschler,vdberg,rehbein}@cl.uni-heidelberg.de

Abstract

We use a convolutional neural network to perform authorship identification on a very homogeneous dataset of scientific publications. In order to investigate the effect of domain biases, we obscure words below a certain frequency threshold, retaining only their POS-tags. This procedure improves test performance due to better generalization on unseen data. Using our method, we are able to predict the authors of scientific publications in the same discipline at levels well above chance.

1 Introduction

Computational authorship identification is a task of great interest for many historical and forensic applications. In order to judge the applicability of current and future authorship identification techniques, they need to have been tested in a variety of realistic settings. As it stands, the accuracy of procedures for automatic authorship attribution varies widely with the setting of the task. Among the variables affecting the accuracy of authorship attribution systems identified by Koppel et al. (2013) are the number of target authors a text is to be attributed to, the presence of an *other*-class in the test set (containing texts not written by any of the authors in the training set), the length of the text segments to be classified, and the amount of training data available.

Another important variable which is frequently unaddressed in the computational authorship attribution literature but which deserves closer attention is the monotonicity or diversity of genres and domains in the data, as well as the domain- and genre-specificity of the writings of individual authors. This work introduces a task setting for authorship attribution that is highly invariant with respect to genre and domain, as well as design ideas for systems adapted to this challenging setting.

We conducted a controlled study on the effects of domain and genre bias on authorship attribution by means of an ablation analysis where words in a text, but not their automatically predicted POS-tag, are obscured at various frequency cutoffs. The aim is the design of a system which can perform authorship attribution of texts which are extremely similar in terms of genre and domain among a large class of target authors, based solely on features extracted from POS-tags and a small core vocabulary. The central research question is how well computational authorship attribution works when based on purely stylometric (as opposed to content) features. In doing so, we shed light on the effect that thematic biases have on results in the area of computational authorship attribution.

2 Related Work

Early work on authorship attribution using statistical methods began as early as the first half of the 20th century (Yule, 1938; Zipf, 1932).[1] Modern authorship attribution was strongly influenced by the work of Mosteller and Wallace (1964) who tried to determine the authors of the Federalist Papers, given a small set of probable candidates. Mosteller and Wallace developed a method based on stylometric features in the texts, such as sentence length, word length, or the distribution of high-frequency function words. For a long time, work on authorship attribution has followed this approach and modeled the task as a closed-set classification problem, assuming that we have access to training data for all the authors in the set.

This setting, however, is highly unrealistic, as has been pointed out by Koppel et al. (2013).

[1] For an overview on modern authorship attribution methods, see (Stamatatos, 2009).

Proceedings of the Workshop on Stylistic Variation, pages 53–58
Copenhagen, Denmark, September 7–11, 2017.

In most realistic scenarios, there will not be a known set of authors to choose from, but an indefinite number of candidates, most of them unknown writers. This means that the closed-set assumption might lead to invalid conclusions, i.e. to consider features as discriminants that are able to model authorship on the closed set, but will not perform well on the large, unseen data that *should be* our test set. In this work, we assume a closed set of authors, however, the set of candidates is large (>800).

Other problems for authorship attribution concern the confusion of author style with genre (Byrnes and Sprang, 2004) and topic (Mikros and Argiri, 2007). The same effects are also relevant for related tasks, e.g. for Native Language Identification (NLI). As shown by Brooke and Hirst (2011), the topic of a document can often bias classification results in an NLI task, even when abstracting away from the context words by using character ngrams. Golcher and Reznicek (2011) reported a similar effect, showing how topic works as a confounding variable when investigating L1 influences in learner language. To assess the real potential of authorship attribution techniques, we need methods that are able to generalize to unseen data, and that are robust against the impact of topic and genre.

