Genre Identification and Network Analysis on Modern Chinese Prose Poetry

Ning, Cheng
chengn20@mails.tsinghua.edu.cn
Tsinghua University, China, People's Republic of

Wei, Zhao
xiye1002@163.com
Chinese Academy of Social Sciences, China, People's Republic of

In the nearly century-long history of modern Chinese literary research, the definition of "modern Chinese prose poetry" has given rise to a considerable amount of discussion, but remains an open question. Since the first “prose poem” was translated by Liu Bannong and published in the*New Youth* (1918), new literary writers began writing "prose poems" in newspapers and periodicals. The prose poem became a popular genre label used to this day. Howe-ver, unlike the early establishment of *vernaculars new poetry*(白话新诗), at what level it can be regarded as an independent literary genre? Or is it simply a by-product of the *new poetry movement*(新诗运动)?

During the last decade, stylistic pattern recognition and network analysis have been used to solve the problem of genre identification (Long and So, 2016), as well as to model discursive communities (Richard, 2019) or translation communities (Long, 2015). In this study, we focus on thousands of translated and composed texts of prose poems published during the China Republican period (1911-1949), and use multi-feature modelling, text clustering, machine learning, and network analysis to make a stylometric and literary sociological examination of the formation of modern prose poems, while trying to enable new theoretical assumptions. Our concerns include, but are not limited to: What stylistic or content features distinguish it from new poetry? Which texts, translators, and authors are more crucial to the shaping of prose poetic practice in the “new literary field”? And so on.

Corpora and Annotation

We collected, OCRed, proofread 1,492 works (including "733 prose poems in translation," "189 classic prose poems," and "570 unknown prose poems") with an original label of *SanwenShi*散文诗 from the archives of Republican periodicals.At the same time, we have digitized the entire*New Chinese Literature Series*, and initially built a cross-referenced corpus of 406 new poems and 216 prose(essays). All tokenized corpora are marked with the symbol “/” for the rhythmic “pause”.

Textual Representation and Genre Classification

Firstly we selected 421 features to represent the text. We chose top 200 high-frequency words calculated by tf-idf as semantic features. For the rhythmic level, we used n-gram to calculate the combination pattern of pause for each text. In addition, metric like "clause repetition rate" was counted to characterize the occurrence of "repetition" in the verse text. The remaining 20 are generic stylistic features, including TTR, sentence length mean, function word ratio, mean dependency distance (MDD) etc.

In the hierarchical clustering experiment with 150 samples, three major categories of prose, prose poetry, and new poetry have been almost clustered out.

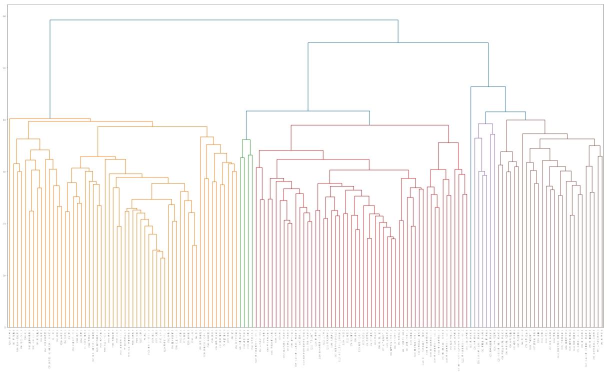

Figure 1. Hierarchical clustering based on 150 randomly sampled texts

Next, a logistic regression 3-class classifier was constructed, and it was able to separate the three with 75% accuracy (tenfold cross-validation). And the performance of 2-class task is even more accurate at 80% or higher, as followed:

	New Poetry	Prose poem
Prose	0.92	0.81
New Poetry	-	0.86

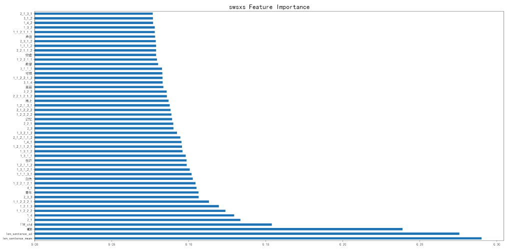

Figure 2. The ranking of the logistic regression coefficient magnitude (prose poem vs. poetry)散文诗和新诗

