Period Classification in Chinese Historical Texts

Zuoyu Tian and **Sandra Kübler**
Indiana University
{zuoytian,skuebler}@indiana.edu

Abstract

In this study, we study language change in Chinese Biji by using a classification task: classifying Ancient Chinese texts by time periods. Specifically, we focus on a unique genre in classical Chinese literature: Biji (literally “notebook" or “brush notes"), i.e., collections of anecdotes, quotations, etc., anything authors consider noteworthy. Biji span hundreds of years across many dynasties and conserve informal language in written form. For these reasons, they are regarded a good resource for investigating language change in Chinese (Fang, 2010). In this paper, we create a new dataset of 108 Biji across four dynasties. Based on the dataset, we first introduce a time period classification task for Chinese. Then we investigate different feature representation methods for classification. The results show that models using contextualized embeddings perform best. An analysis of the top features chosen by the word n-gram model (after bleaching proper nouns) confirms that these features are informative and correspond to observations and assumptions made by historical linguists.

1 Introduction

Traditional studies of language change largely rely on close reading and require researchers to have solid training in understanding historical texts. As a wealth of historical data has been digitized, there has been a surge in recent years in using computational methods to investigate language change (e.g. Popescu and Strapparava, 2015; Hamilton et al., 2016; Rodda et al., 2017). However, most of these studies are limited to a small number of Indo-European languages. Many languages, especially typologically very different ones, are understudied.

One task that often precedes an investigation of language change is automatic text dating, also called period classification or diachronic text evaluation. In this task, a machine learning model's performance in capturing temporal information is evaluated. Such tasks, along with datasets, exist for languages such as Irish (Han and Toner, 2017), Polish (Graliński et al., 2017), Hebrew (Liebeskind and Liebeskind, 2020), and Italian (Menini et al., 2020). To our best knowledge, there is no study focusing on dating historical texts in Chinese.

Therefore, we first built a diachronic Chinese Biji (literally “brush notes") corpus with Biji from four dynasties, spanning about 1300 years. This corpus contains around 33,000 paragraphs from 108 different Biji labeled with dynasty information. Then, we investigate two major questions: 1) Can a classifiers successfully distinguish paragraphs in 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?

For the first question, we are interested in the classification systems' performance on Chinese, since this is the first approach for this language, and written Chinese is renowned for its lack of change from century to century (Norman, 1988). More specifically, we investigate the performance of different feature representation methods (char/word n-gram, word2vec, contextualized embeddings models) in classifying the time periods of a text. These feature representation methods have been shown effective in previous studies of dating historical texts or lexical semantic change (Zampieri et al., 2016; Schlechtweg et al., 2020; Giulianelli et al., 2020).

Our next question looks closer at the interrelation between text dating and language change. The classification results provide useful information, but they cannot directly tell us which features drive language change. Since language change is an accumulating process and usually happens in neighboring time periods (dynasties), we use a task of distinguishing paragraphs from neighboring dynasties to target lexical change. We can then look at the features that ranked highest in this task and

Proceedings of LaTeCH-CLfL 2021, pages 168–177
Punta Cana, Dominican Republic (Online), November 11, 2021.

investigate whether a high relevance for this task corresponds to linguistic insights. Results show that most of the top features are informative and fit into the observation made by historical linguists.

The paper is structured as follows: We discuss related work in section 2. In section 3, we discuss the process of building a Classical Chinese Biji dataset. We introduce the feature representation methods for classifying historical texts in section 4 and the experimental setup in section 5. We report results of the different feature representation methods and present an error analysis in section 6. In section 7, we analyze the classification for neighboring dynasties, focusing on important features drawn from the training data. We conclude in section 8.

2 Related Work

2.1 Studies in Time Period Classification

Studies using language models to capture temporal information in contemporary diachronic texts date back to the 2000s (de Jong et al., 2005; Dalli and Wilks, 2006), they demonstrate that machine learning systems with textual features can successfully predict the publication time span of documents. The shared-task on Diachronic Text Evaluation in SemEval 2015 (Popescu and Strapparava, 2015) provided resources and publicity for the task of automatic historical text dating. The task was to build automatic systems to identify the time period of English news snippets from 1700 to 2010 in three subtasks. Following Popescu and Strapparava (2015), Menini et al. (2020) hosted a shared-task focusing on the same genre and cross-genre issue in dating Italian historical documents. Like SemEval 2015, they also adopted three similar classification standards: year-based, fine-grained, and coarse. There are also studies simplifying the period classification task with larger time intervals. For example, Liebeskind and Liebeskind (2020) distinguish Hebrew texts spanning around 1000 years into four time periods. We can see that tasks in different languages tend to have different classification standards, and this is closely related to the language in question and the goal of the task.