Stamatatos (2017) addresses the problem of topic-sensitivity using text distortion. Before extracting token or character ngram features, he masks all tokens that occur below a certain frequency threshold by replacing either the whole token or each character in the token by an asterisk. He tests his approach in an authorship attribution task on texts from different topics and genres (<15 authors), and in an author verification task on data from the PAN 2014 evaluation campaign (Stamatatos et al., 2014). Stamatatos shows that SVMs trained on the features extracted from the distorted texts outperform previous models in a cross-topic scenario. For topic-specific settings, however, where each author is strongly correlated with a specific topic, his approach yields results below the baseline.[2]

So far, only few studies have employed deep neural networks (NN) for authorship attribution. Ge et al. (2016) used a feed-forward NN language model to classify short transcripts from 18 coursera lectures that are controlled for topic. Rhodes (2015) trained a convolutional neural network (CNN) on word representations to classify medium-sized texts, and Shrestha et al. (2017) applied a CNN to identify the authors of tweets, based on character ngrams. Bagnall (2015) used a multi-headed recurrent neural network (RNN) language model to estimate character probabilities for each author in the PAN 2015 authorship identification task and outperformed all other models. Their results show the promise of deep NN for improving authorship attribution.

Our approach is similar in spirit to that of Stamatatos (2017). We also obscure words that occur below a certain frequency threshold. In contrast to Stamatatos, however, we use a CNN to classify the texts. We test our approach in a more realistic setting where the author has to be chosen from a much larger set of candidates (>800). To disentangle the influence of topic and genre from author style, we test our method on a highly homogeneous set of scientific articles from the areas of computational linguistics and NLP.

3 Datasets and Tools

In our experiments, we used single-author papers from the ACL Anthology Reference Corpus (Bird et al., 2008). The corpus contains scientific papers published in the proceedings of various conferences and workshops in the areas of computational linguistics and natural language processing. The earliest data is from 1965, the latest data is from 2007. We designated all papers published in the year 2006 as development data and all papers published in 2007 as test data, with the remaining data used for training. New authors without publications before this date were not treated any differently from those which were represented in the training data. We only retained publications from authors with at least two single-author papers, although we do not require both or even one of them to be part of the training data. Our dataset contained 808 distinct authors. We discarded the first 10 lines of each document in order to strip publications of author names, email addresses and workplace information. We also removed any lines containing the author's last name (for example, as part of a self-citation or email ad-

[2]The reason for this most probably lies in the closed-class assumption of the setting, and we expect different results for a more realistic test set where the strong correlation between author and topic does not hold.

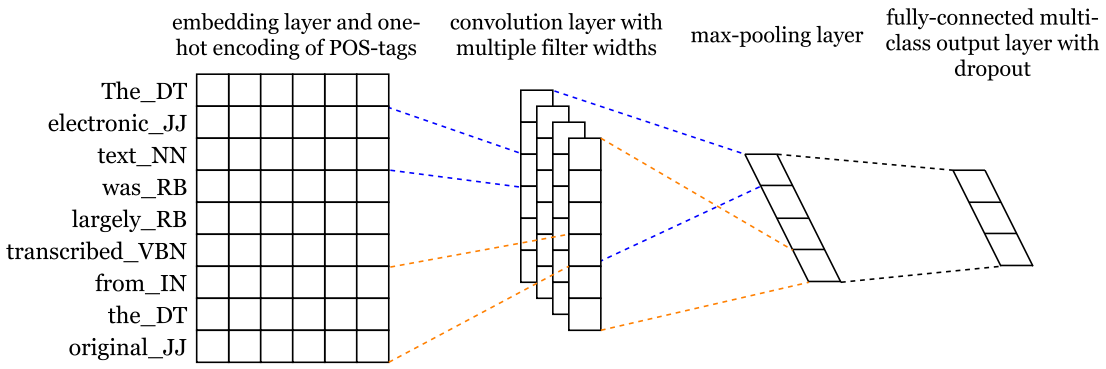

Figure 1: Architecture overview of the convolutional neural network.

dress).[3] We partitioned training, development and test data into segments of 1,500 words each, discarding any segments shorter than 1,500 words at the end of a publication. Authorship prediction is performed on the level of these segments. Table 1 gives an overview of corpus statistics.

