WYWEB: A Classical Chinese NLP Evaluation Benchmark

Anonymous ARR submission

Abstract

For natural language processing (NLP), evaluation benchmarks, such as GLUE, and SuperGLUE, allow researchers to evaluate new models on a set of tasks. For Chinese NLU, the CLUE benchmark brings together more than 10 tasks, benefiting Chinese language researchers. However, CLUE does not apply to Classical Chinese, also known as "wen yan wen"(文言文), which has thousands of years of inheritance attracting researchers from all over the world. For the prosperity of the community, in this paper, we introduce WYWEB evaluation benchmark, which contains eight tasks, implementing sentence classification, sequence labeling, reading comprehension, and machine translation. All of the tasks are designed according to actual requirements of domain researchers and students. The github repository and leaderboard of WYWEB will be released when accepted.

1 Introduction

Classical Chinese, as a written form of Chinese language, had been widely used in the Confucian cultural circle, including China, Japan, Korea, Vietnam, etc (Ye and Tian, 2013; Phong1 and Van2, 2020; Xu, 1995; Zhou, 2009; Jin, 2004). As we know, there are about 400 million words, 3 million ancient articles have been passed down, covering literature, art, history, philosophy, etc, half of which are of great value (Yin et al., 2018). However, in recent centuries, it has become increasingly difficult to understand the language as it has been gradually replaced by modern official languages everywhere. Therefore, it is necessary to introduce efficient NLP technology to process, understand, and research such literature.

In recent years, pre-trained language models such as BERT (Devlin et al., 2019) and BERT-like models (Yang et al., 2019; Dong et al., 2019; Lan et al., 2019; Liu et al., 2019; He et al., 2020; Raffel et al., 2019; Wang et al., 2019d), have shown remarkable performance on NLP benchmarks, including GLUE (Wang et al., 2019c) and SuperGLUE (Wang et al., 2019a). Meanwhile, there are also many efforts (Cui et al., 2020a; Wei et al., 2019; Cui et al., 2021a) in Chinese NLP community, achieving significant improvement on Chinese NLP benchmark, CLUE (Xu et al., 2020), and datasets (Cui et al., 2018; Duan et al., 2019; Cui et al., 2020b). However, since Classical Chinese differs from modern Chinese in writing and grammar, these benchmarks can not be applied to Classical Chinese. Tasks and datasets need to be redesigned to fit Classical Chinese language. Meanwhile, due to the performance is closely related to the corpus data (size, language, domain, etc.) (Qiu et al., 2020), pre-trained language models on modern Chinese corpus can not be perfectly applied to Classical Chinese.

Since previous works (Wang et al., 2021; Yang et al., 2021; Koichi et al., 2022) generally evaluate their models on few, different NLU tasks, the results of which could not be comparable. To facilitate such research in Classical Chinese, it's necessary to design a standard Classical Chinese NLP evaluation benchmark.

In this paper, we introduce WYWEB (Wen Yan Wen Evaluation Benchmark), which will be open, and continually developed as possible as we can. To evaluate how well learned models for Classical Chinese language representation perform, we create and refine 8 tasks for different aspects of the language understanding.

Specifically, for sequence labeling, we design punctuation *PUNC* and named entity recognition *GLNER* tasks to evaluate word separation capability of pre-trained language models. For sentence classification, we design sentence category specification *GJC*, written time specification *TLC* and emotion specification of poems task *FSPC*. Furthermore, we design a reading comprehension task, *IRC*, from exam paper and idiom dictionary. Since machine translation of classical Chinese is also a problem of great concern, we design *WYWMT* task to work on this topic. In addition, a task of token comparison, *Xuci*, is also provided. Details are shown in Section 4 and Appendix B.

In Section 3, we describe the principles we used to design tasks and collect the data.

To better understand the challenges provided by WYWEB, we build a baseline for each task and evaluate several pre-trained models released by the community. The results of experiment in Section 5.5 demonstrate that current state-of-the-art methods are struggling with these tasks. This suggests that those tasks in WYWEB can constitute a useful test-bed for developing and comparing NLP systems for Classical Chinese.

The contributions of our work are summarized as follows:

- We design, create and collect eight Classical Chinese NLP tasks.
- We build an online leaderboard and evaluation tool set for further exploration.
- We conduct a series of experiments with baselines.

