Improved Authorship Attribution with a Combination of Neural and Large-Margin Contrastive Criteria

Hemanth Kandula[1], Haoling Qiu[1], Damianos Karakos[1] , Micheal Selvaggio[2], Hieu Man Duc Trong[3], Thien Huu Nguyen[3] and Brian Ulicny[1]
[1] Raytheon BBN Technologies, Cambridge, MA, USA
[2] Enolink, Cambridge, MA, USA
[3] Department of Computer and Information Science, University of Oregon, Oregon, USA

{hemanth.kandula,haoling.qiu,damianos.karakos,
brian.ulicny}@rtx.com
micheal.selvaggio@enolink.com
{thien,hieum}@uoregon.edu

Abstract. We present a high-performing authorship attribution system consisting of: (i) a novel fusion of two diverse contrastive learning approaches: neural supervised batch contrastive learning [1] and large-margin nearest neighbors learning [2]; (ii) embedding-level and ranking-level combinations of attribution outputs. The ensemble system results in large performance gains (of the order of 30% relative) compared to the best individual authorship attribution systems. The presented system had top performance in multiple official IARPA evaluations on authorship attribution.

Keywords: Authorship Attribution, System Ensemble, Contrastive Learning

1 Introduction

Authorship attribution (AA) determines which author, from a known set, wrote a given text, using features like stylistic nuances, syntax, and word choice. It plays a crucial role in forensic linguistics, criminal investigations, literary analysis, and digital content verification. The distinctive 'signature' of an author transcends lexical choices it can include deeper structures and even subconscious language patterns [1, 2]

The problem we tackle in this paper (which is one of the main tasks in the IARPA HIATUS project) is as follows: given a large corpus D of documents of known authorship, and given one or more documents written by a query author Q , the task is to come up with a ranking of the authors in D such that Q is among the highest-ranked authors. This is akin to the goal of Information Retrieval (IR), which rewards high-rank retrievals of relevant documents. We study two variations on this task: within-genre and cross-genre authorship attribution. The only difference between these conditions is whether all documents in D and Q are written in the same genre or not. As expected (and, as we present in the Results section) the cross-genre condition is much more challenging, as differences in genre can interfere with and even “mask” authorship style.

AA has traditionally focused on stylometric features such as n-grams, punctuation, whitespace usage, function word frequencies, part-of-speech tags. More recently, deep learning approaches, including convolutional neural networks (CNNs) and recurrent

neural networks (RNNs), have gained traction for their ability to learn complex patterns from large datasets [5–7]. These methods incorporate character and word embeddings, allowing for a more nuanced analysis of text. More recent approaches involving BERT-based models have been highly effective, particularly in handling large-scale datasets and diverse writing styles [7–10]. Various ensemble techniques have also been proposed in AA to improve performance and robustness. Some of these methods combine predictions from multiple classifiers, while others utilize voting mechanisms with classifiers like Random Forest and XGBoost [11][12].

In this paper, we propose an ensemble system that combines different attribution systems at both the embedding and ranking levels to improve AA performance. Our approach demonstrates a 32% relative gain over individual component systems. This ensemble approach employs unique learning criteria and training methods to effectively handle variations in document length, genre, and context. Our main contribution is a novel fusion of learning approaches: the batch contrastive learning criterion [1] and the Large Margin Nearest Neighbor (LMNN) criterion [2], which jointly transfrom the document embedding space for improved attribution performance. This novelty, along with a dual-level combination strategy, and a Multi-Similarity Miner to select challenging examples and GradCache for handling large batch sizes, distinguishes our ensemble system from other AA methodologies. Our ensemble system has proven highly effective in various AA scenarios, achieving top performance in multiple official IARPA evaluations on the HIATUS project. The paper is organized as follows: Section 2 outlines our system's structure and describes the methods used, and Section 3 presents the results obtained on within-genre and cross-genre datasets made available under the IARPA HIATUS program.

2 System Design

Our authorship attribution (AA) system employs an ensemble approach that inte-grates various individual AA subsystems. Please see Fig. 1 for a system-level diagram. The neural subsystems are based on a Siamese-BERT embedding architecture, with the subsystems trained using distinct methodologies and learning objec-tives. System training is conducted on a diverse range of datasets, with a focus on authors who have written at least 100 documents. The datasets include RedditMUD, Yelp Reviews, Amazon Product Reviews, and others. To create additional training samples, longer documents are split into smaller ones.