	Publications	Segments
Training	1583	5360
Development	210	620
Test	117	323

Table 1: Corpus statistics for the ACL Anthology dataset.

For POS-tagging, we used the Stanford POS-tagger (Toutanova et al., 2003).[4] In addition to POS-tags, we use the pre-trained word embeddings available from Google[5] trained using the skip-gram objective (Mikolov et al., 2013) as input features for our convolutional neural network. Word frequencies were computed on the News Commentary and News Discussions English datasets provided by the WMT15 workshop.[6]

4 Experiments

For authorship prediction, we used a convolutional neural network (CNN) similar to that of Kim (2014). Each sentence is represented as a padded concatenation of word embedding vectors and POS-tag one-hot encodings. The network then applies a single layer of convolving filters with varying window sizes, and a max-over-time pooling layer which retains only the maximum value. The resulting features are passed to a fully-connected softmax layer to obtain a probability distribution over labels. Figure 1 gives an overview of the model architecture.

We used the implementation of Kim (2014),[7] which we modified in a number of ways. We used static channels only and did not modify the pre-trained word embeddings. Our input feature map contained not only the 300-dimensional word embeddings, but also a one-hot representation of POS-tags. We used 100 convolution filters of length 1, 2 and 3 words each and a batch size of 20 sentences. Like that of Kim (2014), our fully connected layer was trained with dropout. The dropout rate was set to 0.5 during training.

The network scans the entire input text of a segment using a sliding-window approach before applying max-pooling over time and making a prediction of authorship based on the prediction of the softmax layer. We tested the following frequency-cutoff settings:

1. Retain only the 1,000 most frequent words in our large, out-of-domain corpus of English, use their word embeddings as input features alongside a one-hot encoding of their POS-tags as predicted by the Stanford POS-tagger. Replace all other words with an unknown token. Generate a separate random embedding for each combination of the unknown token with a particular POS-tag and, in addition, retain the one-hot encoding of the POS-tags of

[3] As will become apparent, our procedures of obscuring low-frequency words would eliminate most author names anyway, this step is mainly taken to ensure fair comparison with the full-vocabulary baseline.

[4] Among the available models for English, we chose `english-left3words-distsim.tagger`.

[5] Available for download at https://code.google.com/archive/p/word2vec/

[6] Available for download at http://www.statmt.org/wmt15/translation-task.html

[7] Available for download at https://github.com/yoonkim/CNN_sentence

all unknown tokens.

2-4. Same as (1), but retain the 5,000, 10,000 and 50,000 most frequent words, respectively.

5. Retain all words and use their embeddings as input features, including a 1-hot encoding of their POS-tag. Generate a random word embedding for unknown words, as in Kim (2014).

Training was run for a maximum of 50 epochs. After each epoch, we measured the prediction accuracy on the development data. After training was complete, we tested the model parameters with the best development accuracy on the test data.

In addition to evaluating the authorship predictions of the model, we evaluate rank accuracies as well in order to investigate whether the models are able to reduce the list of possible authors for a segment to a short candidate list which contains the correct author. This can be achieved in a straightforward manner by simply sorting the activations of the softmax layer of the convolutional network for a test segment in order to obtain a ranked candidate list.

Our initial research hypothesis was that (1 - 4) would perform significantly worse than (5), while strongly outperforming a random baseline. This would demonstrate that authorship attribution (in a probabilistic sense) is possible based on stylometric features alone, but not to the same level of accuracy as when content clues are used as well.

5 Results

Table 2 gives an overview of the results for outright prediction of authorship. We find that at a frequency cutoff of 50,000 words, our system outperforms a setting in which the full vocabulary is used, while at lower frequency cutoffs performance is slightly reduced. It should be noted that all of our systems far outperform a random assignment of authors, which would be correct in approximately $\frac{1}{808}$ (0.12%) of cases. Performance in terms of accuracy for our best system is thus two orders of magnitude above random assignment.