2 Related Work

2.1 Benchmarks for Pre-trained Language Model

With the rise of pre-training language model, pre-training a model on large corpus and fine-tuning them on downstream tasks becomes general practice in NLP community. To evaluate the ability of pre-trained language models in NLP tasks, SentEval (Conneau and Kiela, 2018), GLUE (Wang et al., 2019c) and SuperGLUE (Wang et al., 2019b) are proposed to provide benchmarks for NLU tasks, making experiments of models comparable. For Chinese NLU, CLUE (Xu et al., 2020) benchmark is proposed with more than 10 tasks, including several sentence classification tasks and several reading comprehension tasks, as well as QA tasks. To evaluate the ability of pre-trained language models in both natural language understanding and generation, CUGE (Yao et al., 2021) is proposed. This benchmark is designed as a hierarchical framework which using multilevel scoring strategy. Meanwhile, to evaluate whether language models can learn a linguistic phenomena of Chinese, Xiang et al. (2021) develops CLiMP which covering 9 major Mandarin linguistic phenomena. QuoteR (Qi et al., 2022) is designed for evaluation of quote recommendation methods. CBLUE (Zhang et al., 2021) is a biomedical language understanding benchmark for Chinese, which mainly focuses on information extraction.

However, for Classical Chinese, there are not so many datasets and benchmarks have been proposed as modern Chinese. Although CCLUE [1] project provides some NLU tasks for classical Chinese, these tasks are not well defined and dataset quality is not as good as other benchmarks.

2.2 Corpus Datasets for Classical Chinese

The largest classical corpus dataset available is Daizhige (殆知阁)[2]. This dataset contains about 3.3 billion tokens of classical Chinese literature which makes classical Chinese corpus not low-resource. Most of pre-training related works use this dataset to train their models.

Ancient Chinese Corpus (ACC) [3] dataset dataset contains the word segmented, POS-tagged data of Zuozhuan (an ancient Chinese history classical book). This dataset is widely used in ancient Chinese studies.

Recently, Zinin and Xu (2020) introduces an open source corpus of Twenty-Four Histories and some other ancient books. Meanwhile, FSPC (Chen et al., 2019) and CCMP (Li et al., 2021) are proposed for ancient poem understanding. While CUGE (Yao et al., 2021) use CCMP as a subtask for classical poetry matching, in this work, we apply the FSPC dataset for poetry emotion recognition.

2.3 Pre-trained Models for Classical Chinese

In Classical Chinese pre-trained language models, SikuBERT and SikuRoBERTa (Wang et al., 2021) are pre-trained BERT/RoBERTa model on the Si Ku Quan Shu (Complete library in the Four Branches of Literature) corpus, and evaluated on 4 tasks, including speech tagging, tokenization, named entity recognition and punctuation which are built from ACC dataset. Meanwhile, based on RoBERTa model, GuwenBERT [4] is pre-trained on Daizhige corpus with continuous training method and is evaluated on several NLU tasks. Other works (Hu Renfen, 2021; Yu et al., 2021; Yang et al.,

[1] https://cclue.top/
[2] https://github.com/garychowcmu/daizhigev20
[3] https://catalog.ldc.upenn.edu/docs/LDC2017T14/
[4] https://github.com/ethan-yt/guwenbert

Task	Train	Dev	Test	Description	Metric	Source
PUNC	90k	20k	20k	Sequence labeling	F1	handcrafted
TLC	28k	6k	6k	Sentence classification	Accuracy	handcrafted
GJC	100k	20k	20k	Sentence classification	F1	handcrafted
XuCi	800	200	200	Token similarity	Accuracy	handcrafted
IRC	3k	1k	1k	Reading comprehension	Accuracy	handcrafted
WYWMT	20k	3k	3k	Machine Translation	BLEU	handcrafted
GLNER	80k	18k	18k	Sequence labeling	F1	GULIAN (2020)
FSPC	3000	1000	1000	Sentence classification	Accuracy	THU-FSPC

Table 1: The statistics of tasks in WYWEB, including the number of dataset, task description, evaluation metric and source. The datasets, except GLNER and FSPC, are handcrafted by us.

2021) also evaluate their models on different few NLP tasks.

3 WYWEB Overview

In this section, principles and methods we applied during construction process of WYWEB are introduced, and a brief overview of tasks is provided in Table 1. First, we describe the process of task design and the principles we follow. Then, we introduce the data selection principles in Section 3.2. After that, we discuss the different character styles in Chinese. Finally, we provide the description of leaderboard and toolkit.

3.1 Task Design Principles

In this work, to assure that the benchmark could evaluate most aspects of pre-trained models and language phenomenons, we design evaluation tasks following best practices of other NLP benchmarks (Xu et al., 2020; Yao et al., 2021; Wang et al., 2019c,b) and suggestions from experts in Classical Chinese. Following the principles of Xu et al. (2020), firstly, these tasks should vary in most aspects of NLP, including text classification, reading comprehension and machine translation etc. Secondly, these tasks should be well defined in the academic community and easily processed for corpus collection. Thirdly, they should be challenging but solvable. Finally, these tasks should be useful for follow-up studies and representative of Classical Chinese natural language understanding tasks.

With the study of thousands of Chinese exam papers and requirements from academia and applications, we construct the tasks, covering most of the regular NLP tasks. In addition to the regular tasks, we designed several tasks specifically for classical Chinese, i.e., punctuation of sentences without punctuation marks, comparison of confusing words and written period classification. These tasks will be introduced in Section 4 and Appendix B.

