XiaoshuoNLP: An NLP Pipeline for Processing Chinese Literary Texts

Kiara Meng Hui Liu* — Cornell University
Xiaomeng Zhu* — Yale University
Carolyn Jane Anderson — Wellesley College

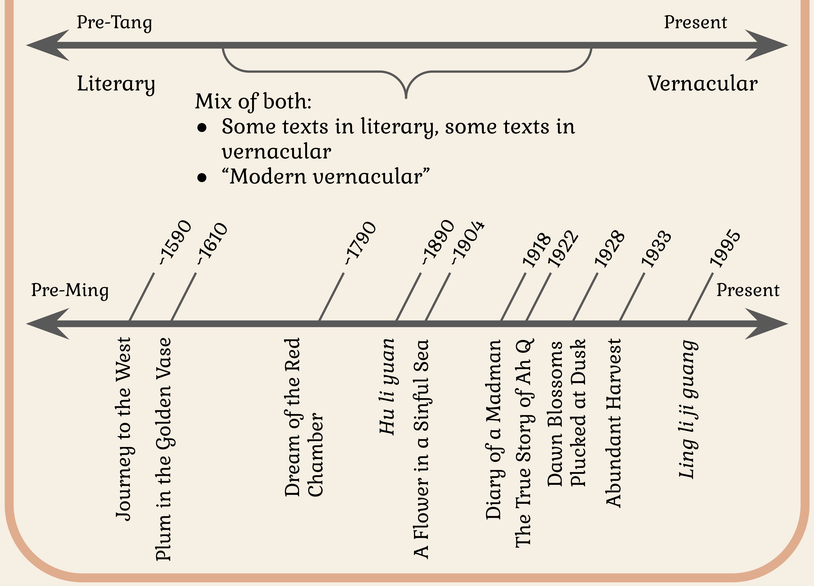
Motivation

There currently exist multiple NLP tools and pipelines for processing English literary texts. They are able to perform tasks such as tokenization, POS-tagging, NER, and coreference resolution (Bamman et al., 2014; Yoder et al., 2021). Specifically designed for processing fiction, these NLP systems allow researchers and scholars to efficiently extract information from literary narratives. However, even though Chinese is one of the most high-resource languages in the world, there is no similar pipeline in Chinese yet, potentially due to its relative lack of annotated data in specifically the fiction genre.

Corpus

From a linguistic perspective, Chinese literature can be separated into two categories: the vernacular style and the literary (classical) style. Literary Chinese was historically used for written documents and literary text. However, it did not reflect spoken language and was inaccessible to most people. The use of Vernacular Chinese in written documents came into popularity during the New Culture Movement (1910-1920). Vernacular Chinese has been widely used ever since.

Many important Chinese literary texts, such as novels from the Ming and Qing dynasties, were written in "Modern Vernacular" Chinese. This version of Vernacular Chinese is generally comprehensible to present-day Chinese speakers, but still contains elements of Literary Chinese, such as shorter token lengths and use of verse.