Frequency Cutoff	Accuracy on DEV	Accuracy on TEST
1,000	11.94%	10.22%
5,000	16.61%	10.53%
10,000	16.45%	9.29%
50,000	15.00%	**13.31%**
None (Full Vocabulary)	15.16%	10.84%

Table 2: Prediction accuracies for the five frequency cutoffs on development as test data (ACL). The best result is marked in boldface.

Freq. Cutoff	$r = 1$	$r = 5$	$r = 10$	$r = 20$	$r = 50$
1,000	10.22%	17.34%	19.50%	26.32%	39.32%
5,000	10.53%	20.43%	26.93%	34.37%	46.75%
10,000	9.29%	20.12%	26.01%	32.20%	49.23%
50,000	**13.31%**	**24.46%**	**30.65%**	**39.32%**	**49.85%**
None	10.84%	19.20%	25.70%	32.82%	44.58%

Table 3: Rank accuracies for different ranks r on holdout test data (ACL). For example, a result of 24.46% at $r = 5$ means that for 24.46% of segments in the test data, the correct author was among the top-5 predicted authors of the model. Best results are marked in boldface.

For ranked prediction, a similar picture emerges. Table 3 gives an overview of results in this setting. At a frequency cutoff of 50,000 words, our model always outperforms the full-vocabulary baseline and lower frequency cutoffs. However, at higher ranks, there is a tendency for lower frequency cutoffs to outperform the full-vocabulary baseline as well, particularly at a cutoff level of 10,000.

6 Evaluation on Benchmark Dataset

In order to enable meaningful comparison of our models to other work, we additionally tested our approach on a commonly used benchmark dataset. We chose Task I of the PAN 2012 authorship attribution shared task,[8] which involves authorship attribution among a closed class of 14 novelists. The training data was again partitioned into segments of 1,500 words. The training procedure was identical to the one employed on the ACL Anthology dataset. We set aside 200 segments as development data, which left 1,694 segments for training. The test data comprised 14 novel-length texts. Prediction on the test data was performed on segments of a maximum length of 1,500 words, although we allowed for shorter segments at the end

[8] http://pan.webis.de/clef12/pan12-web/author-identification.html

of texts. For prediction on the text level, we simply aggregated segment-level predictions by majority vote. Results are summarized in table 4. Overall, we observed a similar effect as on the ACL Anthology dataset: The full vocabulary model performed much worse than models with a frequency cutoff. In contrast to the ACL Anthology dataset, the best results were achieved at a frequency cutoff of 1,000.

Frequency Cutoff	Acc. (Segments)	Acc. (Texts)	
1,000	**52.73%**	**78.57%**	11/14
5,000	50.91%	**78.57%**	11/14
10,000	49.90%	71.43%	10/14
50,000	51.82%	**78.57%**	11/14
None (Full Vocabulary)	48.08%	64.29%	9/14

Table 4: Prediction accuracies on PAN 2012, task I on segment and text levels for different frequency cutoffs. Best results are marked in boldface.

7 Discussion and Conclusions

While perhaps initially surprising, the fact that obscuring infrequent words helps system performance can be explained very well by better generalization: The absence of detailed content information may force the system to focus on stylistic features. All of our models achieved performances above 95% on the training data, demonstrating their large modeling capacity and thus their potential for over-fitting. At a frequency cutoff of 50,000 words, performance was improved on the test data, indicating that the model generalized better to unseen data.

In future work, we would like to include an other-class in order to make our setting even more challenging and realistic. We would also like to investigate which, if any, (automatic or manual) obfuscation techniques can be employed by authors to avoid de-anonymization with techniques similar to ours. Furthermore, we would like to investigate the relationship of authorship and native language identification on the ACL Anthology Reference Corpus, as many scientific publications are written by non-native speakers, which can be expected to influence the ease of authorship attribution on datasets of scientific publications.