Regarding feature engineering in dating texts, lexical features including character and word n-grams are most widely used in classifying time periods. For example, Szymanski and Lynch (2015) found that character n-grams are more effective compared to other features such as POS n-grams, word n-grams, and syntactic phrase-structure rules. Zampieri et al. (2016) built an SVM classifier with word and POS n-grams to classify Portuguese texts from 1600 to 1900. Meanwhile, in the related area of lexical semantic change, many studies use word embeddings trained from deep learning models to model the language change and successfully track the semantic shift (e.g. Hamilton et al., 2016; Rodda et al., 2017; Rodman, 2020). Liebeskind and Liebeskind (2020) first implemented deep learning methods in time period classification tasks. They show that CNN and RNN outperform paragraph vectors and conventional machine learning methods. Recently, contextualized embedding models have been introduced to lexical semantic change studies (Hu et al., 2019; Giulianelli et al., 2020; Tseng et al., 2020), with promising results.

2.2 Digital Resources of Historical Chinese

Even though we see an increase of work on Classical Chinese processing, researchers also are aware of the lack of Chinese diachronic resources, which hinders the research process to some extent (Hamilton et al., 2016; Zinin and Xu, 2020). Unlike other understudied languages, many Chinese texts are already digitized, but only a few digitized texts are free to access and process. Most of the datasets are designed for close reading but not for an NLP purposes. There are well-designed POS tagged diachronic corpora and high-quality digitized resources of historical Chinese such as Zhonghua Database of Chinese Classic Books[1] and Scripta Sinica database[2]. However, they are not available for use in NLP due to copyright restrictions.

The Academia Sinica Classical Chinese Corpus[3] is the most representative diachronic Chinese corpus with sub-corpora for Old Chinese, Middle Chinese, and Early Modern Chinese. The corpus has been tagged with POS annotation. But it can only be accessed via an online search function. Recently, Zinin and Xu (2020) created a corpus of Chinese Dynasty Histories for diachronic research. However, the dynasty histories are mainly literary Chinese and normally written by a small number of authors per dynasty within a short period of time. Thus, they cannot represent the language properties through the whole dynasty.

[1] http://www.ancientbooks.cn/

[2] http://hanchi.ihp.sinica.edu.tw/

[3] http://lingcorpus.iis.sinica.edu.tw/ancient/

2.3 Linguistic Periodization of Chinese

For Chinese, different opinions exist wrt. periodization. The most widely accepted framework, proposed by Wang (1958) and Xiang (1993) and accepted by Dong (2019), splits Chinese into four major time periods: Old Chinese (pre-Qin Dynasty), Middle Chinese (Three Kingdoms and Jin Dynasty to Song Dynasty), Early Modern Chinese (Yuan Dynasty to Qing Dynasty), and Modern Chinese (after 1919 May Fourth Movement). Pan (1989) accepts the four major periods but argues that Early Modern Chinese started in Late Tang Dynasty. In contrast, Ōta (1988) proposes a new language period: Early Chinese, which divides Chinese into five periods. In this periodization system, Early Modern Chinese starts in Qing Dynasty. Thus, the major differences concern the question regarding the starting time of Early Modern Chinese. The answers range from Late Tang Dynasty to Qing Dynasty. Unlike other languages, literary (written) Chinese has undergone little change from dynasty to dynasty (Norman, 1988), hence, historical linguists tend to use materials including vernacular language to discover language change and periodize Chinese. One of the major resources for historical linguists to periodize Middle Chinese and Early Modern Chinese are Biji since they thrived during these time periods, include materials from different aspects of everyday life, and consist of vernacular language.