3.2 Corpora Selection

Since Classical Chinese has a very long history and evolves over time, when designing tasks, we should choose texts that cover as many periods as possible. It is supposed that it is not reasonable enough to treat isolated article as an independent task. For instance, as mentioned in Section 2.3, Wang et al. (2021) evaluate their model on ACC corpus which is built on Zuo Zhuan. However, Zuo Zhuan was written by Zuo Qiuming in East Zhou Dynasty, so that the text features of Zuo Zhuan are relatively simple and are difficult to reflect the multi-faceted characteristics of ancient Chinese. Therefore, in this work we refine datasets like this and combine into well-defined datasets to build uniform sample sets.

Classical Chinese was officially and commonly used as a written language before recent times in East Asia, but now, modern Chinese is the dominant language in China. People learn classical Chinese at school but rarely use it in everyday life except some poems and idioms. We could hardly collect any publicly available NLP datasets compared to modern Chinese. As a result, we design and create most of WYWEB datasets by ourselves. Data collection process for these new datasets is described in Appendix A.

3.3 Character Style Selection

The debate on the character issue in the cultural field is very intense. In this work, we do not want to discuss any tendency, but only take the choice that can reduce the complexity of research.

There are mainly two character styles in the Chinese written language, including Simplified Chinese and Traditional Chinese. Simplified characters

Traditional	Simplified	English
難寫	难写	hard to write
憂鬱	忧郁	melancholy
寧波	宁波	Ningbo

Table 2: Comparison between simplified Chinese with traditional Chinese.

Simplified	Traditional	Pinyin
杯	杯、盃	bei
升	升、昇、陞	sheng
台	臺、台、檯、颱	tai

Table 3: One-to-many from simplified characters to traditional characters (about 20 instances).

Traditional	Simplified	Pinyin
著	著、着	zhe
藉	藉、借	jie
畫	画、划	hua

Table 4: One-to-many from traditional characters to simplified characters (about 400 instances).

are developed from traditional characters to help people out of illiteracy and improve writing efficiency (Wang, 1991b; Su, 2003; Yuang, 1991). A comparison of the two styles is shown in Table 2.

Traditional characters have an indelible position in Chinese culture (Wang, 1991a). All of the classical Chinese documents were written in traditional characters. However, simplified characters have become mainstream among Chinese speakers. The very most of the text data we collect for building our tasks is in simplified style.

Furthermore, there are several styles of traditional characters including Hongkong style, Taiwan style, Japan style, and so on. For example, "里" is a simplified character, but in Hongkong and Taiwan style, they are "裏" and "裡" respectively. Many other examples could be found in these styles. The characters issue of multi-source text is very complicated. Tasks with mixed character styles are very challenging, but it is not easy for researchers to collect sufficient data to train their models for every style. So we consider that unified texts are more acceptable and have lower complexity for researchers.

According to the statistics from the official standard (《general table of simplified Chinese characters》, 《the first batch of variant characters sorting table》, etc.), there are about 400 one-to-many cases in the conversion from simplified to traditional Chinese, while there are only more than 20 cases of one-to-many conversion in the conversion from traditional to simplified Chinese. See Table 3 and Table 4 for some examples. So, the effect of converting traditional Chinese to simplified is more reliable if we wish to unify our corpora. And converting all the characters to simplified ones is not a bad choice, because the simplified characters are unique.

However, there are some issues when we use simplified characters to study classical Chinese literature. For instance, simplified characters unified some characters with different meanings, i.e, multi traditional characters are unified to one simplified one, causing issue like apple (fruit) vs. apple (company). It is considered that contextual models which are trained efficiently could learn all of the meanings. Another issue is, some one-to-many conversions may cause errors. When engineering the task, We manually correct a few words that are prone to errors, such as "乾" to "干", "徵" to "征" and so on.

Overall, we select simplified characters to build all of the tasks. But we think that it also makes sense to create tasks that use traditional characters or both styles. We consider it as an important future work.

3.4 Leader-board

We create a leaderboard on a standalone website for WYWEB. It will be accessible as soon as possible.

3.5 Toolkit

We provide scripts implemented using PyTorch (Paszke et al., 2019) and transformers (Wolf et al., 2020). The followers can evaluate their models easily using this toolkit. Otherwise, models could be uploaded to Hugging Face Model Hub [5] and then get evaluated by contacting us. This toolkit is also released on WYWEB repository.

4 Tasks

In this section, we describe tasks designed for specific aspects of classical Chinese NLP, as well as datasets, respectively. The datasets, except GLNER and FSPC, are created by us. More data examples are shown in Appendix B.