2.1 AA System I

AA System I is an extension of the Siamese-BERT-based architecture used within the LUAR framework [9], trained with a contrastive learning objective. We expand upon this system by implementing GradCache [15], a technique for handling much larger batch sizes than can ordinarily be handled by GPU memory constraints. This system uses a "single-domain" batch training approach, that is, all authors in each batch come from the same domain (Amazon Reviews, etc.). This encourages the model to distinguish between authors within the same domain rather than across different domains.

2.2 AA System II

AA System II follows the same contrastive learning architecture as AA system I, but using hard example selection for training. Two approaches are used to select challenging examples: at the input level and the representation level. At the input level, BM25 is employed to identify hard examples by examining keyword similarity. At the representation level, the Multi-Similarity Miner [16] selects challenging examples by comparing similarities between representations, filtering out easier pairs to focus on more informative ones.

2.3 LMNN Transformation

We utilized The LMNN (Large Margin Nearest Neighbor) approach [2] to work in conjunction the aforementioned neural attribution systems. LMNN "reshapes" embeddings, using Mahalanobis distance to determine similarity. This stage identifies top candidate authors through a retrieval phase and then uses LMNN to optimize the transformation matrix, minimizing squared error. This involves tuning the matrix to bring documents from the same author closer while pushing those from different authors further apart. This process helps the system effectively cluster documents from the same author while maintaining distinct separation between authors.

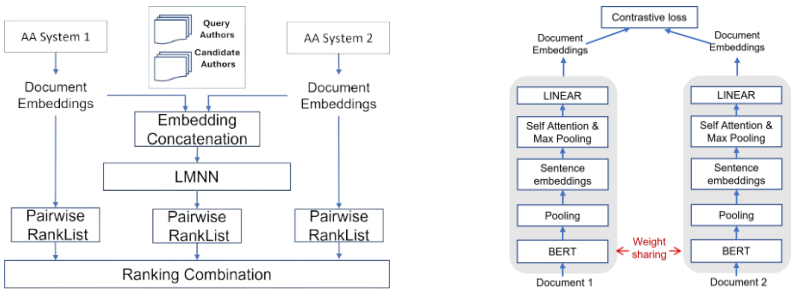

Fig. 1. (a) Ensemble Architecture for Authorship Attribution System (b) Siamese BERT

2.4 Combination of Embeddings

At the embedding level, the system concatenates embeddings from multiple Siamese-BERT models, taking advantage of their unique capture of diverse linguistic and stylistic features. These concatenated embeddings form a more discriminative representation, enhancing the system's ability to identify subtle authorship nuances.

2.5 Combination of Attribution Rankings

In the final stage, the system uses a ranking combination method, integrating pairwise similarity rankings from each AA subsystem and the LMNN output. We adopt a weighted Reciprocal Rank Fusion (RRF)[17], optimized through a grid search, to balance each model's influence. This ensemble approach leverages the individual strengths of each system, resulting in improved authorship attribution.

3 Results and Conclusions

To evaluate our authorship attribution (AA) system, we used Equal Error Rate (EER), a metric derived from Detection Error Tradeoff (DET) curves. DET curves plot false

positive rates against false negative rates for different threshold values in identifying authors. A lower EER indicates better performance. We tested our system using subsets of the IARPA HIATUS Research Set (HRS), released by IARPA for system performance analysis. This dataset contains five distinct genres: board game reviews from BoardGameGeek[1], Global Voices articles[2], Instructables articles[3], Stack Exchange Literature posts[4], and Stack Exchange STEM posts[5]. Each genre has a varying number of query documents and candidate authors, with an average of three documents per author. We tested our system in both per-genre and cross-genre conditions. In per-genre tests, all query and candidate documents are from the same genre; in cross-genre tests, they are from different genres, making attribution more challenging.

Table 1. Results of ensembling AA systems on HRS perGenre and CrossGenre datasets

	Method	WithinGenre	CrossGenre
Siamese BERT	AA System I	0.0333	0.1351
	AA System II	0.0239	0.1000
w/o LMNN	Embedding Ensemble	0.0206	0.0945
	+ Ranking Ensemble	0.0206	0.0818
with LMNN	Embedding Ensemble	0.0147	0.0743
	+ Ranking Ensemble	0.0146	0.0601

The ensemble system for authorship attribution (AA) shows notable performance gains across both within-genre and cross-genre conditions, as shown in Table 1. By integrating multiple AA subsystems, the LMNN and the dual-level combination strategy significantly enhances the system's capacity to discern subtle stylistic differences between authors, leading to considerable performance improvements compared to the best individual systems (38.9% relative gain in the within-genre condition, and 39.9% relative gain in the cross-genre condition).