3 Corpus of Classical Chinese Biji

A reliable dataset is the key to dating historical texts. Taking into account the goal of our task, our dataset needs to satisfy the following requirements: 1) It need to be publicly available, so that it can be used for future studies. 3) The time period information needs to be relatively accurate. 3) As only written form is available for historical texts, we need to focus on a genre that has preserved colloquial language as much as possible. Instead of collecting texts from different genres, we concentrate on a single genre: Biji, to minimize the effect of genre differences.

Biji are “fundamentally a set of discontinuous notes of description and reflection written by the Biji author" (Fu, 2007). As suggested by Fu (2007), Biji can be roughly divided into three types based on structural characteristics: 1) Short stories, anecdotes, or biographies; 2) descriptions and reflections on a wide range of affairs; 3) recordings of a single subject such as history or poetry. This abundance of content that Biji cover also leaves us with the issue of definition: Biji and its boundaries with other genres are always controversial in academia. As we are more interested in those Biji containing more informal language, we focus on the first type of Biji, which contains more conversations. The first type of Biji is normally novel-like, also called Biji fiction. To have a professional data selection standard, during dataset creation, we choose Biji according to the book series *Biji Xiaoshuo Daguan* “the collection of Biji fiction" published by Shanghai Classics Publishing House (Shanghai Classics Publishing House, 2000, 2007a,b,c).

Since these Biji fiction collections are published based on dynasties, this also offers us trustworthy dynasty information about Biji with decisions made by professional editors. Given the difficulties involved in dating Biji fiction (Hanan et al., 1973), we use dynasties as our classes, for several reasons: First, since Biji are a set of discontinuous notes written by an author, the writing time could span dozens of years. It is unrealistic to label a specific publication time. Second, most Classical Chinese databases, such as Scripta Sinica and the Chinese Text Project, label texts with their dynasty instead of the exact publication date. This makes our dataset consistent with the widely used corpus format; Third, the most important reason is that labeling books with dynasties is sensible based on studies in historical linguistics (Wang, 1958; Xiang, 1993; Peyraube, 1996), since dynasties are widely used as an anchor point for periodization. Thus, dynasty labels can easily fit into historical linguists' periodization framework. We collect Biji titles from four books in this series: *Biji Fiction in Tang Dynasty and Five Dynasties, Biji Fiction in Song and Yuan Dynasty, Biji Fiction in Ming Dynasty, and Biji Fiction in Qing Dynasty*. All four books have a total of 140 Biji and cover four dynasties sequentially: Tang Dynasty (618-907), Song Dynasty (960-1279), Ming Dynasty (1368-1644), and Qing Dynasty (1644-1911)[4].

We used the wiki section in the Chinese Text Project[5] as the primary source of data. In this section, digitized texts are submitted by the site users, and they are freely available without copyright restrictions. We searched all Biji titles mentioned in

[4] There are a few Biji written in Five Dynasties and Yuan Dynasties in the book series. These Biji are labeled as Tang Dynasty and Song Dynasty respectively.

[5] https://ctext.org

the books and successfully collected 127 of them. However, we deleted some Biji for the following consideration: 1) Biji without punctuation. The original Biji were written without punctuation, and punctuation was generally inserted during modern editing. Thus we take the presence of punctuation as a sign for careful editing and discard the non-edited Biji. 2) Biji with encoding and formatting issues (mainly paragraph segmentation errors) and a significant amount of annotations within the text. 3) Biji which are not collections of short stories or anecdotes. For example, some Biji mainly aim to explain concepts in previous classical work. A majority of the content is sentences citing from the classical work. Such books clearly carry linguistic features from previous dynasties and are not good representations for the dynasty in which they were published.

After applying these filtering rules, 89 Biji were left. We found two problems with this dataset: 1) Due to the issues described above, some Biji from Tang Dynasty have been deleted, which causes an imbalance of the dataset. 2) Since Biji from Ming and Qing are usually lengthy, the *Collection of Biji Fiction* covers only few Biji from these dynasties. To balance the dataset, we used two open-access data sources, *Daizhige*[6] and *Xuge Shang*[7] to collect additional data. For Tang Dynasty, *Collection of Biji Fiction* already covers most of the existing Biji in Tang Dynasty, so we looked in these new data sources for alternative editions of those Biji with format issues. For Ming and Qing Dynasty, we added books from the Biji category and made sure that all these newly added Biji satisfy our filtering rules. Our final dataset includes 108 Biji.

