The Challenge of Vernacular and Classical Chinese Cross-Register Authorship Attribution

Haining Wang[1], Xin Xie[2] and Allen Riddell[1]

[1] *Indiana University Bloomington, Bloomington, Indiana, USA*
[2] *Shanghai Normal University, Shanghai, China*

Abstract

Ming-Qing fiction is widely regarded as the pinnacle of classical Chinese literature, but over three-quarters of vernacular fictional works were anonymously or pseudonymously composed, frustrating literary-historical research. To begin to address the problem, we propose a cross-register authorship attribution task: recover the authorship of a vernacular Chinese text given classical Chinese writing samples of known authorship. A corpus of eight authors known to have written in both registers was assembled to serve as a testbed. We describe the performance of models using different sets of function character/word frequencies as input features. This standard approach to authorship attribution performs well in the same-register setting but poorly in the cross-register setting. We discuss the degree of vernacularization and the amount of dialog in texts as key factors contributing to the low cross-register accuracy.

Keywords

authorship attribution, Ming-Qing fiction, classical Chinese, vernacular Chinese

1. Introduction

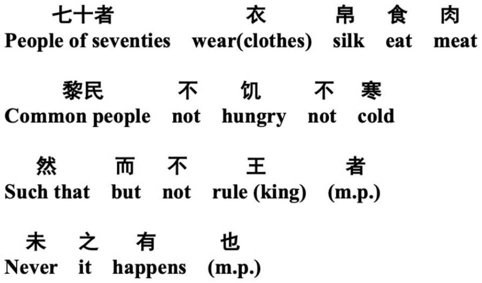
1.1. Backgrounds

Fictional works produced during the Ming (1368 –1644) and Qing (1644 –1912) dynasties are collectively referred as Ming-Qing fiction. Ming-Qing fiction written using vernacular Chinese is widely regarded as the pinnacle of classical Chinese literature. At the time, however, classical Chinese was the privileged register and composing in vernacular Chinese was regarded as unorthodox. For this reason, the authorship of most Ming-Qing vernacular fiction works, including numerous masterpieces, is in question. In one bibliography of Ming-Qing fiction, over three-quarters (ca. 396 in 513) vernacular fiction works were written anonymously or under a pseudonym [23]. The *Golden Plum Vase* (金瓶梅) and the *Marriage Destinies to Awaken the World* (醒世姻缘传) are perhaps the most famous cases. The authorship of the anonymous and pseudonymous Ming-Qing fictional works has puzzled scholars for more than a century.

By analyzing the writing style of candidate authors, authorship attribution enables inferences about the likely author of a text of unknown authorship. Typically, researchers begin by finding, for each candidate author, writing samples that resemble the disputed text in terms of

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

hw56@indiana.edu (H. Wang); rwxiexin@shnu.edu.cn (X. Xie); riddella@indiana.edu (A. Riddell)
https://www.hainingwang.org/ (H. Wang); https://www.ariddell.org/ (A. Riddell)
0000-0002-1196-0918 (H. Wang); 0000-0002-0039-851X (X. Xie); 0000-0002-4967-0879 (A. Riddell)



七十者 衣 帛 食 肉
People of seventies wear(clothes) silk eat meat

黎民 不 饥 不 寒
Common people not hungry not cold

然 而 不 王 者
Such that but not rule (king) (m.p.)

未 之 有 也
Never it happens (m.p.)

Figure 1: An excerpt from the *Mencius* (in classical Chinese) for illustration. A verbatim English translation is annotated below the Chinese. A modal particle is abbreviated as "m.p.".

the previously mentioned factors. In practice, candidate authors are usually already provided to researchers thanks to the labor of literary and cultural historians. The challenge lies in matching writing styles. Numerous factors reportedly influence writing style, including genre [17, 8, 14], topic [15, 16, 9], gender [6, 13], period [5, 1], and culture [3]. Typically researchers begin by finding, for each candidate author, writing samples which resemble—in terms of the previously mentioned factors—the disputed text. Ming-Qing vernacular fiction poses a particular challenge here: candidate authors tended not to sign any vernacular works. In most cases, the texts we have available were written in classical Chinese.

