Automatically Dating Classical Chinese Texts: Preliminary Study on Biji and Buddhist Texts

Zuoyu Tian
Department of Linguistics, Indiana University Bloomington

Introduction

One task that often precedes an investigation of language change is automatic text dating, also called period classification, which evaluates language model's performance in capturing temporal information. However, most of such studies are limited to a few Indo-European languages, to our best knowledge, no one has introduced it into the Chinese NLP community. Furthermore, we are also aware of the lack of Chinese diachronic resources and methods for assessing Classical Chinese language models.

Therefore, we investigate two major questions in this paper:

1) Can a classifiers successfully distinguish paragraphs in Chinese historical texts from different time periods?

2) Do the features used for classification correspond to what we know from a linguistic perspective, and are the features informative for linguists?

Corpus Creation

In this project, we target two grenres: Biji and Buddhist texts.

Biji, literally "brush notes", spans hundreds of years across many dynasties and conserve informal language in written form [1]. We successfully collected 108 Biji across four dynasties. Each paragraph in each Biji is treated as a single instance.

Regarding Buddhist texts, we used texts from Taisho Tripitaka. Different from Biji, here we only focused on conversation appearing in Buddhist texts. For now, we already created a diachronic corpus of conversations within Buddhist texts. But we are still working on experiments. In the following sections, we reported our results on Biji.

Language Period	Dynasty	Instances	Biji
Middle Chinese	Tang (618-907)	4541	33
Middle Chinese	Song (960-1279)	10094	46
Early Modern Chinese	Ming (1368-1644)	9624	16
Early Modern Chinese	Qing (1644-1912)	8898	13

Statistics in Biji Data

Dynasty	Instances	Texts
Eastern Han to Western Jin	9616	220
Eastern Jin & Northern and Southern Dynasties	42697	286
Sui & Tang	22554	231
Song & Yuan	4664	114

Statistics in Buddhist Texts

Feature Representation Methods

we investigate the performance of different feature representation methods (char/word n-gram, word2vec, contextualized embedding models) in classifying the time periods of a text.

Character n-grams: character 1-5 grams with minimal frequency 10.

word n-grams: 1-5 word grams segmented by Jiayan toolkit, with minimal frequency 10.

Word2vec Features: We use a 300-dimension skip-grams with negative sampling (SGNS) character embeddings trained on *Siku Quanshu*. The final feature representation is the average of the individual character vectors in a single instance.

Contextualized Embeddings Features: We extract the representation of the final layer of a BERT-like model: Guwen-RoBertA using sentence transformers[3].

We perform five-fold cross-validation so that we use different books for training and testing. All models are evaluated using scikit-learn with the same algorithm setting each time.

Dynasty Classification

	Acc.	Prec.	Recall	F
SVMs	65.47	65.21	66.21	65.01
Log. Regression	64.63	64.81	65.24	64.47
Naive Bayes	61.38	61.20	63.52	61.02

Result for different algorithms (using SBERT).

Representation	Acc.	Prec.	Recall	F
Baseline	30.44	25.00	7.61	11.67
n-gram$_{\text{character}}$	56.37	57.03	55.52	55.48
n-gram$_{\text{word}}$	58.51	59.48	57.30	57.61
Word2vec	46.14	46.31	44.26	44.60
SBERT	65.47	65.21	66.21	65.01

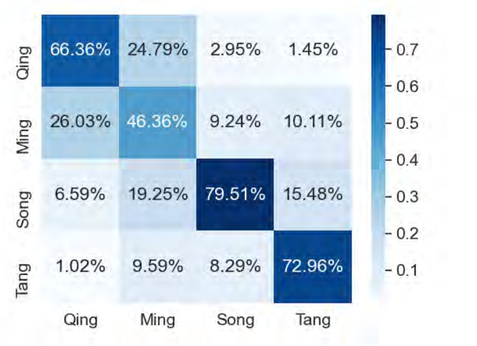
Results of different feature representations.

