Cross-Register Authorship Attribution Using Vernacular and Classical Chinese Texts

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

[1]Indiana University Bloomington
[2]Shanghai Normal University
[1]{hw56|riddella}@indiana.edu rwxiexin@shnu.edu.cn

Today vernacular Chinese fiction from the Ming and Qing dynasties (1368 to 1912) is widely regarded as the pinnacle of Chinese literature. At the time, however, composing in vernacular Chinese was regarded as unorthodox. Classical Chinese was the privileged register. For example, official documents were all composed using classical Chinese. 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 BC). Written vernacular Chinese evolved from this version of Chinese. The differences between the two versions of Chinese are considerable. Relative to vernacular Chinese, Classical Chinese has a "denser" lexicon (words tend to consist of a single character), more frequent part-of-speech ambiguity, more variation in part-of-speech order. Frequently, especially during the Ming and Qing periods, the boundary between vernacular and classical Chinese is not clear. Many texts mix the two registers together in various ways.[1] 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 blurs further when classical grammar was mixed with the vernacular lexicon at the end of the Qing dynasty.

In the Ming and Qing dynasties, most vernacular fiction, including some masterpieces, were published anonymously or under a pseudonym. The authorship of these novels has puzzled scholars for more than a century.

By characterizing writing styles of specific authors, authorship attribution enables inferences about the likely author of texts of unknown authorship. Typically, authorship attribution begins with an assumption that, for a given text of unknown authorship, there is a set of candidate writers, one of whom wrote the text. 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 influence writing style, including genre (Koppel et al., 2007, Sapkota et al., 2016, Stamatatos, 2018), topic (Markov et al., 2017, Sapkota et al., 2014, Stamatatos, 2017),

[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) and Modern Latin (ca. 1500-1900). The analogy is not exact, of course.

gender (Herring and Paolillo, 2006, Rubin and Greene, 1992), and politics (Chen, 2021). Typically researchers begin by finding, for each candidate author, writing samples which resemble—in terms of the previously mentioned factors—the disputed text. Ming and 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 from candidates were written in classical Chinese.

Cross-register (vernacular/classical) authorship attribution has received limited attention. Yet developing reliable cross-register authorship attribution techniques will be required to resolve the long-standing debates about disputed authorship of many vernacular Chinese novels, such as the *Golden Plum Vase* (金瓶梅) and the *Marriage Destinies to Awaken the World* (醒世姻缘传). Our paper will explore the possibility of using classical Chinese texts to pin down pseudonymously and anonymously composed vernacular works. Specifically, we will evaluate simple authorship attribution techniques in a cross-register setting. Because all of the texts in our corpus have known authors, we will be able to characterize the difficulty of the task.