1.2. Classical Chinese, Vernacular Chinese, and Cross-Register authorship attribution

Classical Chinese can be understood as preserving the grammar and semantics of Chinese as it was used before the Qin period (i.e., before 221 BCE). Classical Chinese predominates in official texts and texts written by members of the educated class throughout the imperial period (221 BCE - 1912 CE) [20]. For example, most of official documents were composed using classical Chinese.

Written vernacular often emerged from written dialog in classical works. This way of writing developed into various genres. For example, *Bianwen* (变文) paraphrases canonical Buddhist texts using speech-like writing. And *Huaben* (话本) describes actors' movements and scripts when performing. It was not until the middle sixteenth century before vernacular written Chinese became a recognized literal register [4].

The differences between the two versions of written Chinese are considerable. First, the classical lexicon tends to use single characters, while vernacular words often use pairs of characters. Take an excerpt from the *Mencius* as an example (Figure 1). The *Mencius* is a classic of Confucianism composed in classical Chinese. In the example, "饥" (being hungry) is used in isolation, but in vernacular is expected to be collocated with "饿" ("饥饿") to express the same meaning. Second, classical has more frequent part-of-speech ambiguity. "衣" (clothes) is a noun in vernacular most of the time, but it functions as a verb when used before "帛" (silk), meaning "wear." Third, word order in the classical register is more variable. In most cases, Chinese uses subject-verb-object order. In the classical clause "未之有也", "之" is the object and appears before the linking verb "有". This order is unconventional in vernacular Chinese.

Frequently, especially during the Ming and Qing periods, the boundary between vernacular and classical Chinese is not clear. Many texts mix the two registers in various ways. For example, dialog in classical texts often resembles the vernacular equivalent. Vernacular fiction also has a tradition of opening and closing a chapter with classical verse. The boundary blurred further when classical grammar was mixed with the vernacular lexicon at the end of the imperial period. In addition, written Chinese shares the same convention of having no obvious break markers (corresponding to punctuation) between "sentences" and delimiting paragraphs with empty spaces after the ending of a previous paragraph.[1]

In this study, we consider the task of cross-register authorship attribution using texts of known authorship. We aim to infer the authorship of a vernacular text given classical texts by the same author. Developing reliable cross-register authorship attribution techniques will be required to resolve the long-standing debates about disputed authorship of vernacular fictions, such as the *Golden Plum Vase* and the *Marriage Destinies to Awaken the World*, and roughly four hundred works as we found in our survey.

2. Corpus

2.1. Description

We assemble a corpus of eight authors known to have written in both registers. All authors lived between 1570 and 1870. All but one are from southern China, and all authors are men.[2] The imbalance of gender and region reflects relevant social and economic circumstances of the period.[3]

The corpus contains 4.2 million characters of fictional and non-fictional prose, although these genres are not always distinct.[4] Topics are diverse, including jokes, the care of pregnant women, history, opera commentary, war diaries, and personal reflections. All works address a general audience.

2.2. Collecting and Preprocessing

In practice, only authors who have written in each register and whose surviving works have machine-readable editions are considered.[5] We refrain from picking texts that are disputed or use rhyme. We also avoid texts which are revisions of pre-existing texts or mixtures of vernacular glosses with classical grammar. Table 1 shows the texts in the corpus.

After downloading all texts, we performed the following preprocessing:

1. We checked all texts for flaws and missing parts by consulting other digital editions and print editions.[6] If a character in a text lacks a Unicode code point, we used the modern

[1] Readers familiar with the evolution of Latin may gain some appreciation of how the registers differed by considering the lexical and syntactic differences between Classical Latin (75 BCE to 300 CE) and Modern Latin (ca. 1500 - 1900 CE). The analogy is not exact, of course.

[2] We spent 20 hours searching for woman authors to include, but we were unable to find an author with available texts. We gathered the candidate authors by consulting two bibliographies of Ming-Qing novels[23, 7]. We welcome suggestions for candidates to include in a future, expanded version of the corpus.

