Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5889–5893
Marseille, 20-25 June 2022

CCTAA: A Reproducible Corpus for Chinese Authorship Attribution Research

Haining Wang, Allen Riddell
Indiana University Bloomington
Bloomington, IN, USA
{hw56|riddella}@indiana.edu

Abstract

Authorship attribution infers the likely author of an unsigned, single-authored document from a pool of candidates. Despite recent advances, a lack of standard, reproducible testbeds for Chinese language documents impedes progress. In this paper, we present the Chinese Cross-Topic Authorship Attribution (CCTAA) corpus. It is the first standard testbed for authorship attribution on contemporary Chinese prose. The cross-topic design and relatively inflexible genre of newswire contribute to an appropriate level of difficulty. It supports reproducible research by using pre-defined data splits. We show that a sequence classifier based on pre-trained Chinese RoBERTa embedding and a support vector machine classifier using function character n-gram frequency features perform below expectations on this task. The code for generating the corpus and reproducing the baselines is freely available at `https://codeberg.org/haining/cctaa`.

Keywords: authorship identification, authorship attribution, stylometry, reproducibility, Chinese

1. Introduction

Authorship attribution attempts to infer the authorship of an unsigned, single-author document by analyzing candidate authors' writing style. It has wide applications in fields such as literary history, intellectual history, and online forensics. In this study, we propose a standard authorship attribution testbed for contemporary Chinese prose. The Chinese Cross-Topic Authorship Attribution (CCTAA) corpus features banning content shortcuts and supporting reproducibility.

1.1. Backgrounds

Problematic content shortcuts Recent work in authorship attribution has witnessed the introduction of a variety of new models, including models featuring deep neural networks. Although these models report impressive results, most of these models are evaluated using questionable test corpora. Often authors in test corpora tend to write about a limited set of topics. For example, a user, who we will refer to as "John", in the IMDb corpus may only comment on action movies. A model trained on both content and function words is virtually certain to learn an association between John writing a document and a document featuring action-movie-specific words. This association may be sufficiently strong—if, say, no other user writes about the particular kind of action movie John prefers—that a model may "learn" nothing about John's writing style that will facilitate authorship attribution of other unsigned documents by John. The model may work well on John's writings in the IMDb corpus but may break down when presented with a blog post by John reflecting on current events. Ideally, test corpora should be designed to prevent models from leveraging topical information. Particularly inappropriate are corpora where topic or role-specific words are *reliably* correlated with particular writers (e.g., the Enron email and IMDb corpora). Findings from recent papers illustrate the mentioned concern. Zhu and Jurgens (2021) found good performance when using only content words on *ad hoc* Amazon and Reddit corpora with a model based on pre-trained RoBERTa and BERT models. However, with a cross-topic task, Altakrori et al. (2021) reports inferior performance of pre-trained RoBERTa and BERT models relative to a support vector machine (SVM) classifier using common word n-gram frequency. The findings suggest topical information is indicative of one's identity in homogeneous corpora, but not as useful in topic-diverse corpora. A desirable testbed that encourages models relying only on topic-independent telltale signs should have training and testing samples from different topics.

Supporting reproducibility Using an appropriate corpus is the first step. It is also important that researchers be able to obtain—ideally at low cost—corpora and reproduce results of others. This requires the use of carefully-prepared, standard datasets with fixed train, validation, and test splits (Rendle et al., 2019; McFee et al., 2018; Stodden et al., 2014). Using cross-validation to measure a model's performance often frustrates reproduction efforts because splits are not recorded.

A corpus with fixed training, validation, and testing splits allow a transparent, fair comparison between proposed models. Reporting ad hoc data splits with random seeds is undesirable due to the dynamics of external dependencies (e.g., a particular random number generator) (Lin and Zhang, 2020). Using independently-derived pre-defined data splits—picked by a different researcher—also reduces the risk of an experiment overstating results by using a "lucky" split. Second, fewer lessons can be learned if models are compared with different datasets or with different versions of the same dataset (Bittner et al., 2019). Making sure datasets are identical is integral to reproducible research.

1.2. Contribution

We present a standard testbed for authorship attribution on contemporary Chinese prose that has appropriate level of difficulty and meets the demands of reproducible research. Our contributions are listed below.

1. The CCTAA corpus is the first standard testbed for authorship attribution on contemporary Mandarin prose.

2. The CCTAA corpus is designed to encourage the development of models which focus narrowly on topic-independent writing style.