[5] https://huggingface.co/

4.1 Single Sentence Classification Tasks

GJC This task aim to work on the problem of ancient book classification which has been discussed since ancient times. The Si Ku Quan Shu had formed a classification method of four parts of Jing, Shi, Zi, Ji (Confucian classics, historical records, philosophical writings, and miscellaneous works), and 40 categories. This is the authoritative method till now. The largest classical Chinese corpus dataset Daizhige extends the method to 10 collections. Since this corpus is actually the basis of most of classical Chinese NLP research, we apply this method to design our text classification task following CCLUE. Texts from each category of the Daizhige project are selected with a specified proportion and split into the evaluation dataset. See Appendix B.4 for details.

TLC Since ancient books have been handed down over a period of more than 2,000 years, it is a very meaningful and challenging task to identify the writing time of ancient books according to the characteristics of the text. Chang et al. (2021) propose that identifying written time of literature is helpful for understanding works. Being classified according to the period, ancient Chinese is generally divided into ancient (Pre-Qin and Han Dynasty), mid-ancient (Jin Dynasty to Song Dynasty) and late-ancient (Yuan, Ming, Qing Dynasty) (Wang, 2004). Furthermore, in the process of the development of classical Chinese, each dynasty has its own unique characteristics (Li et al., 2013). In such background, we collect about 300 hundred ancient books and famous articles which have exact time of writing, and sample a reasonable number of paragraphs from the texts. Each sample of this dataset has a coarse-grained label (period) and a fine-grained label (Dynasty) forming a hierarchical structural. See Appendix B.3 for details.

FSPC FSPC (Fine-grained Sentiment Poetry Corpus) is an emotion recognition task for ancient rhythmic poetry. The dataset is created by THUAIPoet (九歌) group (Chen et al., 2019). Sentiments are annotated into 5 classes, i.e. negative, implicit negative, neutral, implicit positive, and positive. THUAIPoet designs a reasonable annotation mechanism to ensure annotations follow similar standards during the work process. See Appendix B.5 for details.

4.2 Sequence Labeling Tasks

PUNC This task is designed for text punctuation. Since there are not any punctuation marks in traditional Chinese literature, discriminatory of sentence punctuation is important for reading ancient books. Even though ancient Chinese researchers have made great efforts in the proofreading and sorting out of ancient books, there are still a large number of ancient books without punctuation waiting to be solved (Qi, 2022; Li, 2002). So that punctuation task is useful for classical Chinese researchers. Therefore, all related works evaluate their models mainly on this task.

To make sure the time distribution of the corpus as uniform as possible, we select history books as source data for this task including 二十四史(the Twenty-Four Histories), 春秋(The Spring and Autumn Annals), 战国策(Strategies of the Warring States Period) and so on. The corpus contains historical books from the Zhou Dynasty to the Republic of China, which cover nearly three thousand years (1046 BC to 1927). All of the books are concatenated and shuffled by paragraph, and then split into reasonable datasets. See Appendix B.5 for details.

GLNER This is a named entity recognition task with a dataset created by GULIAN (2020). Texts of the dataset are selected from ancient books and some other relevant literature. There are two kinds of entities in this dataset, i.e., classical book name and other which including human name, location name, etc. Since the entity category is of coarse grain size, it is expected to implement a new labeling work to refine this dataset in the future. See Appendix B.2 for details.

4.3 Sentence Pair Tasks

XuCi This task is designed to determine whether two function words in a sentence pair have the same meaning and usage. Function words (Xu ci in Chinese) have no real meaning and generally cannot be used as a single sentence element (Liu et al., 1995). They are very important in classical Chinese but easily confused. Relevant topics is part of the basic knowledge for Chinese students which appears in the college entrance exam every year. We collect sentence pairs with function words from examination papers with help of middle school teachers to construct this dataset. See Appendix B.6 for details.

4.4 Reading Comprehension Tasks

Reading comprehension tasks are usually in the form of choosing the best from multiple options. We implement the following task according to this best practice.

IRC This task is designed aiming to solve idiom comprehension which is a very important part of classical Chinese learning and tests every year in the college entrance exam.

Idiom is one of the major features of Chinese culture. Most of the idioms are long-standing fixed phrases, derived from ancient classics or writings, historical stories, and oral stories. For idiom comprehension, there are other tasks (Zheng et al., 2019) ready. However, they are mainly aiming to test modern Chinese texts with idioms. To focus on classical Chinese, we implement this task as: given an Idiom and its origin (most are in classical Chinese), select the best explanation from four options. See Appendix B.7 for details.

4.5 Sequence to Sequence Tasks

WYWMT Machine translation of Classical Chinese is a problem of great concern. Classical Chinese is a very concise written language, so it's not easy for everyone to understand. Scholars often translate classical Chinese into modern Chinese with notations to make it easier for people to read. We consider it as an in-language translation or rewriting task because the source and target could share the same vocabulary and some semantic features. Since the evaluation metric of this seq2seq task is different from that of other NLU tasks, we separate this task from others to make a stand alone leaderboard.

This dataset is filtered and calibrated from hundreds of translated classical Chinese books collected from multiple channels. Since allusions and quotations appear frequently in classical Chinese, and these references may have a time span of thousands of years, it's not easy to construct a very well-established dataset by ourselves. So this dataset is used for evaluation purposes and far from training a good machine translation system. See Appendix B.8 for details.