[3] At the time, education for women received less attention. And, southern China, such as Zhejiang and Jiangsu, was relatively wealthier and developed.

[4] Distinguishing between fictional and non-fictional historical narratives is difficult or, in some cases, impossible.

[5] The existence of an edition in a machine-readable format usually indicates a canonical author. This introduces a bias towards authors who were well known at the time or who subsequently became well known.

[6] If multiple versions exist, we choose the one which has fewer characters missing or obvious errors.

variant. We refer to zdic.net and Unicode*pedia* to check whether a character falls outside of the Chinese Unicode set (between \u4e00 and \u9fff). We keep outside characters if they are valid Chinese characters or punctuation, deleting them otherwise.[7] Characters are rarely deleted.

2. We remove title, heading, table of contents, preface, postscript, and editorial comments, as well as rhymed verse and prose if they are not part of the main body. Blank lines, redundant spaces, and indentation marks are removed too.
3. All the texts are automatically converted into UTF-8 encoded simplified Chinese using the Python package "hanziconv" (v.0.3.2).
4. Texts are then segmented into roughly 1,000-character chunks without breaking clause-level structures. Modern publishers punctuate ancient Chinese texts, which originally did not have punctuation. We leverage these delimiters introduced by editors to avoid breaking clauses when segmenting.
5. We eliminate all punctuation in the next step to restore the original formatting.
6. For authors who have only one work in a register, we split the work into two parts.[8]

After performing these steps, we organize all chunks by register under each candidate's directory with informative file names.

The corpus consists entirely of texts in the public domain and is available at https://zenodo.org/record/5513043.

3. Method

3.1. Feature Set and Algorithm

Given our research question—Is it possible to recover the authorship of a vernacular text using classical texts as training data?— we choose function words/characters for features because they have been shown to be useful stylistic markers [22, 24]. We transcribed two published lists of function words—one for classical and the other for modern—as the feature sets.[9] The classical feature set contains 479 function characters [21]; the modern feature set has 819 function words (262 character unigrams, 545 bigrams, 10 trigrams, 2 tetragrams) [18].[10] We also use a feature set which is the union of the two feature sets (the "combined" feature set).

A linear support vector machine (SVM) is chosen as the classifier. We use LIBSVM's implementation [2], wrapped by scikit-learn [10]. We use the default cost parameter ($C = 1.0$). Features are standardized by dividing by feature standard deviations after deducting the means.[11]

3.2. Setup

The goal of our cross-register task is to recover the authorship of a vernacular text based on classical writing samples. To this end, we set up two experiments. In the first experiment, we

[7] Some valid Chinese characters fall outside of the aforementioned Unicode range.

[8] We do this to prevent severe inflation in calculating same-register accuracy. See justification for this treatment in the Appendix.

[9] Chinese vernacular function words overlap heavily with modern Chinese's. Also, there is no function word dictionary built for vernacular Chinese specifically, to the best of our knowledge. The modern function word list [21] is particularly comprehensive.

[10] We released a Python package ("functionwords") on PyPI to help others use these lists.

[11] A pilot study shows a standard logistic regression with L2 regularization (regularization parameter 1.0) achieves similar accuracy as SVM [19].