3. The CCTAA corpus supports reproducible research with fixed data splits.

The rest of the paper is organized as follows. In Section 2, we describe the corpus in detail. In Section 3, we explain how we organize and preprocess the corpus. We then run two baseline models on the corpus in Section 4. Finally, we briefly discuss the implications and suggest the future directions in Section 5.

2. Corpus Description

The CCTAA corpus contains single-author newswire articles using simplified Chinese characters from 500 reporters affiliated with the Xinhua News Agency. Each author appears in all three splits, upholding the closed-set assumption of many authorship attribution models. For training, every author contributes multiple passages which consist of one or more paragraphs, with cumulatively no fewer than 5,000 characters. Authors have exactly one sample in the validation and testing sets. Examples in the two sets have more than 400 characters. See the corpus summary in Table 1.

Split	Number of Authors	Characters per Author (s.d.)	Passages per Author (s.d.)
Train	500	5305 (247)	11 (2)
Validation	500	460 (208)	1 (0)
Test	500	471 (226)	1 (0)

Table 1: Corpus summary. Character count does not include spaces.

Six topics are found in all splits: international news, culture, entertainment, financial, political news, and sports. (We describe the topic classification procedure in section 3.2.) See the topic distribution across splits in Figure 1. Importantly, with a cross-topic design, a candidate's training topic(s) *will not* appear in their validation or testing examples, as the CCTAA's title advertises. The topic of validation and testing examples for an author may or may not be the same, though the topic distribution in the two data sets are very similar. Further, reports associated with the entertainment topic are comparatively scarce in the training examples; models will need to correctly predict the authorship of entertainment news articles (in validation and test sets) in order to perform well.

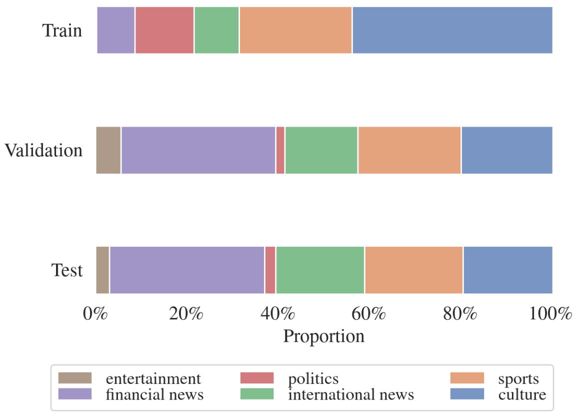

Figure 1: Topic distribution across data splits. The topic distribution in the validation and testing sets are similar but different from that found in the training samples. The validation and testing sample of the same reporter may or may not be identical.

Note, the location of news events is commonly associated with specific reporters. (That is, reporters' "beats" are often location-specific.) We found reporters from Xinhua News Agency local branches tend to report provincial news (i.e., news occurring in a particular city or region). The correlation between identity and geography may allow topic-relying models to cheat via geography-related words, which hinders a models' generalization on style. The correlation issue is addressed by manually excluding the co-occurrence of provincial location between training and the other splits for half of the reporters. For instance, for a reporter with a large portion of news reported happened in Jinan (the capital of Shandong Province), the reporter's validation and testing samples have ca. 50% chance of containing news from Shandong province.

3. Corpus Organization

3.1. Xinhua Newswire

The corpus is extracted from the Chinese Gigaword Second Edition (Gigaword-2E) (Graff, David and Chen, Ke and Kong, Junbo and Maeda, Kazuaki, 2005) published by the Linguistic Data Consortium (LDC). Gigaword-2E includes written newswire articles published by Xinhua News Agency from the year 1991 to the year 2004. As the largest state news agency, Xinhua covers diverse topics, from politics, financial, to sports and entertainments. Typical Xinhua news articles feature formal language use. Most of the news articles are short, of ca. 400 characters. The genre is relatively rigid compared to common documents found in literary and historical forensic scenarios. But making the task more challenging will do no harm in the present context.

3.2. Preprocessing

Only coherent reports on a particular topic or event, namely the "story" type in the Gigaword-2E, are included. Heuristics are used to extract reporters who published at least ten single-author documents and contribute at least 10,000 to 100,000 characters in aggregate (after removing texts shorter than 100 characters).[1]

Name Duplication Authorship attribution identifies individuals through their writing style. A good corpus for such purposes should not be polluted by writings from multiple authors who happen to have the same name (e.g., share family and given names). We therefore seek to minimize the problem of duplicate names. First, the reporters who have the same name as others were removed after consulting an official list of reporters found on the Xinhua website.[2] Second, we removed reporters that have popular given names at the time (e.g. "建国" and "芳").

