Learning Universal Authorship Representations

Rafael A. Rivera Soto **Olivia Miano** **Juanita Ordonez** **Barry Chen**
Lawrence Livermore National Laboratory
{riverasoto1,miano3,ordonez2,chen52}@llnl.gov

Aleem Khan **Marcus Bishop** **Nicholas Andrews**
Johns Hopkins University
{akhan141,marcus.bishop,noa}@jhu.edu

Abstract

Determining whether two documents were composed by the same author, also known as *authorship verification*, has traditionally been tackled using statistical methods. Recently, authorship representations learned using neural networks have been found to outperform alternatives, particularly in large-scale settings involving hundreds of thousands of authors. But do such representations learned in a particular domain transfer to other domains? Or are these representations inherently entangled with domain-specific features? To study these questions, we conduct the first large-scale study of cross-domain transfer for authorship verification considering zero-shot transfers involving three disparate domains: Amazon reviews, fanfiction short stories, and Reddit comments. We find that although a surprising degree of transfer is possible between certain domains, it is not so successful between others. We examine properties of these domains that influence generalization and propose simple but effective methods to improve transfer.

1 Introduction

Authorship verification offers a useful capability for a number of problems, such as plagiarism detection, moderation of user-generated content, historical authorship attribution, and forensic analysis. Authorship verification techniques have traditionally relied on modeling stylometric linguistic properties, such as punctuation and whitespace usage and the frequencies of function words, parts-of-speech, and character n-grams (Stamatatos, 2009; Stolerman et al., 2014). More recently, end-to-end models built with convolutional neural networks (Shrestha et al., 2017; Andrews and Bishop, 2019) and recurrent neural networks (Boenninghoff et al., 2019) have been proposed. However, neural methods introduce a tradeoff: although they obviate the need of manual feature design by *learning* relevant features, these features are not explicitly identified. Not knowing how a system arrived at its authorship conclusions makes it difficult to assess how closely tied these features are to the domain from which the training data was drawn.

In addition, neural methods are less applicable in low-resource domains due to their requirement of large training datasets. On the other hand, if authorship features could be learned in a domain-independent fashion, it would reduce the need for in-domain training sets by exploiting *transfer* between domains: authorship representations could be learned from a large but out-of-domain corpus and subsequently deployed in a target domain. In prior work, Barlas and Stamatatos (2020) perform a study on cross-domain author verification in a closed world of 21 authors. In contrast, we consider an open-world setting with several orders of magnitude more authors. With a view towards addressing the deeper question of whether neural approaches are capable of learning *universal* authorship representations that are effective in unseen domains, we conduct a systematic study involving three disparate domains: Amazon reviews, fanfiction, and Reddit comments.

We propose the combination of an attention-based architecture and a contrastive training objective in §2 that achieves a new state-of-the-art in ranking performance. In §4.1 we conduct zero-shot transfer experiments among various domains, finding large discrepancies in transfer performance. In §4.2 we evaluate how properties of the training data affect transfer, including the number of training authors and the topic distribution of the data. In §4.3 we train models on the unions of multiple domains, in some cases resulting in improved generalization. Finally, in §4.4 we propose a simple masking technique that we find improves transfer. Overall, we find that neural models do *not* in general capture universal authorship features, and that the training data must be carefully controlled in order to learn the desired invariances.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 913–919
November 7–11, 2021. ©2021 Association for Computational Linguistics

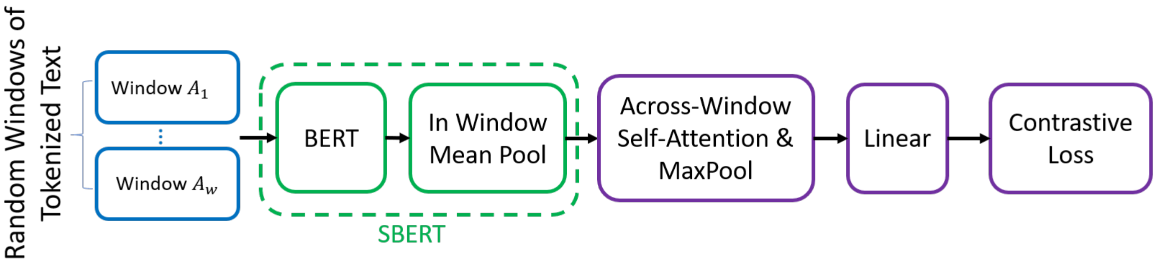

Figure 1: The proposed model for learning authorship embeddings consists of the aggregation of w text embeddings using SBERT and max-over-time pooling.