Table 1
Corpus Description

Author	Title	Register	Genre	Length
Feng Menglong 1574 - 1646	*Yandu Diary*	classical	non-fiction	13,752
	Smart Ideas Pandect	classical	fiction	43,376
	Common Words to Alert the World	vernacular	fiction	329,382
	Eternal Stories to Awaken the World	vernacular	fiction	457,017
	Illustrious Words to Instruct the World	vernacular	fiction	317,583
	Zhan GuoYing Romantic Story	vernacular	fiction	65,839
Ding Yaokang 1599 - 1669	*History of Order*	classical	non-fiction	61,908
	Golden Plum Vase, A Continuation	vernacular	fiction	295,103
Qi Biaojia 1603 - 1645	*Opera Review from Yuanshan Studio*	classical	non-fiction	10,120
	Qu Review from Yuanshan Studio	classical	non-fiction	23,245
	Diary of 1644 & 1645	vernacular	non-fiction	54,755
Li Yu 1611 - 1680	*Leisure Tales*	classical	non-fiction	132,681
	The Carnal Prayer Mat	vernacular	fiction	81,833
	Twelve Mansions	vernacular	fiction	123,036
	Silent Play	vernacular	fiction	110,827
Chu Renhuo 1635 - ?	*Hard Gourd Collection*	classical	non-fiction	604,762
	Romance of the Sui and Tang Dynasties	vernacular	fiction	530,065
Wu Jingzi 1701-1754	*Wenmu Collection*	classical	non-fiction	21,881
	The Scholars	vernacular	fiction	273,546
Du Gang ca.1742 - ca.1800	*Romance of the Northern Dynasties*	classical	fiction	217,164
	Romance of the Southern Dynasties	classical	fiction	160,125
	Amusing and Awakening	vernacular	fiction	109,196
Zhang Yaosun 1808 - 1863	*Pregnancy & Childbirth, A Revision*	classical	non-fiction	25,407
	Dream of the Red Chamber, An unfinished Twenty-Chapter Complement	vernacular	fiction	141,440

Note: The length field indicates the character count after clearing all punctuation. The standalone rhymed prose and verse in *Hard Gourd Collection* and *Wenmu Collection* are pruned.

consider authorship attribution in a scenario when a reasonable amount of classical training text is available. In the second experiment, we consider a scenario in which extensive classical training material is available.

We do not have a golden rule for how many classical Chinese characters constitute a large amount of training material. For English language authorship attribution, Rao, Rohatgi, et al. [12] recommends around 6,500 English words as adequate. We estimate the corresponding character count in classical Chinese by counting English words used in the first ten stories of Herbert Giles' translation of *Strange Stories from a Chinese Studio* (聊斋志异) [11] (ca. 14,190 words) and Chinese characters of the corresponding plots in the original work (ca. 9,350 Chinese characters). The English-word-to-Chinese-character ratio is roughly 1.5. We finally decide to fit the model with ca. 4,000 classical characters from each author for the "limited data" scenario. In the other experiment—the "abundant data" scenario—we give the model more training material., ca. 10,000 classical characters. For both settings, we evaluate