8 Ethical Considerations

Our work demonstrates that convolutional neural networks have the potential to assign the correct author to very similar documents with somewhat remarkable accuracy well above chance. Although the performance of our particular system does not justify a use in legal or forensic settings, as more than 85% of predictions were still incorrect, the public should be made aware that stylistic features, in combination with modern natural language processing methods such as convolutional neural networks have significant potential to de-anonymize text, even when authors write about similar or related topics, and in an ostensibly factual, impersonal register. Since many people value their anonymity as authors, particularly when publishing text online, they should be made aware of the risk that current and future language technology holds for their ability to publish texts anonymously.

For the use of computational authorship attribution as part of historical research, reliable data about the accuracy of such methods is important to good scientific practice. Our work should thus be of interest to historians using such methodologies. In the future, as more powerful techniques are developed, more forensic uses of authorship identification may be justified. Policymakers, legal professionals and the public should have a realistic appraisal of the reliability of authorship identification as a technology in order to make informed judgments about if and when its use could be appropriate. Testing authorship identification technology in difficult, realistic settings such as the one of this work is important to tracking technological progress in this area and giving the public a realistic appraisal of the potential for use and abuse of computational authorship attribution.

References

Douglas Bagnall. 2015. Author Identification using multi-headed Recurrent Neural Networks—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*. Marrakech, Morocco.

Julian Brooke and Graeme Hirst. 2011. Native Language Detection with Cheap Learner Corpora. In

Conference of Learner Corpus Research. Louvain-la-Neuve, Belgium, LCR2011.

Heidi Byrnes and Katherine A. Sprang. 2004. Fostering advanced L2 literacy; A genre-based, cognitive approach. In Heidi Byrnes and Hiram H. Maxim, editors, *Advanced foreign language learning: A challenge to college programs*, Boston: Heinle Thomson, pages 47–85.

Zhenhao Ge, Yufang Sun, and Mark J. T. Smith. 2016. Authorship Attribution Using a Neural Network Language Model. *CoRR* abs/1602.05292. http://arxiv.org/abs/1602.05292.

Felix Golcher and Marc Reznicek. 2011. Stylometry and the Interplay of Topic and L1 in the Different Annotation Layers in the Falko Corpus. In *Quantitative Investigations in Theoretical Linguistics*. Berlin, Germany, QITL 4.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1746–1751.

M. Koppel, J. Schler, and S. Argamon. 2013. Authorship Attribution: Whats Easy and Whats Hard? *Journal of Law and Policy* 21(2):317 – 332.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119.

George K. Mikros and Eleni K. Argiri. 2007. Investigating Topic Influence in Authorship Attribution. In Benno Stein, Moshe Koppel, and Efstathios Stamatatos, editors, *SIGIR 07 Workshop Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*. PAN 2007.

Frederick Mosteller and David L. Wallace. 1964. Inference and Disputed Authorship: The Federalist. *Journal of the American Statistical Association* 58(302):275–309.

Dylan Rhodes. 2015. Author Attribution with CNNs. Technical report. http://cs224d.stanford.edu/reports/RhodesDylan.pdf.

Prasha Shrestha, Sebastian Sierra, Fabio A. Gonzalez, Manuel Montes y Gmez, and Thamar Solorio. 2017. Convolutional Neural Networks for Authorship Attribution of Short Texts. In *Proceedings of the EACL*. EACL, Valencia, Spain. https://www.aclweb.org/anthology/E/E17/E17-2106.pdf.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60(3):538–556. https://doi.org/10.1002/asi.v60:3.

Efstathios Stamatatos. 2017. Authorship Attribution Using Text Distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for the Computational Linguistics*. Valencia, Spain, EACL 2017.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. Overview of the Author Identification Task at PAN 2014. In Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*. Edmonton, Alberta, Canada.

G. Udny Yule. 1938. On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship. *Biometrika* 30:363–390.

George K. Zipf. 1932. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press.