All immediate author information is then removed from the text field for each sample. The Gigaword-2E corpus is well formatted: the byline is wrapped in parentheses following the headlines and datelines. In essence, we remove everything before the closing parenthesis of the author field. Otherwise we leave everything as is for a better reproduction of the corpus, including many newline markers ("`\n`") that researchers may want to remove.[3]

Topic Classification We apply a Chinese RoBERTa Model (Liu et al., 2019) already fine-tuned on the Chinanews corpus (Zhao et al., 2019; Zhang and LeCun, 2017).[4] We reuse the topic classes of Chinanews and merged its "mainland China politics" and "Hong Kong - Macau politics" into "politics" due to the similarity of the topics. The headline and the first sentence of each news item are *separately* labeled with the model. Articles where the predicted label disagrees are removed. Finally, predicted topics are manually reviewed. Incorrect classifications are removed. Only a small portion of the labels are incorrectly predicted and have been corrected by two human judges.

An example article Every sample in the CCTAA corpus contains the fields of `id`, `split`, `author`, `topic`, and `text`. The author and text fields are in simplified Chinese, otherwise in English. `id` field matches exactly the identifiers found in the Gigaword-2E corpus. An example is shown below.

> `id:` `XIN_CMN_20030520.0191`
> `split:` `Train`
> `author:` 记者朱少华
> `topic:` `International News`
> `text:`
>
> 沙特阿拉伯外交大臣费萨尔20日在此间表示,沙特希望联合国安理会能够一致通过修改后的取消对伊拉克制裁决议草案。
>
> 费萨尔当天下午在与到访的丹麦外交大臣默勒会谈后对新闻界说,现在伊拉克迫切需要的是恢复社会治安,取消制裁。他说,沙特希望安理会各方在讨论与战后伊拉克局势有关的所有问题时都能完全达成一致。
>
> 费萨尔4月中旬曾表示,沙特暂时不赞成取消对伊拉克的制裁,原因是安理会取消制裁应当有一个程序问题,而且现在伊拉克没有合法政府,由谁来代表伊拉克接受新的决议也是个问题。

We released Python scripts to help reproduce the CCTAA corpus from the Gigaword-2E corpus and compute two simple baselines. A check on the checksum of a newly created CCTAA will be performed to make sure the corpus is distributed identically. The scripts are hosted at `https://codeberg.org/haining/cctaa`.

4. Baseline

We run two baseline models on the CCTAA corpus.

4.1. SVM

We adopt a naïve model using a linear SVM as the algorithm and function character n-gram frequency as the feature set. We choose function character n-grams for features because they are typically free of obvious meaning (e.g. "倘若", English for "if") and have shown to be useful (Kestemont, 2014; Koppel et al., 2006) in ascribing authorship of Chinese prose (Wang et al., 2021; Zheng et al., 2006). We reused a function character n-gram list transcribed from a Chinese function word dictionary (Wang, 1998).[5] The feature set has 819 common function character n-grams found in classical and modern Chinese, including 262 unigrams, 545 bigrams, ten trigrams, and two quadgrams. Spaces

[1] We only consider articles from "reporters" ("记者"). Articles from "correspondents" ("通讯员") adjunct to Xinhua are excluded. Correspondents contribute only a small fraction of the news. They often represent voices from other institutions, making their writings less likely to fit the single-author assumption. Also, short text authorship attribution is not within our scope.

[2] `http://www.xinhuanet.com/reporter/list1.htm`

[3] The Gigaword-2E corpus has wrapped lines for every ca. 30 characters with newline markers that we left intact.

[4] The pre-trained model is provided by the Hugging Face (`https://huggingface.co/uer/roberta-base-finetuned-chinanews-chinese/tree/main`).

[5] The list is provided by the `functionwords| library (v.0.8)` on PyPI (`https://pypi.org/project/functionwords/`).

and content between double quotation marks are removed using heuristics before feature extraction.
SVM is a familiar classification technique. It is chosen for its simplicity—we anticipate others will be able to reproduce our results exactly. We use a linear multi-class SVM with a penalty parameter $C = 1.0$. Each document's feature vector is normalized by the sum of its elements. Features are then standardized by dividing by feature standard deviations after deducting the means. For the consideration of reproduction, this model is the only model whose results we suggest other researchers attempt to reproduce exactly.