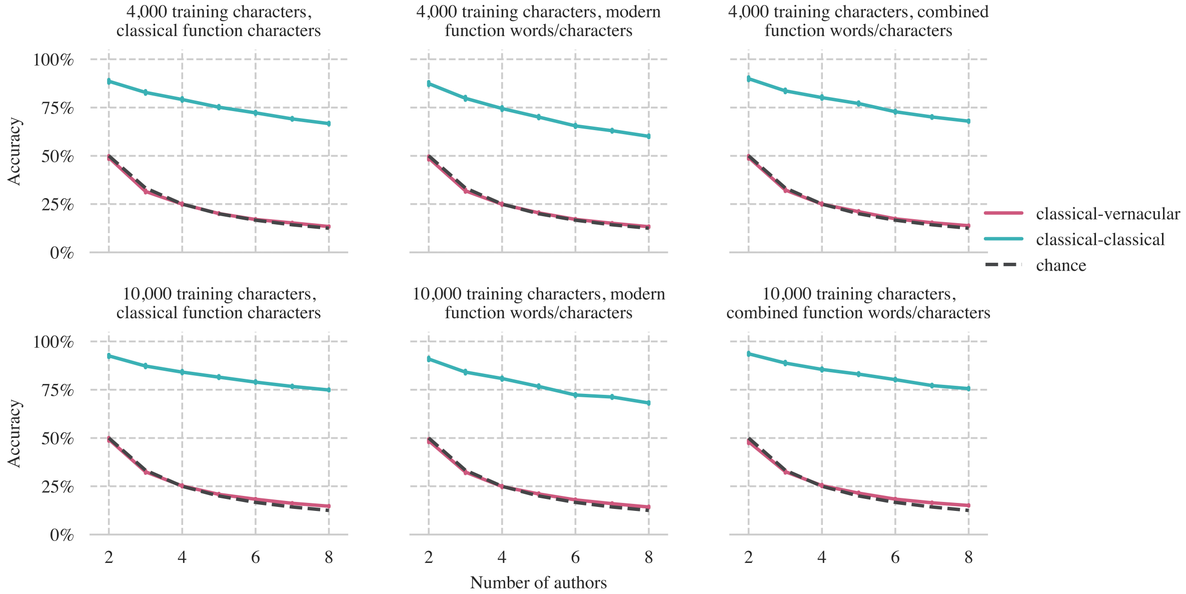


Figure 2: Assigning classical and vernacular Chinese texts using a standard linear SVM trained with classical Chinese. With different feature sets, we trained the classifier with 4,000 and 10,000 classical characters for the top row and the bottom row respectively.

the model on its ability to predict the authorship of ca. 1,000-character vernacular texts. Predictive success is measured using accuracy.

The experimental procedure is described using the limited data scenario. For a given candidate size from two to eight, we randomly choose four classical chunks from each author to fit the classifier. With the same trained classifier, we make predictions on test samples from different registers. The first prediction is made on ca. 1,000 vernacular characters, where the samples are randomly chosen from each author's vernacular writing. We care about the classical-vernacular task the most because it shows how well a classifier trained with classical texts can successfully infer vernacular texts' authorship.

The other prediction, the classical-classical task, uses 1,000 classical characters from each author. The classical-classical task's testing samples are chosen from documents or parts of documents that are not in the training set to avoid inflating the accuracy (see the Appendix for a discussion of this concern). This task works as a baseline by indicating the level of accuracy the same classifier can perform with "vanilla" authorship attribution with classical Chinese. We also use random chance—$\frac{1}{candidate\ size} \times 100\%$—as another baseline. The accuracy of the cross-register task should be bounded from above by the classical-classical accuracy and bounded from below by chance. The experiment is performed 2,000 times for every candidate size.

4. Results

We predict the authorship of 1,000-character vernacular texts and 1,000-character classical Chinese texts with three feature sets under the limited data and abundant data scenarios. The classical-vernacular and classical-classical experiments use a standard linear SVM trained on

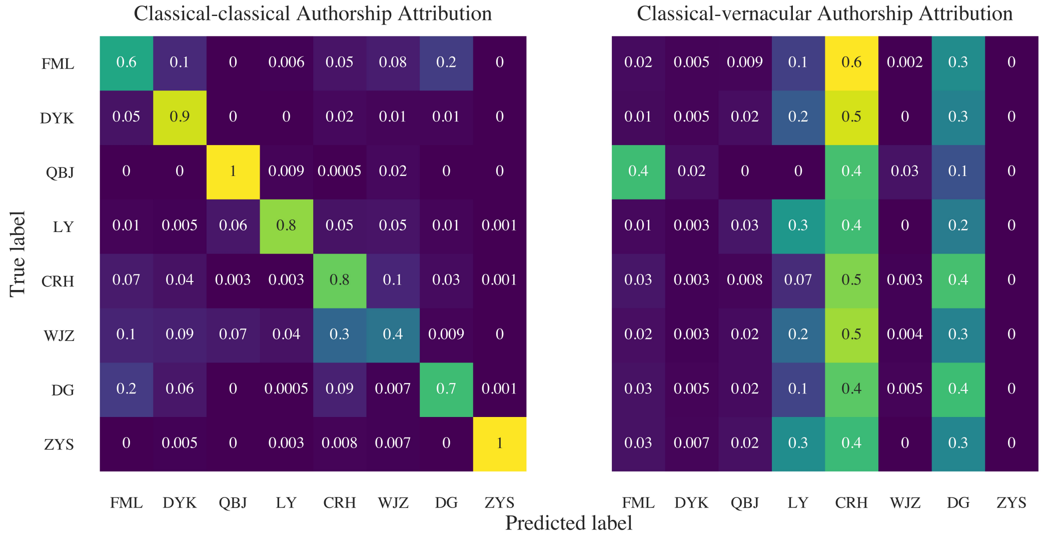


Figure 3: The confusion matrices of the classical-classical and the classical-vernacular task, normalized by rows. Rows indicate the labels of true authors' initials. Columns are for the initials of predicted authors. The result is calculated by training a standard linear SVM on ca. 10,000 classical Chinese with the combined feature set. Values are rounded to one significant digit.