Since Biji are usually a collection of short stories, we essentially treat each story as one instance in the corpus. We also notice that stories can be lengthy in Biji from Ming and Qing Dynasties; for such books, we treat each paragraph as a single instance. To balance the length of each instance, we discard those too long (larger than 1000 characters) or too short instances (shorter than 50 characters). Given our selection process, the quality of our dataset may not be as high as professional Classical Chinese databases since we could not manually examine the correctness of each text.

Table 1 show the statistics of the corpus. The corpus is mainly balanced, but we can see that instances from Tang Dynasty are significantly fewer than the other three dynasties. An example of instance in the corpus is shown in Table 2.

Language Period	Dynasty	Instances	Biji
Middle Chinese	Tang	4541	33
Middle Chinese	Song	10094	46
Early Modern Chinese	Ming	9624	16
Early Modern Chinese	Qing	8898	13

Table 1: Statistics for the Biji Corpus. The periodization of Chinese is based on Wang (1958).

4 Feature Representation Methods

In this section, we present the feature representations that we use for time period classification. These feature representations can be categorized into three major types: n-grams, word embeddings, and contextualized embeddings features.

4.1 n-gram features

n-gram features are widely used in dating historical texts and achieve a high performance (Zampieri et al., 2016). We use n-grams models to capture two types of information: character- and word-level.

Character n-grams Segmentation is always a concern for preprocessing Chinese, especially Classical Chinese. Modern Chinese segmentation tools are not ideal for segmenting Classical Chinese due to major difference in vocabularies. By building character-based models, we can naturally avoid the problem of segmentation. Furthermore, in Classical Chinese, many words are monosyllabic, hence a character n-gram can be an easy yet effective way to characterize Classical Chinese. We use character 1-5 grams and set the frequency threshold to 10.

Word n-grams: For a better understanding of the performance of word and character representations in Classical Chinese, we also construct word n-gram features. We use Jiayan Toolkit[8], a publicly available Classical Chinese segmenter designed for different language periods[9]. It is an unsupervised HMM segmentation tool trained on *Siku Quanshu* 'Complete Library in Four Sections'. We use word 1-5 grams and the same frequency threshold as for

[6] http://www.daizhige.org/

[7] https://www.xuges.com/gdwx/xs/bj/index.htm

[8] https://github.com/jiaeyan/Jiayan

[9] Stanford NLP toolkit Stanza also supports the segmentation in Classical Chinese, but it is trained on an Ancient Chinese corpus. The vocabulary in Ancient Chinese is very different from that in our dataset.

Chinese	陆长源以旧德为宣武军行司马，韩愈为巡官，同在使幕，或讥年辈相悬。 周愿曰：大虫、老鼠，俱为十二相属，何怪之有?
Gloss	Lu Changyuan was promoted to the general of Xuanwu Army due to previous good deed, and Han Yu was the censor of that army. Someone gloated over the generation gap between them. Zhou Yuan said: Both tiger and mouse are zodiac animals, why do you think this is weird.
Book	Wuzazu (Ming Dynasty)

Table 2: Example of a instance in the corpus

the character-based model. To evaluate the segmentation performance, we randomly sampled 50 sentences and checked them manually, the final accuracy is 94.6%.

4.2 Word2vec Features

As discussed in section 2, word2vec models have been shown to be effective in capturing language change across many languages (Schlechtweg et al., 2020). Thus, we conjecture that word embeddings will be good sources for classifying language periods. We are interested in how informative the embeddings vectors are in comparison to other feature types in the task of dating historical texts, which is different from the problem addressed in previous studies.

Assuming that the small size of our dataset would negatively influence the quality of embedding models trained on the dataset, we use pretrained character embeddings instead. We expect that these pretrained features will help us capture subtle semantic relations and infrequent phrases. We use Gensim[10] with 300-dimension skip-grams with negative sampling word embeddings (SGNS) trained on *Siku Quanshu* 'Complete Library in Four Sections' (Li et al., 2018), *Siku Quanshu* is the largest encyclopedia in Ancient China, it contains 3503 books, around 800 million tokens (character). The final feature representation is the average of the individual character vectors in a single instance.