This paper demonstrates the effectiveness of ensemble techniques in authorship attribution, resulting in top performance in official evaluations. The success of this approach (even in the difficult cross-genre condition) suggests that doing work on diverse AA and ensemble methods is crucial for obtaining state-of-the-art performance.

Acknowledgments. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

1 https://boardgamegeek.com/
2 https://globalvoices.org/
3 https://www.instructables.com/
4 https://literature.stackexchange.com/
5 https://academia.stackexchange.com/questions/tagged/stem

References

1. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised Contrastive Learning. In: Advances in Neural Information Processing Systems. pp. 18661–18673. Curran Associates, Inc. (2020).
2. Weinberger, K.Q., Saul, L.K.: Distance Metric Learning for Large Margin Nearest Neighbor Classification. Journal of Machine Learning Research. 10, 207–244 (2009).
3. Zheng, R., Qin, Y., Huang, Z., Chen, H.: Authorship analysis in cybercrime investigation. In: Chen, H., Zeng, D.D., Madhusudan, T., Miranda, R., Schroeder, J., and Demchak, C. (eds.) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 59–73. Springer-Verlag (2003). https://doi.org/10.1007/3-540-44853-5_5.
4. Holmes, D.I.: The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing. 13, 111–117 (1998). https://doi.org/10.1093/llc/13.3.111.
5. Ruder, S., Ghaffari, P., Breslin, J.G.: Character-level and Multi-channel Convolutional Neural Networks for Large-scale Authorship Attribution, http://arxiv.org/abs/1609.06686, (2016). https://doi.org/10.48550/arXiv.1609.06686.
6. Bagnall, D.: Author Identification using Multi-headed Recurrent Neural Networks, http://arxiv.org/abs/1506.04891, (2016). https://doi.org/10.48550/arXiv.1506.04891.
7. Fabien, M., Villatoro-Tello, E., Motlicek, P., Parida, S.: BertAA : BERT fine-tuning for Authorship Attribution. In: Bhattacharyya, P., Sharma, D.M., and Sangal, R. (eds.) Proceedings of the 17th International Conference on Natural Language Processing (ICON). pp. 127–137. NLP Association of India (NLPAI), Indian Institute of Technology Patna, Patna, India (2020).
8. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, http://arxiv.org/abs/1908.10084, (2019). https://doi.org/10.48550/arXiv.1908.10084.
9. Rivera-Soto, R.A., Miano, O.E., Ordonez, J., Chen, B.Y., Khan, A., Bishop, M., Andrews, N.: Learning Universal Authorship Representations. In: Moens, M.-F., Huang, X., Specia, L., and Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 913–919. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.70.
10. Tyo, J., Dhingra, B., Lipton, Z.C.: Siamese Bert for Authorship Verification. In: CLEF (Working Notes). pp. 2169–2177 (2021).
11. Custódio, J.E., Paraboni, I.: An Ensemble Approach to Cross-Domain Authorship Attribution. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Heinatz Bürki, G., Cappellato, L., and Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 201–212. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_17.
12. Abbasi, A., Javed, A.R., Iqbal, F., Jalil, Z., Gadekallu, T.R., Kryvinska, N.: Authorship identification using ensemble learning. Sci Rep. 12, 9537 (2022). https://doi.org/10.1038/s41598-022-13690-4.
13. Tennyson, M.F., Mitropoulos, F.J.: A Bayesian Ensemble Classifier for Source Code Authorship Attribution. In: Traina, A.J.M., Traina, C., and Cordeiro, R.L.F. (eds.) Similarity

Search and Applications. pp. 265–276. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-11988-5_25.
14. Czibula, G., Lupea, M., Briciu, A.: Enhancing the Performance of Software Authorship Attribution Using an Ensemble of Deep Autoencoders. Mathematics. 10, 2572 (2022).
15. Gao, L., Zhang, Y., Han, J., Callan, J.: Scaling Deep Contrastive Learning Batch Size under Memory Limited Setup. In: Rogers, A., Calixto, I., Vulić, I., Saphra, N., Kassner, N., Camburu, O.-M., Bansal, T., and Shwartz, V. (eds.) Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021). pp. 316–321. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.repl4nlp-1.31.
16. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019).
17. Cormack, G.V., Clarke, C.L.A., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 758–759. Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1571941.1572114.