2 Model

Objective We learn a function f mapping a collection of documents to $\mathbb{R}^{512}$ under which any two collections composed by the same author have higher cosine similarity than two collections composed by different authors. We fit f using contrastive learning (Goldberger et al., 2004; Khosla et al., 2020), which proceeds as follows. Given a mini-batch[1] of collections x_i indexed by a set I, we denote the L_2 normalization of $f(x_i)$ by z_i and maximize

$$\sum_{i\in I} \frac{1}{|P(i)|} \sum_{j\in P(i)} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{k\in I\setminus\{i\}} \exp(z_i \cdot z_k/\tau)}$$

where $P(i) \subseteq I$ denotes the set of indices other than i of collections by the same author as x_i and τ is a temperature parameter, which we set to 0.01. This formulation includes several common metric learning losses as special cases, such as triplet loss (Schroff et al., 2015).

Architecture We propose the attention-based authorship verification model illustrated in Figure 1, an elaboration of the SBERT model developed for learning semantic sentence embeddings (Reimers and Gurevych, 2019). The following description assumes that the input is a collection of short documents, as in Andrews and Bishop (2019). If instead the input is a single long document, we regard it as a collection by subdividing into paragraphs.

At training time our model randomly samples excerpts consisting of $N = 32$ consecutive byte-pair encoded tokens (Kudo, 2018) from each of w contiguous documents, where w varies from 1 to 16 by taking $w = \lceil 1 + 15x \rceil$, where $x \sim \mathrm{Beta}(3, 1)$. Previous studies have shown that randomly sampling windows improves model generalizability and provides a simple way to handle very long documents (Boenninghoff et al., 2020), while Khan et al. (2021) show that sampling the number w of documents improves a model's ability to handle collections of documents of arbitrary size at inference time.

We pass the w excerpts to SBERT, which consists of a BERT model followed by an attention-weighted mean pooling of the features of each excerpt. This results in w feature vectors, each of dimension 768, the dimension of BERT's hidden feature. We apply self-attention to the w vectors, apply maxpooling, and finally project the result to $\mathbb{R}^{512}$ through a linear layer. We use a pretrained SBERT model (Reimers and Gurevych, 2019) but update all model parameters during training. Prior work has explored self-attention models for authorship attribution (Saedi and Dras, 2020; Fabien et al., 2020; Barlas and Stamatatos, 2020) with mixed success compared to simpler convolutional models. These systems have utilized either the output or the classification token of BERT as the basis for learning authorship embeddings. However, we find that SBERT's attention weighted averaging of the hidden activations leads to better performing authorship embeddings.

At test time we use more of the available text to evaluate the model. Namely, we set $w = 16$ for the Amazon and Reddit domains (see §3.1, §3.3). For the fanfiction dataset (see §3.2) each short story is regarded as a collection of documents by subdividing into paragraphs; at test time we take w to be the number of paragraphs, which essentially amounts to encoding the entire work.

3 Data

Our experiments use the three anonymous domains described below. For each domain we hold out a portion of the domain's data to be used to evaluate the performance of models trained on various

[1] We construct mini-batches of 256 collections by sampling two collections by each of 128 randomly chosen authors, which ensures that each mini-batch contains at least 128 pairs with matching authors.

combinations of the three domains. The remainder of the data is used for training. For domains that supply timestamps, we stipulate that the held-out evaluation data must be *future* to the training data in order to assess robustness to ephemeral aspects of writing, such as topic (Andrews and Bishop, 2019). The evaluation metrics described in §4 require that the evaluation corpora each be further divided into two disjoint sets, the *queries* and the *targets*. For domains that include timestamps, we also stipulate that all the targets be future to all the queries.