4.2. Pre-trained RoBERTa

We use a RoBERTa sequence classification model (Liu et al., 2019; Zhao et al., 2019) pre-trained on Chinese corpora as the second baseline model. The final layer of the model is fine-tuned using the CCTAA training samples.[6] The model is chosen due to its popularity in Chinese sequence classification.

4.3. Results

Table 2 shows the proportion of correct prediction over all predictions ("accuracy") of the baseline models on the CCTAA corpus. Both models beat random guessing (0.2%). The RoBERTa achieves a better testing accuracy of 18.0%. The SVM model performs comparatively poorly.

Model	Test Accuracy
Linear SVM	3.0%
RoBERTa	**18.0%**
chance	0.2%

Table 2: Running two baseline models on CCTAA. Chance stands for random guessing. The best performance is marked in bold.

4.4. Discussion

The accuracy of the SVM model that is known to work on other Chinese corpora (Riddell et al., 2021; Wang et al., 2021) is far lower than expected. None of the models can be evaluated as 'useful' in real-world applications (where having high confidence in the correct author is desired). The findings challenge the expectation of excellence performance one might have after reading recent research on authorship attribution. This may reflect some feature peculiar to newswire articles or to the cross-topic design of the CCTAA corpus.

5. Conclusions and Future Work

In this paper, we introduce the CCTAA corpus which has been designed to evaluate Chinese language authorship attribution models. It is, to our knowledge, the first standard testbed for contemporary Mandarin prose that supports reproducible research. The CCTAA corpus has been carefully constructed to encourage models to use only non-topical style information—in keeping with the traditional aims of authorship attribution research. We document relatively weak performance of two well-known models, confirming that the task poses a challenge.

6. Acknowledgements

We thank to Sandra Kübler, Zuoyu Tian, and Yujing Wang for resources, suggestions, and assistance with corpus preprocessing.

7. Bibliographical References

Altakrori, M. H., Cheung, J. C. K., and Fung, B. (2021). The topic confusion task: A novel scenario for authorship attribution. *arXiv preprint arXiv:2104.08530*.

Bittner, R. M., Fuentes, M., Rubinstein, D., Jansson, A., Choi, K., and Kell, T. (2019). mirdata: Software for reproducible usage of datasets. In *ISMIR*.

Kestemont, M. (2014). Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66.

Koppel, M., Akiva, N., and Dagan, I. (2006). Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology*, 57(11):1519–1525.

Lin, J. and Zhang, Q. (2020). Reproducibility is a process, not an achievement: The replicability of ir reproducibility experiments. *Advances in Information Retrieval*, 12036:43.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

McFee, B., Kim, J. W., Cartwright, M., Salamon, J., Bittner, R. M., and Bello, J. P. (2018). Open-source practices for music signal processing research: Recommendations for transparent, sustainable, and reproducible audio research. *IEEE Signal Processing Magazine*, 36(1):128–137.

Rendle, S., Zhang, L., and Koren, Y. (2019). On the difficulty of evaluating baselines: A study on recommender systems. *arXiv preprint arXiv:1905.01395*.

Riddell, A., Wang, H., and Juola, P. (2021). A call for clarity in contemporary authorship attribution evaluation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1174–1179.

Stodden, V., Leisch, F., and Peng, R. D. (2014). *Implementing reproducible research*. CRC Press.

Wang, H., Xie, X., and Riddell, A. (2021). The challenge of vernacular and classical Chinese cross-register authorship attribution. In *Proceedings of the*

[6] The pre-trained model can be found on the Hugging Face (`https://huggingface.co/uer/chinese_roberta_L-12_H-768`).

Conference on Computational Humanities Research 2021, pages 299–309.

Wang, Z. (1998). *Modern Chinese Dictionary of Function Words*. Shanghai Lexicographical Publishing House.

Zhang, X. and LeCun, Y. (2017). Which encoding is the best for text classification in chinese, english, japanese and korean? *arXiv preprint arXiv:1708.02657*.

Zhao, Z., Chen, H., Zhang, J., Zhao, X., Liu, T., Lu, W., Chen, X., Deng, H., Ju, Q., and Du, X. (2019). Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). 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(3):378–393.

Zhu, J. and Jurgens, D. (2021). Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. *arXiv preprint arXiv:2109.03158*.

8. Language Resource References

Graff, David and Chen, Ke and Kong, Junbo and Maeda, Kazuaki. (2005). *Chinese Gigaword Second Edition LDC2005T14*. Linguistic Data Consortium.