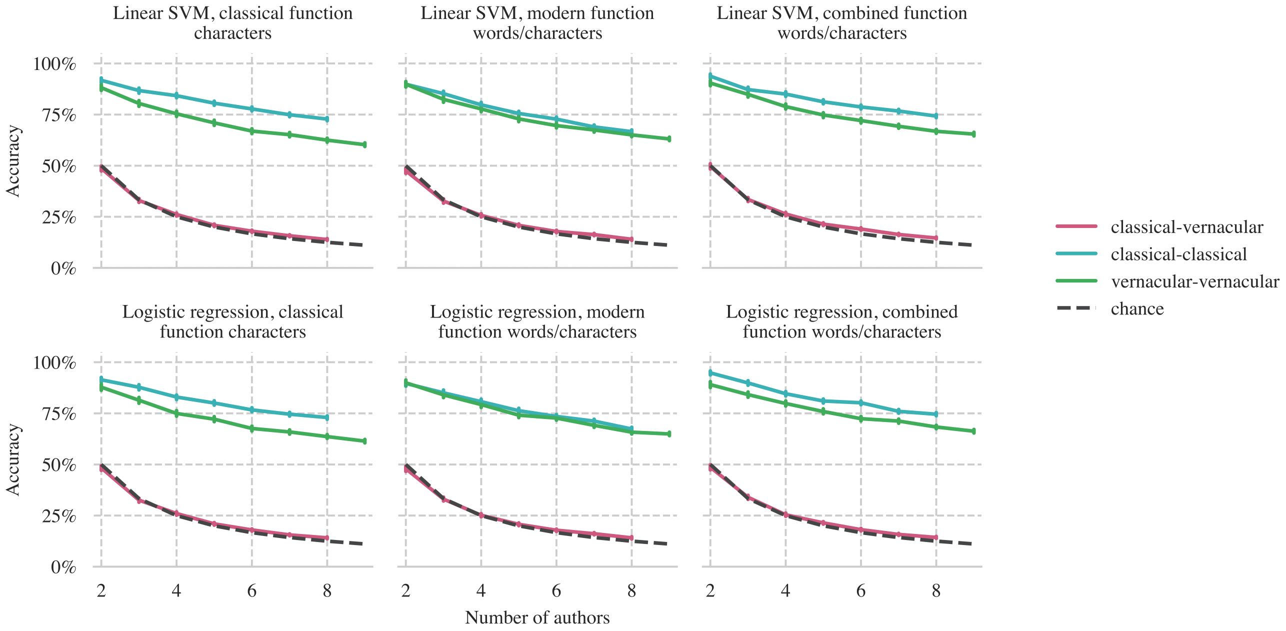

Figure 1: Assigning ca. 1,000-character vernacular Chinese texts using linear SVM and logistic regression trained with ca. 8,000 classical Chinese. The classical-vernacular and classical-classical experiments use models trained on the same training texts but evaluated using texts written in different registers. Error bars indicate a standard deviation calculated using 1,000 rounds.

To address the research problem, we organize a corpus of nine authors known to have written in both registers. All authors lived during the late-Ming to late-Qing period (between 1570 and 1870). All but one are from southern China, and all authors are men. We spent 20 hours searching for woman authors to include, but we were unable to find an author with available texts. In our search, we used Hu (2008) and Zhang (2005). We welcome suggestions for candidates to include in a future, expanded version of the corpus. The imbalance of region and gender reflects relevant social and economic circumstances of the period.

Our corpus contains 4.9 million Chinese characters (for more details, see this Google spreadsheet). All the texts are double-checked, converted into simplified Chinese, and segmented into roughly 1000-character chunks without breaking phrase-level structures.

As a preliminary measure of the difficult of cross-register attribution, we use a standard authorship attribution setup. Function words/characters are chosen for features because they have been shown to be useful stylistic markers (Yu, 2012, Zheng et al., 2006). We identify function words/characters using two published function words/characters dictionaries. The classical function character set contains 479 function unigrams (Ziqiang, 1998); the modern feature set has 819 function words (262 character unigrams, 545 bigrams, ten trigrams, two 4-grams) (Hai et al., 1996). There are 165 unigrams in both sets.[2] We also use a feature set which is the combination of the two feature sets. For the classifier, we opt for a linear support vector machine (SVM) and a logistic regression model for their efficiency and simplicity.[3]

Cross-register accuracy is calculated by predicting the authorship of a text consisting of ca. 1,000 vernacular characters using a model trained on ca. 8,000 classical characters from each candidate author. Because available texts from different writers vary considerably, a framework that can automatically decide how many candidates can participate in a specific experiment was developed. The candidates set in each experiment ranges from two to eight.[4]

Two same-register (classical-classical and vernacular-vernacular) tasks' accuracy are computed for comparison. We make sure testing and training texts are from different works, or at least from a different part, to prevent inflation in calculating same-register accuracy. (For authors who have only one document in a register, we segment the work into two parts.) The experiment is repeated 1,000 times for every possible candidate size.

Figure 1 shows that function words/characters perform similarly across the various tasks. This result indicates that inferring vernacular writing style from classical pieces is hard, if not impossible, using standard authorship attribution models based on function word frequency. We will examine the confusion matrices and attempt to account for the reason why the task is so difficult in a future work.

References

Chang, C.-C. and C.-J. Lin
2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen, P.
2021. Coexistence of classical chinese and vernacular chinese in fiction writing. In *A Historical Study of Early Modern Chinese Fictions (1890—1920)*, Pp. 123–145. Springer.

Hai, W., Z. Changcai, h. Shan, and W. Keying
1996. *Classical Chinese Dictionary of Function Characters*. Peking University Press.

[2] We released a Python package ("functionwords") on the PyPI to automate the process.

[3] We use scikit-learn (v.0.24.1)'s API for SVM (Chang and Lin, 2011) and logistic regression (with L2 regularization). We use the default cost parameter (1.0) for SVM and the default regularization parameter (1.0) for logistic regression. Features are normalized by dividing by the training samples' standard deviation after deducting the mean from them.

[4] The available classical texts for Ling Mengchu are limited. His texts feature in the vernacular-vernacular task only.

Herring, S. C. and J. C. Paolillo
2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Hu, W.
2008. *Bibliography on Women in Antiquity*, 3rd. edition. Shanghai Lexicographical Publishing House.

Koppel, M., J. Schler, and E. Bonchek-Dokow
2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8(Jun):1261–1276.

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

Rubin, D. L. and K. Greene
1992. Gender-typical style in written language. *Research in the Teaching of English*, Pp. 7–40.

Sapkota, U., T. Solorio, M. Montes, and S. Bethard
2016. 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)*, Pp. 2226–2235.

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

Stamatatos, E.
2017. 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*, Pp. 1138–1149.

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

Yu, B.
2012. Function words for chinese authorship attribution. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, Pp. 45–53.

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

Zheng, R., J. Li, H. Chen, and Z. Huang
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.

Ziqiang, W.
1998. *Modern Chinese Dictionary of Function Words*. Shanghai Lexicographical Publishing House.