5 Baselines

5.1 Baseline Implementation

Substantial works have shown that pre-trained models have achieved great success (Qiu et al., 2020; Han et al., 2021) on NLP tasks. For baselines, we implement models for all tasks in WYWEB using pretrained models, i.e., adding a specified prediction head on the model output for every task respectively.

Sequence Labeling: Get hidden states of the last layer of the model encoder and pass them to a classifier to get sequence labels.

Sentence Classification: Get pooled out of model encoder, i.e., hidden state of [CLS] token, and pass that to a classifier to get sequence labels.

Reading Comprehension: Encode every option concatenated with paragraph-question, and pass the hidden states to a shared classifier to get a score. The best one will be the answer.

Token Similarity: Encode sentence pairs the same as a sentence classification task, and get the hidden state vector of the corresponding token. Marking the vectors to be compared as $\mathbf{u}$ and $\mathbf{v}$, we use $\{\mathbf{u}\ ; \mathbf{v}; |\mathbf{u} - \mathbf{v}|\}$ to represent the similarity score.

Machine Translation: We implement this task as sentence pair with a prefix attention mask to adapt BERT style models. To save inference time cost, the sequence output of the target sentence is greedy, partial auto-regressively predicted and decoded. Note that this implementation is untypical for the sequence to sequence models. Because fewer parameters and evaluation resource cost are needed, we believe that this approach is more able to reflect the capabilities of the model itself.

All the experiments are implemented using PyTorch (Paszke et al., 2019).

5.2 Pretrained Models to Be Evaluated

GuwenBERT GuwenBERT has three versions, including GuwenBERT-base, GuwenBERT-large, GuwenBERT-fs-base. While GuwenBERT-base and GuwenBERT-large are trained based on RoBERTa-wwm-ext (Cui et al., 2021b) modern Chinese pre-trained model and then continue trained on classical Chinese corpus, GuwenBERT-fs-base is trained purely on classical Chinese corpus. Note that all three models above are pre-trained in RoBERTa style actually.

RoBERTa-classical-chinese RoBERTa-classical-chinese has two versions, RoBERTa-classical-chinese-base-char (RoBERTa-CCBC),

Models	Avg.	Sequence Labeling		Sentence Classification			Token Sim.	Reading Comp.
		PUNC	GLNER	GJC	FSPC	TLC	Xuci	IRC
Human	89.2	92.4	94.3	90.5	80.2	89.0	85.5	92.5
GuwenBERT-base	79.3	82.5	82.8	84.8	61.3	85.1	71.7	86.8
GuwenBERT-large	80.1	83.1	86.1	84.9	58.5	**87.6**	**73.4**	**87.8**
GuwenBERT-base-fs	79.3	82.9	84.8	84.2	61.0	86.7	70.0	85.3
RoBERTa-CCBC	79.4	82.5	84.7	84.5	59.5	85.0	73.2	86.1
RoBERTa-CCLC	80.2	82.8	86.1	84.7	58.6	87.1	74.9	86.9
SikuBERT	77.9	80.8	82.8	82.2	60.9	82.4	70.4	85.8
SikuRoBERTa	78.1	81.4	82.8	82.5	**62.2**	83.8	68.5	85.8
DeBERTa-base	**80.3**	**83.3**	**86.7**	**85.2**	61.1	86.7	72.4	86.7
RoBERTa-wwm-ext	76.4	78.8	79.8	81.3	59.2	78.3	71.0	86.2

Table 5: Baseline results.

RoBERTa-classical-chinese-large-char (RoBERTa-CCLC). This is a RoBERTa model pre-trained on Classical Chinese texts, derived from GuwenBERT-base. Character-embeddings are enhanced into traditional/simplified characters (Koichi et al., 2022).

SikuBERT, SikuRoBERTa These models are pre-trained on the verified high-quality "Siku Quanshu" (Wang et al., 2021). Note that these two models are pre-trained on traditional Chinese. In fine-tuning phase, we convert simplified Chinese corpus into traditional.

DeBERTa-base Based on the structure of DeBERTa (He et al., 2020), we pre-trained the model on DaiZhiGe corpus from scratch.

RoBERTa-wwm-ext This model is trained with BERT (RoBERTa) structure (Cui et al., 2021b) and whole word masking.

It is noting that there are not as many pre-training models of Classical Chinese as modern Chinese. We collect all models of Classical Chinese which are accessible to evaluate and take them as baselines. More details of these models can be found in Appendix C

5.3 Experiment Setting

We fine-tune pre-trained models mentioned above by adding a classifier with the same architecture respectively. For each task, we train 3 runs, and the model with best dev score is used for the test report. When the learning rate decreases to a specified small value or the performance do not improve for 5 evaluations, the training is stopped. Details of hyper-parameters are shown in Appendix D

5.4 Human Performance

For all tasks, we evaluate human performance following the principle of SuperGLUE (Wang et al., 2019b): extract 30 samples in the training phase, and then sample 100 items from the test set in the testing phase. We collect test results from three annotators and calculate the human performance. The annotators are all college students majoring in ancient Chinese. The results are shown in Table 5 and Table 6.