3.1 Amazon reviews

We restrict the corpus of Amazon product reviews described in Ni et al. (2019) to the reviews composed by the 135K authors contributing at least 100 reviews each. We organize the reviews by author and use those composed by 100K randomly chosen authors for training and those by the remaining 35K for evaluation, in which we take half the reviews by each author as queries and remainder as targets.

3.2 Fanfiction stories

Our fanfiction dataset is derived from the training set released with Bevendorff et al. (2020), which was collected by crawling `fanfiction.net`. The dataset consists of 278,169 stories by 41,000 distinct authors. Our evaluation split consists of stories by the 16,456 authors contributing exactly two stories each to the collection, one story to serve as the query and the other as the corresponding target. Our training set consists of the stories by the remaining authors.

3.3 Reddit comments

We use the corpus of Reddit comments collected by Baumgartner et al. (2020) for training, restricting to 1M authors contributing at least 100 comments each. Specifically, the training split is the same as in Khan et al. (2021). For evaluation we use the evaluation dataset proposed by Andrews and Bishop (2019), which is disjoint from and future to our training split. This is the largest of the three training sets used in this work, although we consider the effect of reducing the number of training authors in §4.2.

4 Experiments

Metrics We assess performance using two standard retrieval metrics. First, *recall-at-8 (R@8)* is the probability that, after ranking the targets according to their cosine similarity to a randomly selected query, the single correct target appears among the top eight results. We also report the *mean reciprocal rank (MRR)*, which does not require the choice of a fixed threshold. For both metrics, larger scores are preferable.

Baselines For comparison we also train the following baseline models using the same data. First, we consider the TF-IDF vector representation of the concatenated text content of a document collection. We use single words as tokens for this model. We also consider a variation of the proposed model with the SBERT component replaced with convolutions, as described in Andrews and Bishop (2019). An important distinction between the two neural architectures is that the convolutional model operates on fixed windows of nearby subwords, in contrast with the global perspective afforded by SBERT. For transfer, the limited expressiveness of the convolutional model may provide a helpful inductive bias, and therefore serves as a strong baseline.

Reproducibility Our source code and scripts to reproduce our main experiments are available at `https://github.com/noa/uar`.

4.1 Zero-shot transfer

For each domain, we train a domain-specific model and evaluate its performance on the held-out evaluation sets of all three domains. The results are shown in Table 1. Surprisingly, we see large variations in the transfer performance of the three models, with the model trained on Reddit and evaluated on the other two corpora achieving more than 80% of the R@8 performance of the respective domain-specific models, which is particularly notable in the zero-shot setting. On the other hand, models trained on Amazon or fanfiction perform consistently worse than the Reddit model on out-of-domain evaluations. We also remark that the proposed model outperforms the convolutional baseline model in terms of R@8 in all but one case. In particular, the in-domain Reddit results are better than those reported in Khan et al. (2021), establishing a new state-of-the-art.

4.2 Domain properties affecting transfer

The effect of the number of authors. In order to explore the extent to which the larger size of the Reddit dataset accounts for its excellent transfer performance, we train the proposed model on a

			Training Dataset					
			Reddit		**Amazon**		**Fanfic**	
			R@8	**MRR**	**R@8**	**MRR**	**R@8**	**MRR**
Evaluation Dataset	**Reddit**	**P**	65.61	50.37	23.72	15.43	12.10	7.64
		C	56.32	42.38	6.30	9.70	5.74	3.90
		T	10.34	6.77	7.65	5.03	6.97	4.63
	Amazon	**P**	68.91	55.59	82.54	68.99	28.91	20.06
		C	60.20	47.60	74.30	60.60	34.90	25.90
		T	43.70	35.50	31.61	24.86	21.46	16.45
	Fanfic	**P**	41.58	30.61	36.35	26.45	50.89	41.20
		C	40.66	30.98	24.99	17.98	47.98	39.02
		T	25.22	18.72	26.04	19.37	31.37	22.53

Table 1: Zero-shot transfer results for the proposed model (**P**), the convolutional model (**C**), and a TF-IDF baseline (**T**).