4.3 Contextualized Embeddings Features

Instead of using static embeddings, in contextualized language models such as BERT (Devlin et al., 2019), the same word is represented differently according to its contexts. This is promising for lexical change detection, but it has not been applied to the classification of time periods. Here we use Sentence BERT (SBERT) (Reimers and Gurevych, 2019) to embed the whole instance. Using siamese and triplet network structures, Sentence BERT can efficiently generate the sentence embeddings from pre-trained models of the BERT model family. From the different pretrained models, we chose Guwen-RoBertA[11], a RoBertA model (Liu et al., 2019) trained on a large collection of Classical Chinese, which is ideal for our tasks. The final representation of each instance is the final layer of the Guwen-RoBertA extracted by sentence-transformers[12].

5 Experimental Settings

Since Biji show considerable differences in length, we split them into single paragraphs for classification. However, we need to avoid having instances from the same author distributed over training and testing data since the system may learn to classify authorship instead of time period. Therefore, we perform five-fold cross-validation with a split based on books rather than individual paragraphs. In this way, we use different books for training and testing. We additionally balance the number of paragraphs from each dynasty in each fold as much as possible. For the baseline, we use the majority class, i.e., all instances are labeled as Song Dynasty.

Since our aim is to compare the performance of different feature representation methods, we need to investigate all models using the same machine learning algorithm. We compare three machine learning algorithms widely used in classification tasks: SVM, Logistic Regression, and Naive Bayes. All the classification experiments use the Python package scikit-learn (Pedregosa et al., 2011).

6 Dynasty Classification

In this section, we compare the machine learning algorithms and different feature representation meth-

[10] https://radimrehurek.com/gensim/index.html

[11] https://github.com/Ethan-yt/guwenbert

[12] https://www.sbert.net/

	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

Table 3: 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

Table 4: Results of different feature representations.

ods, and we conduct an error analysis. In order to determine which machine learning algorithms perform well, we first compare different machine learning algorithms using the same contextualized embeddings features[13] and report the results in Table 3. Among the three algorithms, the SVM performs best; it outperforms Logistic Regression by 0.5 percent wrt. the F score. Hence, we assume SVMs to be the best choice for dating historical texts, and all the following results are reported using SVMs.

Table 4 shows the result for different feature representations. We see that SBERT outperforms all the other representations and achieves an F score of 65.01. This shows that contextualized features are more informative in the task of dating historical texts compared to other feature representations. However, if we compare our results to the results in a similar classification task using 4 time periods in Hebrew (Liebeskind and Liebeskind, 2020), the Chinese task is somewhat more challenging for classification systems. In Hebrew, a model using Naive Bayes with 20,000 selected n-gram features achieved an F score of 74.43, i.e., almost 10 points higher than our results. Possible reasons include: 1) Despite our filtering, some instances tell stories from previous dynasties, which is difficult for the classifier to interpret correctly. 2) In some folds, the test set only contains instances from few Biji of certain dynasties, and the skewed authorship makes these dynasties difficult to classify correctly. For example, in some folds, most of the test instances of a certain dynasty are from one Biji. If the classifier performs poorly on this Biji, the final accuracy will be low. A striking case is the Qing Dynasty, for which the F scores of the SBERT model vary from 45% to 85% among the five folds.

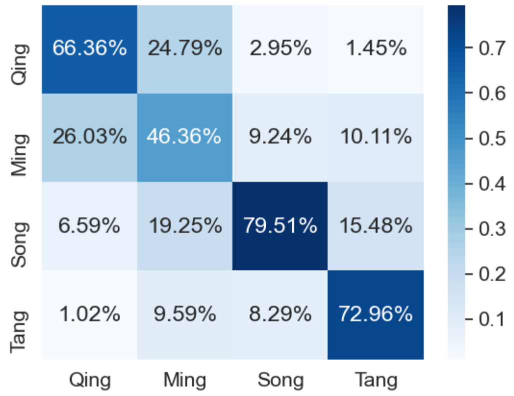

Figure 1: Confusion matrix of system using contextualized embeddings. Row: true label; column: predicted.