5.5 Benchmark Results

The results of our baseline models are reported in Table 5. As evaluation metrics of sequence to sequence tasks are different from NLU tasks, WYWMT task for each model is evaluated independently with several seq2seq metrics and BLEU is used as the primary metric. Results of WYWMT is shown in Table 6.

5.6 Baseline Analysis

From the results, it can be seen that some regular patterns, i.e. "the bigger (model scale and batch size), the better"; "the more (data and train steps), the better" appear as described in other experiments.

DeBERTa-base (He et al., 2020) performs best on this benchmark showing that the model structure and training strategy are both effective. Note that this model is pretrained just according to default settings of DeBERTa V2 English version without convolution layer and purely on classical Chinese. Some techniques that have obvious effects in Chinese are not used, such as Whole Word Masking (Cui et al., 2021b), etc.

All models pretrained on classical Chinese get better scores than the model chinese-roberta-wwm-ext (Cui et al., 2021b) which was pretrained on

Model	BLEU	chrF2	TER*	ROUGE-1	ROUGE-2	ROUGE-L
Human	45.6	44.2	34.4	77.4	50.7	76.2
guwenbert-base	**40.1**	**38.1**	37.5	**72.5**	**46.0**	**70.3**
guwenbert-large	38.8	37.2	38.1	70.1	43.7	67.7
guwenbert-base-fs	36.3	35.2	39.2	68.3	41.2	65.7
roberta-CCBC	39.1	37.1	36.8	71.4	44.9	69.3
roberta-CCLC	39.8	38.0	36.4	71.6	45.3	69.3
SikuBERT	38.8	36.2	37.9	72.0	45.5	69.8
SikuRoBERTa	39.1	36.5	37.7	72.2	45.7	70.0
DeBERTa-base	39.5	37.8	**35.9**	71.9	44.2	68.7
chinese-roberta-wwm-ext	38.0	35.8	39.1	69.9	43.2	66.7

* Translation Edit Rate

Table 6: WYWMT results.

modern Chinese corpus. Similarly, models trained on both classical Chinese and modern Chinese perform better on tasks involving both scripts.

For FSPC task, which composed of ancient Chinese rhythmic poems, SikuRoBERTa (Wang et al., 2021) performs the best. The authors claim that they pretrained the model using a high-quality classical Chinese corpus of Si Ku Quan Shu, which has a much smaller scale but better quality than Daizhige. Because these poems are very different from general texts, we think that models could learn better ancient word representation using this type of corpus.

The two large models yield similar scores to DeBERTa-base but much better than other smaller ones. However, large version models have 3 times more parameters than DeBERTa-base.

On WYWMT task, we find that GuwenBERT-base achieves the best score. It is supposed that its pretraining strategy works well. The strategy is:

- Initialize the transformer model parameters from a pretrained model without the embedding layers;
- train the model by freezing transformer encoder layers to translate modern Chinese knowledge to classical;
- update all parameters of the model.

Applying this pretraining strategy, the model could learn a good representation of both modern and classical Chinese. So that it could get the best score on the translation task.

Comparing with human performance, all the models have a big gap with the artificial results, especially on tasks GLNER, XuCi, and IRC which require a lot of implicit knowledge.

One limitation of our evaluation is, models we collected are all BERT or RoBERTa style and are lacking some variety. Furthermore, models we evaluated maybe not achieve the best score in this baseline due to difference among them. However, they are fine-tuned with similar hyper parameters, so that the results are comparable as expected.

6 Conclusions and Future Work

In this paper, we introduce a NLP benchmark for classical Chinese, which contains 8 NLP tasks and datasets respectively to help researchers to evaluate and Analyse NLP models. Also, we created a leaderboard online for the community.

Comparing with general benchmarks, this work is restricted. Also, the study of ancient Chinese is a highly specialized subject, so the professionalism of this benchmark may need to be further improved. On the other hand, there is a big gap between the performance of the classical Chinese models on this benchmark with other leader-boards. Better models are needed to handle more linguistic features of classical Chinese.

Furthermore, to resolve traditional and simplified character issue, traditional style tasks are meaningful to researchers. We consider it as a future work of the community.

Classical Chinese is a treasure of the entire human cultural history. We contribute this work with the hope of helping the entire community to be more prosperous. This work will be an open, community-driven project which improves with the advancement of technology.

7 Limitations

In this work, we contribute an evaluation benchmark for classical Chinese NLP tasks. However, our work has several limitations due to lacking expertise knowledge and data.