		Training Dataset		
		Amazon ∪ Fanfic	**Amazon ∪ Reddit**	**Fanfic ∪ Reddit**
Evaluation	**Reddit**	20.40	60.58	56.35
		13.11	45.46	41.20
	Amazon	79.60	84.84	53.51
		64.83	72.09	39.63
	Fanfic	51.84	40.69	57.51
		41.79	30.49	46.32

Table 2: Transfer performance with multi-domain training, where each cell shows R@8 above MRR.

subset of the Reddit corpus containing the posts of 100K randomly selected authors. This model obtains a R@8 of 36.7 when tested on the fanfiction dataset, which is marginally better than the 36.35 obtained by the model trained on Amazon, which also has 100K training authors. At the other extreme, we extend the Amazon corpus by relaxing the requirement that its authors contribute at least 100 reviews each. Training on the 250K authors contributing 75 reviews results in only a 1.1% improvement in R@8 on fanfiction, while training on the 500K authors contributing 50 reviews results in a 1.03% improvement over the previous model.

We conclude that the larger size of the original Reddit dataset partially explains its transfer performance. However, the converse is true only to a limited extent in the case of Amazon, where we see only marginal gains from increasing the number of authors. We now propose an explanation for this disparity.

The effect of topic diversity Reddit authors appear to write about a wider range of topics than authors of the other two domains. This hypothesis is borne out through the in-domain evaluations of Table 1 by the fact that the degree of improvement of the neural models **P** and **C** over **T** is greater for Reddit than it is for Amazon or fanfiction, noting that we regard **T** as a proxy for a topic model. This suggests that the Reddit model relies less on topic similarity to distinguish authors, something we expect to be beneficial for transfer, since there is little topical overlap among our domains.

4.3 Multi-domain training

Combining data from multiple domains is a natural way to increase the amount of training data available. Furthermore, requiring the model to *simultaneously* explain data from multiple domains may result in representations that generalize better.

Procedure For each pair of domains, we train a model by constructing mini-batches of 256 randomly selected document collections with examples split evenly between the two domains.[2] We assume that the authors contributing to the first domain are disjoint from those contributing to the second, so a pair of collections can share an author only if they are drawn from the same domain.

Discussion The results are shown in Table 2. Comparing with the single-domain results shown in Table 1, we find that introducing Reddit data always improves transfer over the Amazon and fanfiction domains alone. Conversely, introducing fanfiction always hurts transfer, while introducing Amazon improves transfer in one case but not the other. Furthermore, introducing Reddit to in-domain evaluations similarly improves performance. In fact, our best results on the Amazon and fanfiction domains come from this experiment by introducing Reddit data, with improvements of 2.3 and 6.62 points in R@8 compared to training on in-domain data alone. In summary, introducing the Reddit dataset is helpful for improving both transferability and in-domain performance, suggesting that domains that transfer effectively when used alone are also likely to improve in-domain performance when combined with data from another domain.

[2] We also considered more complicated sampling schemes, such as sampling collections according to the sizes of their respective datasets, but found these methods equally effective as sampling uniformly.

		Evaluation Dataset					
		Reddit		**Amazon**		**Fanfic**	
		R@8	**MRR**	**R@8**	**MRR**	**R@8**	**MRR**
Mask Prob	0.00	23.72	15.43	82.54	68.99	36.35	26.45
	0.05	23.23	15.10	81.46	67.50	35.68	26.00
	0.15	23.49	15.30	81.82	68.20	36.70	27.00
	0.30	25.11	16.60	82.65	69.50	**36.70**	**27.10**
	0.45	**26.36**	**17.50**	**82.65**	**69.50**	36.27	26.90
	0.60	26.31	17.41	81.06	67.84	34.38	25.33

Table 3: Impact on zero-shot transfer performance of random masking at training time.