the same training texts but evaluated on texts written in a different register (see Figure 2).

The classical-vernacular accuracy barely deviates from chance in the limited data scenario. Classical-vernacular accuracy is very slightly better than chance when trained on 10,000 character texts. In contrast, the classical-classical task, trained on the same data, performs far better than chance. With the optimal setting (combined feature set and 10,000 characters training data), the mean accuracy for classical-vernacular, classical-classical, and chance are 25.3%, 83.4%, and 24.5%, in turn.

The function characters/words feature sets play a trivial role in determining the cross-register accuracy but affect the classical-classical accuracy.

Confusion Matrix Confusion matrices show a similar pattern. For brevity, we only draw the confusion matrices under the abundant data scenario with the combined feature set assigning authorship given eight candidates (see Figure 3). The confusion matrices are normalized by rows (true labels).

In the confusion matrices, each row represents the true author, and each column represents the predicted author. Taking Feng Menglong (FML) as an example, the first row (labeled by "FML") indicates the empirical probability of the linear SVM predicting the author indicated by the bottom labels when the true author is FML. In the classical-classical task (the left panel), FML has a probability of 0.6 to be predicted correctly; Du Gang (DG), Ding Yaokang (DYK), Wu Jingzi (WJZ), Chu Renhuo (CRH), and Li Yu (LY) have probabilities of 0.2, 0.1, 0.08, 0.05, and 0.006 to be misclassified as FML, respectively. However, CRH, DG, and LY are more likely to be predicted to be FML than the true author FML when it comes to a cross-register situation (the right panel).

In the classical-classical task, the values on diagonal are higher than random guess ($\frac{1}{8} = 0.125$). Rarely is the classifier confused by authors with a similar writing style in the classical register. The most difficult case is predicting WJZ's classical texts. Though CRH is a competitive candidate with a probability of 0.3 being assigned, the probability of correct prediction (0.4) is highest.

In the right panel, the probability of correct classification (the diagonal values) are much lower for classical-vernacular authorship attribution. Instead, CRH tends to be predicted as the author, regardless of the true author. DG, LY, and FML also attract incorrect attributions. Other authors receive little attention from the classifier.

A follow-up experiment Since CRH and FML are among favorite candidates in the cross-register task and both are known for their vernacular works, we investigate whether vernacularization plays a part in the main experiment. We made a follow-up experiment by removing the most favorite author from the candidate pool one by one. The queue follows the popularity (CRH, DG, LY, and FML, in turn).

The result shows that the next favored author keeps taking the lead by removing the most preferred. For instance, after removing the most popular author (CRH), the second popular author (DG) draws almost all the attention of the classifier when seven candidates are present.

5. Discussion

Can we successfully infer the authorship of an unsigned vernacular text from classical texts with a standard authorship attribution technique? Our finding shows that inferring the author of a vernacular text using classical text as training data is challenging based on function characters/words frequency. Increasing training sample size is of limited help.

We speculate that the difficulty of the cross-register task lies in the elusive degree of "vernacularization" in classical texts. It is entirely possible that one's classical style is "more vernacular" than others because the transition of Chinese from classical to vernacular unfolded gradually over time. For example, Chu Renhuo (CRH), the author most likely to be predicted by the model in the cross-register settings, is well-known as a vernacular novelist. In his classical writing, the *Hard Gourd Collection*, CRH documents many anecdotes about the composition of doggerel verse and rhymed verse (*Qu*) with extensive use of the function character "了" (i.e., "今宵过了" and "见了微微笑").[12] And "了" as a modal particle rarely appears in written classical Chinese, if at all.