The next best feature type is n-gram features. We notice that word n-grams outperform character n-grams by more than 2 points in F score. A possible reason may be the increase of disyllabic words after the Tang Dynasty since disyllabic words increase in frequency from Old Chinese to Middle Chinese and from Middle Chinese to Early Modern Chinese (Norman, 1988). Therefore, word n-grams become more informative than simple character n-grams even when taking segmentation errors into consideration. Surprisingly, using word2vec embeddings as features is not very effective in this task. The F score of word2vec features is 44.6, i.e., about 10 points lower than that of character n-grams. This implies that word2vec embeddings are not very effective in capturing language change outside a neural network architecture.

Figure 1 shows the confusion matrix of the SVMs with contextualized embeddings, averaged over the five folds. It shows that the most challenging time period is Ming Dynasty. Less than 50 percent of its instances are predicted correctly. About 26% instances from Ming Dynasty are wrongly labeled as Qing Dynasty, and we see the same trend for the Qing Dynasty instances as well. Notice that this similarity may be due to the property of language usage and similar content (see section 7). Most instances from Song and Tang are labeled correctly, and we also see that they are seldom labeled as Qing Dynasty. This indicates that Qing and Ming dynasties are similar to each other, but

[13] We also examined the n-gram features and found similar trend of contextualized embeddings. Thus, we only report results from our best model.

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

Table 5: 10 Biji with the lowest prediction accuracy.

Song and Tang are each very distinct from any other dynasty.

We have a closer look at the 10 Biji with the lowest accuracy, see Table 5. We can roughly group these 10 books into two categories: 1) The book is written at the end the beginning of dynasty. Examples are *Jianhuji* (Qing), *Nanbu Xinshu* (Song), *Xihu Mengxun* (Ming), *Yongtong Xiaopin* (Ming), *Jianghuai Yiren Lu* (Song),*Wuzazu* (Ming), and *Chibeioutan* (Qing). 2) The book is a collection of short stories from different dynasties such as *Zhinang* (Ming) and *Qingshi* (Ming). Regarding the first category, language change is continuous and happens across dynasties. Thus books written either early in a new dynasty or very late in the previous dynasty will share many language characteristics and are thus difficult to classify. For example, many books from the period of Late Ming and Early Qing are wrongly predicted. According to the periodization assumption made by historical linguist Fang (2010), Ming and Early Qing should belong to the last stage of Early Modern Chinese while Mid Qing to Late Qing constitutes the transition of Early Modern Chinese to Modern Chinese. Content similarity may be another reason for wrong labeling and can explain the low accuracy of Biji in the second category. For example, *Qingshi* collected love stories from pre-Qin to Ming Dynasties, which explains why many instances are wrongly labeled as Tang and Song Dynasty. We can also see the mixture effect of language and content similarities. *Jianhuji* is written in the Early Qing Dynasty but mainly includes anecdotes from Ming Dynasty, and we see that most of the instances are predicted as Ming Dynasty. *Wuzazu* is a similar case. This book is written in the Late Ming Dynasty, and one of the major parts of the book is about the author's thoughts about Song Dynasty's strategy. Therefore, many instances are wrongly labeled as Song and Qing Dynasty. A different case is *Duyizhi*. It was written in Tang Dynasty, but most parts of the original book are missing (Liu and Wang, 2016). The existing edition was published in Ming Dynasty. However, when we look into the classifier's prediction, we find that many instances are wrongly labeled as Ming Dynasty. This implies that the texts may contain language features from the editing during Ming Dynasty, and our classifiers identify such language features.

7 Neighboring Dynasties Classification

We now turn to the second research question, investigating whether the features that our classifier uses correspond to linguistic insights. Instead of classifying the four dynasties, we now look at the classification of neighboring dynasties, which is where the effects of language change should be the most visible. Since contextualized embeddings features are difficult to interpret, we use n-gram features. However, a first trial showed that many of the high ranking features were person names, etc., which are inherently linked to a specific dynasty. Thus, we decided to bleach proper nouns, i.e., we mask words tagged as proper nouns by Jiayan Toolkit. Then, we use chi-square feature selection on the bleached n-gram model, and select the 1,000 highest ranked features for the investigation.