When designing the tasks, we got a lot of inspiration from the middle school Chinese test paper. thousands of test papers are collected in order to extract data for NLP tasks. During the work process, we learn that it is difficult to extract a sufficient number of questions of a single type. The main difficulty is due to the variety of questions on the test papers and the mixture of the language of classical and modern Chinese. Finally, we create Xuci task and IRC task from the test papers and related literature but failed to create solvable machine reading comprehension and natural language inference tasks.

When working on some datasets which has less corpus, i.e, the Xuci task, we find it very difficult to calibrate existing samples or create new ones. This problem also exists in other tasks. For instance, the category rule we followed in the GJC task is not certified by authoritative experts, so this method is not completely reliable if viewed by experts of classical Chinese.

In this work, tasks for the more aspects of grammar phenomenon are lacking. For an evaluation benchmark, it is actually far from enough.

It's expected that more classical Chinese experts and researchers join this work in the future to solve the above problems.

On the other hand, we lack a diagnostic dataset compared to other benchmarks. This is because similar data (NLI corpus generally) are even more difficult to retrieve. However, this benchmark works for NLP researchers even though the diagnostic dataset is missing. This issue is also expected to be solved in future work.

References

Ernie Chang, Yow-Ting Shiue, Hui-Syuan Yeh, and Vera Demberg. 2021. Time-aware ancient chinese text translation and inference. *arXiv preprint arXiv:2107.03179*.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Cheng Yang, Wenhao Li, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020a. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021a. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021b. Pre-training with whole word masking for chinese bert.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*.

Yiming Cui, Ting Liu, Ziqing Yang, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. 2020b. A sentence cloze dataset for Chinese machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6717–6723, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *China National Conference on Chinese Computational Linguistics*, pages 439–451. Springer.

GULIAN. 2020. "gulian cup" ancient book document named entity recognition competition of ccl 2020.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Zhu Yuchen Hu Renfen, Li Shen. 2021. Knowledge representation and sentence segmentation of ancient chinese based on deep language models. *JOURNAL OF CHINESE INFORMATION PROCESSING*, 35(4).

Jishi Jin. 2004. A brief critical summuary of the chinese language education in rok. *DongJiang Journal*, 21(1).

Yasuoka Koichi, Wittern Christian, Morioka Tomohiko, Ikeda Takumi, Yamazaki Naoki, Nikaido Yoshihiro, Suzuki Shingo, Moro Shigeki, and Fujita Kazunori. 2022. Designing universal dependencies for classical chinese and its application. *Journal of Information Processing Society of Japan*, 63(2).

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Guangjie Li, Xiaomei Gao, and Xiulan Cui. 2013. 汉语发展史研究. 黑龙江大学出版社.

Guoxin Li. 2002. The development and task of chinese ancient book resources digitization. *Journal of Academic Libraries*, (1):21–26.

Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. Ccpm: A chinese classical poetry matching dataset. *arXiv preprint arXiv:2106.01979*.

Jian Liu, Guangshun Cao, and Fuxiang Wu. 1995. 论诱发汉语词汇语法化的若干因素. 中国语文, (3):161–169.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Nguyen Xuan Phong1 and Vu Hong Van2. 2020. Taoism in vietnam during the northern colonial period and some notes when studying taoism in vietnam. *Journal of Natural Remedies*, 21(8(1)):342–352.

Fanchao Qi, Yanhui Yang, Jing Yi, Zhili Cheng, Zhiyuan Liu, and Maosong Sun. 2022. QuoteR: A benchmark of quote recommendation for writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 336–348, Dublin, Ireland. Association for Computational Linguistics.

Jianglei Qi. 2022. 古籍知识服务平台发展策略. *Chinese Editors Journal*, (2):60–65.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Peicheng Su. 2003. A review of the simplified chinese characters. *JOURNAL OF PEKING UNIVERSITY*, (1):121–128.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019b. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019c. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Dongbo Wang, Chang Liu, Zihe Zhu, Jiangfeng Liu, Haotian Hu, Si Shen, and Bin Li. 2021. Sikubert 与sikuroberta：面向数字人文的《四库全书》预训练模型构建及应用研究. *Library Tribune*.

Li Wang. 2004. 汉语史稿. 中华书局.

Ning Wang. 1991a. 汉字的优化与简化. *Social Sciences in China*, (1):69–80.

Ning Wang. 1991b. 论汉字简化的必然趋势及其优化的原则. *Language Planning*, (2):26–31.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019d. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. In *COLING*.

Qiuhan Xu. 1995. 汉字在日本. 中国文化研究, pages 135–139+6.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zinong Yang, Ke-jia Chen, and Jingqiang Chen. 2021. Guwen-unilm: Machine translation between ancient and modern chinese based on pre-trained models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 116–128. Springer.

Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, et al. 2021. Cuge: A chinese language understanding and generation evaluation benchmark. *arXiv preprint arXiv:2112.13610*.