Also, the amount of dialog—another likely source of confusion—also varies across works. Dialog parts make a classical text resemble a vernacular text. For instance, Du Gang's the *Romance of the Northern Dynasties* contains a large portion of dialog. This factor also relates to genre. Fiction and diaries tend to have more speech-like prose relative to history and poetry. The "signal" ordinarily picked up on by function words/characters is no longer detectable given this variability.

Future work might experiment with new, purpose-built feature sets which can mitigate the problem created by dialog and vernacularization. A larger corpus, especially one with limited dialog in classical texts, would also be valuable.

Acknowledgments

We thank the anonymous reviewers for their insightful comments on the manuscript of our

[12] Qu is one of the most colloquial form among rhymed verse forms in Chinese classical literature.

paper.

References

[1] H. Baayen, H. van Halteren, A. Neijt, and F. Tweedie. “An experiment in authorship attribution”. In: *6th JADT*. Vol. 1. 2002, pp. 69–75.

[2] C.-C. Chang and C.-J. Lin. “LIBSVM: A library for support vector machines”. In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011), 27:1–27:27.

[3] P. Chen. “Coexistence of Classical Chinese and Vernacular Chinese in Fiction Writing”. In: *A Historical Study of Early Modern Chinese Fictions (1890–1920)*. Springer, 2021, pp. 123–145.

[4] L. Ge. *Out of the margins: the rise of Chinese vernacular fiction*. University of Hawaii Press, 2001.

[5] A. Glover and G. Hirst. “Detecting stylistic inconsistencies in collaborative writing”. In: *The new writing environment*. Springer, 1996, pp. 147–168.

[6] S. C. Herring and J. C. Paolillo. “Gender and genre variation in weblogs”. In: *Journal of Sociolinguistics* 10.4 (2006), pp. 439–459.

[7] W. Hu. *Bibliography on Women in Antiquity*. Ed. by H. Zhang. 3rd. Shanghai Lexicographical Publishing House, 2008.

[8] M. Koppel, J. Schler, and E. Bonchek-Dokow. “Measuring Differentiability: Unmasking Pseudonymous Authors.” In: *Journal of Machine Learning Research* 8.6 (2007), pp. 1261–1276.

[9] I. Markov, E. Stamatatos, and G. Sidorov. “Improving cross-topic authorship attribution: The role of pre-processing”. In: *International Conference on Computational Linguistics and Intelligent Text Processing*. Springer. 2017, pp. 289–302.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. “Scikit-learn: Machine Learning in Python (Version 0.24.1)”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[11] S. Pu. *Strange Stories from a Chinese Studio (Volumes 1 and 2)*. Kelly & Walsh, 1880. URL: https://www.gutenberg.org/files/43629/43629-0.txt.

[12] J. R. Rao, P. Rohatgi, et al. “Can pseudonymity really guarantee privacy?” In: *USENIX Security Symposium*. 2000, pp. 85–96.

[13] D. L. Rubin and K. Greene. “Gender-typical style in written language”. In: *Research in the Teaching of English* (1992), pp. 7–40.

[14] U. Sapkota, T. Solorio, M. Montes, and S. Bethard. “Domain adaptation for authorship attribution: Improved structural correspondence learning”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 2226–2235.

[15] U. Sapkota, T. Solorio, M. Montes, S. Bethard, and P. Rosso. “Cross-topic authorship attribution: Will out-of-topic data help?” In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 1228–1237.

[16] E. Stamatatos. “Authorship attribution using text distortion”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 2017, pp. 1138–1149.

[17] E. Stamatatos. “Masking topic-related information to enhance authorship attribution”. In: *Journal of the Association for Information Science and technology* 69.3 (2018), pp. 461–473.

[18] H. Wang, C. Zhao, S. Huang, and K. Wu. *Classical Chinese Dictionary of Function Characters*. Peking University Press, 1996.