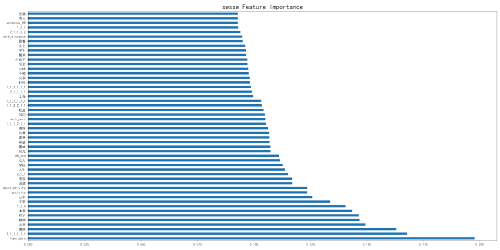

Figure 3. The ranking of the logistic regression coefficient magnitude (prose poem vs. prose)散文诗和散文

It is not difficult to find that, in addition to some explicit features, such as average sentence length, sentence length (SD), some pause combination patterns possess high contribution degrees. This confirms our previous understanding that rhyme forms based on certain patterns of pause combinations are important for establishing the stylistic independence of modern Chinese prose poetry. In other words, it is probably that the prose poem can serve as an independent genre because it possesses a unique rhythmic morphology that distinguished from the “new poetry”.

To explore this issue at a finer granularity, we transform all n-gram pause combinatorial patterns of new poetry and prose poems into feature vector model (1339 and 1547 dimensions, respectively), and adopt cosine distances to calculate the similarity of each text.As a result, if new poems are represented in a way that computes the "between-sentences n-gram combinations", new poems with similar "prosody" can be discovered. In contrast, prose poems need to be calculated as "n-gram pause combinations within pauses" in order to discover prose poems with similar rhythmic patterns. This result demonstrates that the key to the formation of the new poetry rhythmic style is the way the pauses are combined, which may not related to the number of syllables between each pause. In the case of prose (poem), the pattern of pause (including the number of syllables) within a clause needs to be taken into account.

Stylistic network analysis

In view of above differences in feature extraction algorithms, we further construct stylistic similarity networks for poems and prose poems on the basis of multiple but different features sets. In the network of prose poetry, the texts with the highest weighted degree can be considered as the model of “typical modern Chinese prose poem”. In fact, this group of unnoticeable "small poems" (小诗) overturns our impression of prose poetry in the Republican pe-riod.Finally, we also added the authorial attributes as dimensions for similarity calculation. Examining the modularity of network, it reveals some potential connections between authorial styles, and in particular the influence of translation on major writers, and of classic writings on a large number of unknown authors. In this regard, a comparison of poetry networks reveals differences that are likely due to the large number of "multiple translations of one poem" in prose poetry. In the future, we will try more network analysis methods to explore the reasons.

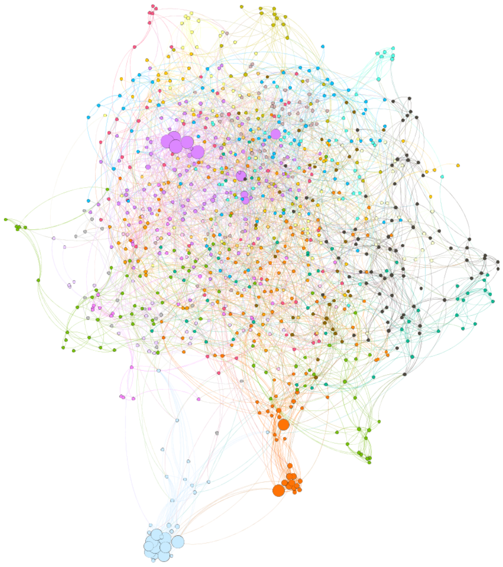

Figure 4. Prose poems stylistic similarity network (1917-1949)

Bibliography

Allison, S., Heuser, R., Jockers, M., Moretti, F. and Witmore, M. (2011). *Quantitative Formalism: An Experiment* , Stanford Lab Pamphlet Series. http://litlab.stanford.edu/?page id=255.

Long, H., & So, R. (2016). Literary Pattern Recognition: Modernism between Close Reading and Machine Learning., *Critical Inquiry* 42 (2), 235-267.

Long, H. (2016). Fog and steel: M apping communities of literary translation in an information age, *The Journal of Japanese Studies* , 41(2), 281–316.

Long, H., & So, R. (2013). Network Science and Literary History. *Leonardo,* 46(3), 274–274.

Richards, E. (2019). The Concept of Culture in the Humanities after the Digital Turn: The Stylometric Paradigm., Vielfalt der Disziplinen - Einheit des Kulturbegriffs?, Symbole, Annäherungen, Perspektiven, Diskurse,ed. CarolineKolisang, Bielefeld: Transkript Verlag, 2019, pp. 147-162.