4.4 Random masking data augmentation

Motivation Although we did not observe significant overfitting in our experiments, we did observe poor transfer with Amazon and fanfic as the source domains. This may be attributable in part to *underspecification*: the expressivity of the neural network results in fitting domain-specific predictive patterns instead of generalizable features of authorship (D'Amour et al., 2020). To improve transfer, we explore strategies to impede the learning of idiosyncrasies of the training domain. Specifically, we consider a simple domain-agnostic data augmentation strategy consisting of random subword dropout. If domain-specific features are masked with sufficient frequency, it may discourage reliance on those features and thereby improve transfer performance. Inspired by Stamatatos (2018) we also experimented with using word frequency to identify and mask only topic words, but found it difficult to identify suitable word frequency thresholds.

Procedure We take Amazon to be the source domain. At training time, we fix $p \in \{0.05, 0.15, 0.30, 0.45, 0.60\}$ and replace each subword of each Amazon review with the distinguished `<mask>` symbol with probability p in each training epoch. We use the standard SBERT subword tokenizer, which uses a byte-pair encoding (Sennrich et al., 2015). The resulting model is then applied to the unmasked, held-out document collections from all three domains.

Discussion The results are reported in Table 3. For each of the three evaluation domains, the best performance is obtained with the use of the proposed data augmentation. The effect is particularly notable with Reddit as the target domain, improving R@8 by 2.64 and MRR by 2.07. Nonetheless, the absolute performance remains low relative to in-domain performance, so random masking by itself is not sufficient to overcome domain-specific biases. Although we did not experiment with other source datasets, we expect random masking to also be effective in the multi-domain setting of §4.3. We also experimented with masking entire words rather than subwords, which resulted in very similar results.

5 Conclusion

Our main finding is that in general, neural authorship models are *not* universal, with models trained on some domains exhibiting significantly better transfer than those trained on others. Specifically, we find that Reddit exhibits consistently good transfer performance and improves performance when incorporated in multi-domain training, which we attribute to both topic diversity and the size of the dataset. In addition, the proposed attention-based model establishes a new state-of-the-art for the large scale, open-world setting (Andrews and Bishop, 2019; Khan et al., 2021) which to our knowledge, is the first instance where a self-attention model has been shown to consistently outperform simpler, yet highly effective convolutional-based architectures for authorship verification.

In future work, it would be interesting to attempt to explicitly disentangle the learned representations into content and style, for example, by using domain adversarial training objectives (Ganin et al., 2016; Bischoff et al., 2020). The data augmentation proposed in §4.4 represents a promising first step towards more sophisticated data augmentation methods. Additional data augmentation methods, particularly if applied together during training and incorporated into the loss calculation via multiple independent *views* of the data (Khosla et al., 2020), have the potential to further improve verification performance.

Acknowledgements

We thank the EMNLP reviewers for their helpful comments. Part of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DEAC52-07NA27344. This work was also supported by the Human Language Technologies Center of Excellence at Johns Hopkins University.

References

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1684–1695, Hong Kong, China. Association for Computational Linguistics.

Georgios Barlas and Efstathios Stamatatos. 2020. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations*, pages 255–266. Springer International Publishing.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel Pardo, Paolo Rosso, Guenther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. *Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection*, pages 372–383.

Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2020. The importance of suppressing domain style in authorship analysis. *arXiv preprint arXiv:2005.14714*.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 36–45. IEEE.

Benedikt Boenninghoff, Julian Rupp, Robert M. Nickel, and Dorothea Kolossa. 2020. Deep bayes factor scoring for authorship verification. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing*. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.

Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. 2004. Neighbourhood components analysis. *Advances in Neural Information Processing Systems*, 17.

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. 2021. A deep metric learning approach to account linking. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Chakaveh Saedi and Mark Dras. 2020. Siamese networks for large-scale author identification. *arXiv preprint arXiv:1912.10616v2*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Prasha Shrestha, Sebastián Sierra, Fabio A. González, Manuel Montes y Gómez, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *EACL*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2018. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and technology*, 69(3):461–473.

Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. 2014. Breaking the closed-world assumption in stylometric authorship attribution. In *IFIP International Conference on Digital Forensics*, pages 185–205. Springer.