*Equal contributions

Pipeline

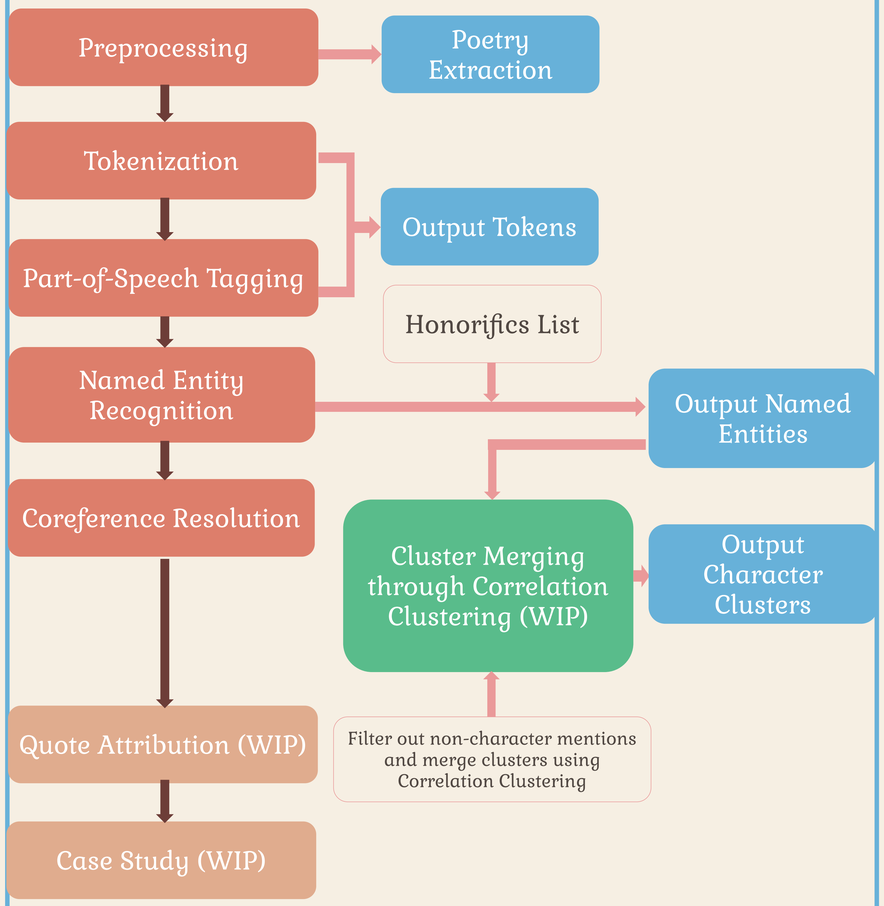


Table 1: Evaluation of 6 NLP toolkits on 4 tasks

	jieba	LAC	**HanLP** 👑	jiagu	THULAC	pkuseg
TOK	824	675	563	628	641	626
POS	0.59	0.35	0.24	0.31	0.30	0.31
	/	/	0.13	0.19	/	/
NER	/	0.71	0.79	0.16	0.48	/
COR	/	/	See Table 2 & 3	/	/	/

Major Tasks

Poetry extraction: Many Ming-Qing novels included lines of verse mixed within the prose, which could negatively affect model performance. Chinese poems often involve matching character counts, so we developed a rule-based pre-processor that extracts all poems.

Tokenization: Tokenizing Chinese words can be challenging because word boundaries are not indicated by spaces as in English. We evaluated 6 models and found that HanLP resulted in the highest performance when compared to our annotations using the minimum edit distance algorithm.

Named entity recognition: Named Entity Recognition is also more difficult in Chinese than in English given that Chinese does not capitalize proper names. To compare across different NER conventions, we compiled a list of 200+ common Chinese honorifics.

Coreference resolution: This can be difficult in some Chinese texts, especially older ones, where characters often have several names and can be referred to in multiple ways.

Table 2: Coreference recall scores with 5 metrics on 4 context lengths for *Abundant Harvest*

Metric ———— Section	MUC	B^3	BLANC	CEAF_m	CEAF_e
4	0.857	0.797	0.688	**0.537**	**0.276**
34	**0.929**	**0.893**	**0.837**	0.500	0.172
234	0.736	0.820	0.802	0.467	0.185
1234	0.739	0.833	0.825	0.444	0.171

Table 3: Coreference recall scores with 5 metrics on 4 context lengths for *Huliyuan*

Metric ———— Section	MUC	B^3	BLANC	CEAF_m	CEAF_e
4	**0.727**	**0.862**	**0.781**	**0.586**	**0.331**
34	0.641	0.610	0.489	0.393	0.231
234	0.615	0.660	0.514	0.354	0.196
1234	0.548	0.600	0.452	0.351	0.214

Future Work

Quote Attribution: This will be a component that identifies the speaker of every quote. There are two steps to this task: the first is quotation identification, and the second is speaker attribution. The former can be achieved through searching for quotation marks via simple regular expressions, while the latter would likely require us to train our own model.

Case Study: We plan to evaluate the performance of our entire pipeline by conducting a case study on some previously unseen text. We also hope to conduct a case study on a translated work, comparing the performance of our pipeline on the Chinese version to the performance of a similar English pipeline on the English version.