Table 6 shows that contextualized embeddings models perform best for this task as well. We expect bleached models to perform worse than standard word n-grams since many informative features are deleted. This is the case when distinguishing the first two neighboring dynasty pairs but the opposite is the case when distinguishing Ming Dynasty and Qing Dynasty. Across three pairs of neighboring dynasty classification, all models show the same trend that distinguishing instances between Tang and Song Dynasty is easiest for the systems, then Song and Ming Dynasty. The F score of the Ming and Qing pair is the lowest. This supports the view that Ming Dynasty and Qing Dynasty belong to the same language period. 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 (Pan, 1989) and Song Dynasty (Wang, 1958). Our results support the late Tang Dynasty assumption proposed by Pan

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

Table 6: Results of comparing neighboring dynasties.

Tang vs Song		Song vs Ming		Ming vs Qing	
n-gram	gloss	n-gram	gloss	n-gram	gloss
曰	say	生	born/man	女	woman/unmarried woman
云	say/cloud	而	and	鬼	ghost
则天	person name	矣	final particle	某	certain thing or person
盖	because	俱	all	狐	fox
相国	prime minister	余	first person pronoun	我	first person pronoun
焉	final particle	也	final particle	Proper Noun	proper noun
作	make	女	woman/unmarried woman	一	one
馀	rest	不	no	妇	married women
于	on	云	say/cloud	曰	say
予	first person pronoun	上	top/superior	君	king
用	use	夫人	madam/wife	作	make
正	proper	宋	Song Dynasty/ family name	上	top/superior
知 location	govern place	宰相	prime minister	来	come
亦	also	知 location	govern place	声	sound
令	make	馀	rest	汝	second person pronoun
以	by means of	十余	around ten	忽	suddenly
则	but	情	feeling	笑	smile
刺史	provincial governor	御史	imperial censor	笑 曰	smile and say
员外	landlord	天 顺	regnal year	主人	master
皆	all	死	die	家	home

Table 7: Top 20 features from the bleached n-gram model when classifying neighboring dynasties.

(1989), since instances are more distinguishable between Tang Dynasty and Song Dynasty.

To look into the features selected by the classifiers, we list the top 20 features extracted from the bleached n-gram models[14], shown in Table 7. These features contain two words that should have been bleached but are listed due to segmentation and tagging error[15]. However, most of these features are very informative and fit historical linguists' observations well: We see the lexical change of the first-person pronoun. Many words are used as first-person singular pronouns in literary Chinese, including 我, 余 and 予. In Modern Chinese, 我 is the only first-person singular pronoun left. From the table, we can see that 我 is listed among the top features only in the comparison between Ming and Qing Dynasty. This shows that it becomes more and more prominent as the time period approaches Modern Chinese (Wang, 2018b). Meanwhile, the results also imply that 予 can distinguish instances from Tang Dynasty and Song Dynasty, and the usage of 余 is different in Song and Ming Dynasty.

Another example concerns the features 皆 and 俱. For both words, the main sense is 'all (adverb)'. Our results show that 皆 is useful in distinguishing instances from Tang to Song Dynasty while 俱 is helpful for classifying Song and Ming Dynasty. These results can be compared to a study by Wang (2018a), who found that the usage of 皆 decreases time, and both 皆 and 俱 are substituted by 都 gradually in modern Chinese. However, the change of 皆 and 俱 is not addressed by Wang (2018a). In contrast, surprisingly, 都 does not appear in our top 20 features. The differences need to be investigated more closely in the future. We also see cultural changes reflected. For example, 相国 is a common title for Prime Minister from Qin to Tang Dynasty, but after Tang Dynasty, 宰相 is used to refer to the Prime Minister. However,

[14] The rank is calculated by averaging the scores of a feature across folds.

[15] 则天 is the given name of a famous female ruler in China, but it is tagged as adverb. 天 顺 is the regnal year in Yuan Dynasty but is wrongly segmented.

even after bleaching proper nouns, there remain features containing indirect information about the dynasty. For example, 鬼 and 女 are the names of frequently used main characters in Biji from Qing Dynasty. This makes them informative features, and they rank as the top features in the comparison of instances from Ming and Qing Dynasty.

8 Conclusion & Future Work

In this paper, we have approached the task of dating Chinese historical texts and examined the performance of different feature representations and machine learning algorithms. The first public available diachronic Biji dataset is created for tracking language change[16]. Our main findings are that even though dating historical Chinese texts is a challenging task for different types of feature representation methods, we still find informative features concerning language change, which correspond to historical linguists' observations. However, currently our methods only capture lexical change. In future work, we will investigate models for syntactic change.