[19] H. Wang, X. Xie, and A. Riddell. “Cross-Register Authorship Attribution using Vernacular and Classical Chinese Texts”. In: *DH Benelux 2021*. Zenodo, 2021. DOI: 10.5281/zenodo.4886596.

[20] L. Wang. *History of Chinese Linguistics*. Zhonghua Book Company, 2013.

[21] Z. Wang. *Modern Chinese Dictionary of Function Words*. Shanghai Lexicographical Publishing House, 1998.

[22] B. Yu. “Function words for Chinese authorship attribution”. In: *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. 2012, pp. 45–53.

[23] B. Zhang. *Five Hundred Kinds of Ming and Qing Novels*. Shanghai Lexicographical Publishing House, 2005.

[24] R. Zheng, J. Li, H. Chen, and Z. Huang. “A framework for authorship identification of online messages: Writing-style features and classification techniques”. In: *Journal of the American society for information science and technology* 57.3 (2006), pp. 378–393.

A. Measuring same-register accuracy inflation

When processing the corpus, we performed a “create-two-from-one” strategy on the classical works for authors who only have one work. We did that because training and testing with texts from the same work can inflate the accuracy by overfitting a model to pick up topical information. Although function words are nominally content-free, genre and content can influence function word rates. Indeed, this problem is perhaps clearer in Chinese than it is in authorship attribution work using other languages (where the problem also exists). For example, 殊 (“very”) is a common classical function character, and it is part of the name of a Buddha, Mañjuśrī (文殊). This may increase the likelihood of assigning a novel in which Mañjuśrī appears to an author who frequently uses 殊 as a function character.

To probe if our strategy prevents accuracy inflation, we ran another experiment only applying candidates who have at least two works in the classical register and compare it with the main experiment’s same-register accuracy (already with the “create-two-from-one” strategy used). Only three authors have at least two classical works. We calculated the same-register accuracy with different feature sets under the same setting corresponding to the limited data scenario of the main experiment (See Figure 4). By comparing the three-author settings, in one case strictly using different works and in the other case applying “create-two-from-one” strategy, we estimate that the accuracy inflation is low, less than 3%. Same-register accuracy is far better than chance.

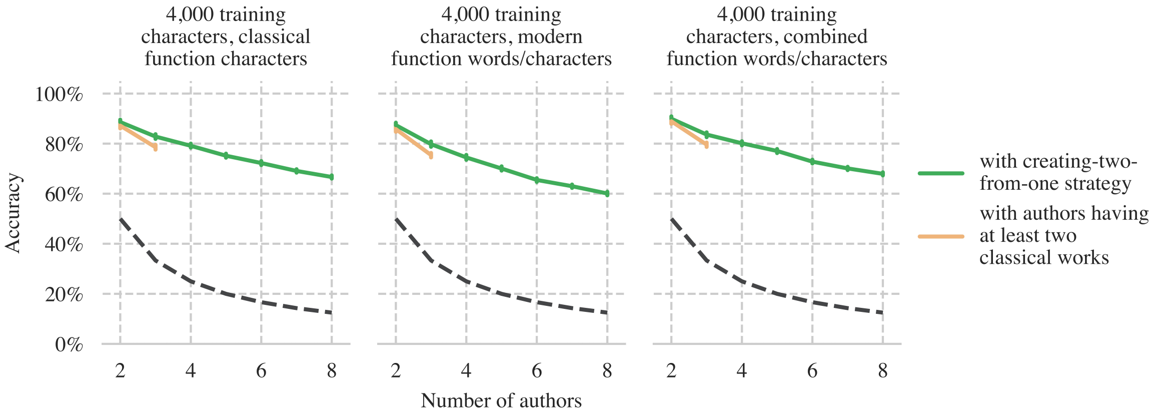


Figure 4: Comparison of the same-register accuracy computed based on authors of at least two classical works and the main experiment. Only FML, QBJ, and DG have at least two classical works and participate in this experiment. The setting mirrors that of the main experiment under a limited data scenario.