- Among the three algorithms, the SVM performs best.
- SBERT outperforms all the other representations, which means that contextualized features are more informative in the task of dating historical texts.
- The confusion matrix shows that Qing and Ming dynasties are similar to each other, but Song and Tang are each very distinct from any other dynasty
- Most of books with lowest accuracy can be grouped into two types: 1) the book is written at the late period of one dynasty or the early period of one dynasty; 2) the book is a collection of short stories from different dynasties

Error Analysis

Book	Chinese N.	Acc.	Dynasty
Jianhuji	坚瓠集	0.70	Qing
Nanbu Xinshu	南部新书	19.25	Song
Xihu Mengxun	西湖梦寻	21.09	Ming
Yongtong Xiaopin	涌幢小品	27.20	Ming
Jianghuai Yiren Lu	江淮异人录	28.00	Song
Wuzazu	五杂组	28.40	Ming
Duyizhi	独异志	32.00	Tang
Zhinang	智囊	32.10	Ming
Chibeioutan	池北偶谈	36.20	Qing
Qingshi	情史	36.40	Ming

10 Biji having the lowest prediction accuracy.



Confusion matrix of system using contextualized embeddings. Row: true label; column: predicted.

Feature Selection

Tang vs Song	Song vs Ming	Ming vs Qing
曰	生	女
云	而	鬼
则天	矣	某
盖	俱	狐
相国	余	我
焉	也	Proper Noun
作	女	一
馀	不	妇
于	云	曰
予	上	君
用	夫人	作
正	宋	上
知location name	宰相	来
亦	知location name	声
令	馀	汝
以	十余	忽
则	情	笑
刺史	御史	笑曰
员外	天顺	主人
皆	死	家

Top 20 features from the bleached n-gram model when classifying neighboring dynasties.

Since contextualized embeddings features are difficult to interpret, we use n-gram features but with proper nouns being bleached. Additionally, we also implement chi-square feature selection on the bleached n-gram model.

Discussions

Our results provide new evidence in a long standing debate: For the cutting point of Middle Chinese and Early Chinese, there are several assumptions, two consider the late Tang Dynasty[2] and Song Dynasty[4]. Our results support the late Tang Dynasty assumption proposed by Pan (1989), since instances are more distinguishable between Tang Dynasty and Song Dynasty.

Most of these features are very informative and fit historical linguists' observations well: for example, the lexical change of the first-person pronoun 我, 余 and 予. We also see the lexical change reflecting cultural factor like *prime minister* 相国 and 宰相.

Neighboring Dynasties Classification

Representation	Tang vs Song			Song vs Ming			Ming vs Qing		
	Prec.	Recall	F	Prec.	Recall	F	Prec.	Recall	F
SBERT	86.19	85.26	85.42	80.81	80.11	80.02	69.49	69.13	68.76
N-gram$_{\text{word}}$	81.44	77.92	79.09	77.66	76.72	76.52	59.56	59.13	58.59
N-gram$_{\text{bleach}}$	78.98	74.44	75.83	72.81	72.40	72.37	65.08	64.11	63.69
N-gram$_{\text{bleach}_{1000}}$	77.42	73.24	74.60	70.77	70.15	70.02	62.55	61.55	60.92

Results of comparing neighboring dynasties.

References

1 Yixin Fang. 2010. *Zhonggu Jindai Hanyu CIHUIXUE (Vocabulary Studies of Middle and Early Modern Chinese)*. The Commercial Press, Beijing

2 Li Wang. 1958. *Hanyu Shigao (Manuscript of the History of Chinese Language)*. Science Press, Beijing.

3 Nils Reimers and Iryna Gurevych. 2019. Sentence- BERT: Sentence embeddings using Siamese BERTnetworks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

4 Shanghai Classics Publishing House. 2007. *Qingdai Biji Xiaoshuo Daguan (Brushnotes in Qing Dynasty)*. Shanghai Classics Publishing House, Shanghai.

5 Yunzhong Pan. 1989. *Hanyu Cihui Shi Gaiyao (Summary of Chinese Vocabulary History)*. Shanghai Classics Publishing House, Shanghai.