References

Angelo Dalli and Yorick Wilks. 2006. Automatic dating of documents and temporal text classification. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 17–22.

Franciska de Jong, Henning Rode, and Djoerd Hiemstra. 2005. Temporal language models for the disclosure of historical text. *Humanities, Computers and Cultural Heritage*, page 161.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Hongyuan Dong. 2019. *A History of the Chinese Language*. Routledge.

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

Daiwie Fu. 2007. The flourishing of Biji or pen-notes texts and its relations to history of knowledge in Song China (960-1279). *Extrême-Orient Extrême-Occident*, pages 103–130.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973.

Filip Graliński, Rafał Jaworski, Łukasz Borchmann, and Piotr Wierzchoń. 2017. The RetroC challenge: How to guess the publication year of a text? In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 29–34.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, Berlin, Germany.

Xiwu Han and Gregory Toner. 2017. Dating texts by multi-class classification with sliding time intervals. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE.

Patrick Hanan et al. 1973. *The Chinese Short Story: Studies in Dating, Authorship, and Composition*, volume 21. Harvard University Asia Center.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143.

Shmuel Liebeskind and Chaya Liebeskind. 2020. Deep learning for period classification of historical Hebrew texts. *Journal of Data Mining & Digital Humanities*, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Zehua Liu and Junde Wang. 2016. The research of writer and edition of duyizhi. *Journal of Qilu Normal University*, 6.

Stefano Menini, Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2020. DaDoEval@EVALITA 2020: Same-genre and cross-genre dating of historical documents. In *7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. EVALITA 2020*, pages 391–397. Accademia University Press.

[16] The dataset is available at https://github.com/zytian9/Chinese-Biji-Corpus.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

Tatsuo Ōta. 1988. *Chūgokugoshi Tsūkō [Comprehensive Studies of the History of the Chinese Language]*. Tokyo: Hakuteisha.

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

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research*, 12:2825–2830.

Alain Peyraube. 1996. Recent issues in Chinese historical syntax. In *New Horizons in Chinese Linguistics*, pages 161–213. Springer.

Octavian Popescu and Carlo Strapparava. 2015. SemEval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Martina A Rodda, Marco SG Senaldi, and Alessandro Lenci. 2017. Panta rei: Tracking semantic change with distributional semantics in Ancient Greek. *IJ-CoL. Italian Journal of Computational Linguistics*, 3(3-1):11–24.

Emma Rodman. 2020. A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, 28(1):87–111.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23.

Shanghai Classics Publishing House. 2000. *Tang Wudai Biji Xiaoshuo Daguan (Brushnotes in Tang Dynasty and Five Dynasties)*. Shanghai Classics Publishing House, Shanghai.

Shanghai Classics Publishing House. 2007a. *Mingdai Biji Xiaoshuo Daguan (Brushnotes in Ming Dynasty)*. Shanghai Classics Publishing House, Shanghai.

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

Shanghai Classics Publishing House. 2007c. *Songyuan Biji Xiaoshuo Daguan (Brushnotes in Song Dynasty and Yuan Dynasties)*. Shanghai Classics Publishing House, Shanghai.

Terrence Szymanski and Gerard Lynch. 2015. UCD: Diachronic text classification with character, word, and syntactic n-grams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Yu-Hsiang Tseng, Shu-Kai Hsieh, Pei-Yi Chen, et al. 2020. Computational modeling of affixoid behavior in chinese morphology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2879–2888.

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

Min Wang. 2018a. Hanyu zongkuo fuci de laiyuan (the source of Chinese blanket range adverbs). *Xiandai Yuwen (Modern Chinese)*, 2018(6):4–11.

Weihui Wang. 2018b. *Hanyu Hexinci de Lishi yu Xianzhuang Yanjiu (The History and Current State of Core Vocabulary in Chinese)*. The Commercial Press, Beijing.

Xi Xiang. 1993. *Jianming Hanyu Shi (A Brief History of Chinese Language)*. Higher Education Press, Beijing.

Marcos Zampieri, Shervin Malmasi, and Mark Dras. 2016. Modeling language change in historical corpora: The case of Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4098–4104.

Sergey Zinin and Yang Xu. 2020. Corpus of chinese dynastic histories: Gender analysis over two millennia. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 785–793.