Shaofei Ye and Zhiyong Tian. 2013. On the origins of vietnamese ancient history. *Southeast Asian and South Asian Studies*, (2):83–89.

Xiaolin Yin, Ming Fang, and Wenfan Shen, editors. 2018. 中华传世藏书. 浙江人民出版社有限公司.

Jingsong Yu, Yi Wei, and Yongwei Zhang. 2021. Automatic ancient chinese texts segmentation based on bert. *JOURNAL OF CHINESE INFORMATION PROCESSING*, 33(11).

Xigui Yuang. 1991. 从纯文字学角度看简化字. *Language Planning*, (2):20–22.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. *arXiv preprint arXiv:2106.08087*.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. Chid: A large-scale chinese idiom dataset for cloze test. *arXiv preprint arXiv:1906.01265*.

Bin Zhou. 2009. A comprehensive study on the history presented in a series of biographies written in chinese in japan. *Journal of Historiography*, pages 98–104.

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, Marseille, France. European Language Resources Association.

A Data Collection Process

In this appendix, we describe principle and methodology applied when we create new datasets.

A.1 Data Collection

In Section 3.2, principle of data selection is discussed. We collect data from multi channels, including Dazhige, ACC, Gushiwenwang (for WYWMT task) [6], and many other web sites, dictionaries and so on. When selecting candidate sentences, we apply rules as: (1) having refined punctuation marks; (2) having more than 4 words in classification tasks; (3) being originally simplified Chinese character style preferred.

For GJC task, because some books/articles may appear in more than one categories, they are disregarded to avoid confusion.

A.2 Annotation and Quality Checks

For the annotation work, annotators are required to be familiar with ancient Chinese. Some rules annotators following are: (1) dropping out confusing sentences and sentences; (2) double-checking rarely used words and dropping out sentences with uncertain rarely used words; (3) removing unnecessary symbols except specified punctuation marks.

Quality checks for WYWMT Translating classical Chinese sentence to modern Chinese is challenging. We follow Guzmán et al. (2019) to filter texts collected from internet. In addition, because classical Chinese sentences are usually short, the limitation of sample length is set to 5 to 200 characters.

B Data Examples and Statistics

In this section, we use a "[SEP]" mark to denote separation between two parts of a sample. And we try to translate the classical sentence to English to make it easier to understand.

B.1 PUNC

This dataset is in sentence pair TSV format. Every sample is a pair of source text and label sequence as shown following. We choose eight punctuation marks as prediction target in this dataset. Statistics are show in Figure 1 Figure 2 and Table 7.

壬戌诏定科举流寓人名额蒙古色目南人各十五名汉人二十名

[6] www.gushiwen.cn

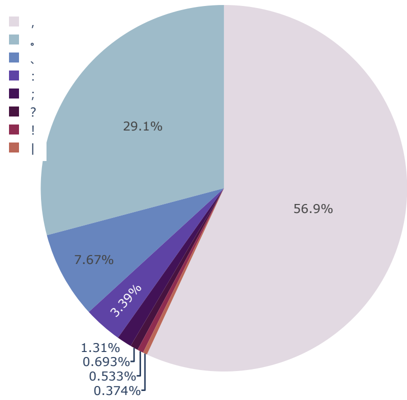

Figure 1: Percentage of punctuation marks to be predicted of PUNC dataset.

OO，OOOOOOOO，O、O、OOOOO，OOOO。
On the 24th, an imperial edict was issued to establish the quota of expatriates in the imperial examination, 15 for Mongolians, 15 for colored-eyes and 20 for Han Chinese.

谢肇《北河纪》八卷《纪余》四卷除坛西郊坎其击鼓百灵至止结作主
OOOOOOOOOO，OOOOO|OOO，OOO。OOO，OOO。
Xie Zhaozhe wrote eight volumes of Beihe Ji and four volumes of Ji Yu. At the altar in the western suburbs, playing drums, hundreds of gods stopped here and became the leader of the alliance.

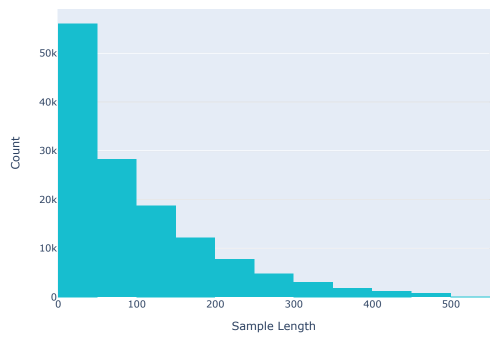

Figure 2: Statistic of Sample Length of PUNC dataset.

B.2 GLNER

This dataset is in JSON format as shown bellow. Every sample consists two keys which are "text"

Total Samples	135156
Mean Sample Length	100
Min Sample Length	5
Max Sample Length	510

Table 7: Statistic of Sample Length of PUNC dataset.

Total Samples	18762
Mean Sample Length	210
Min Sample Length	28
Max Sample Length	510

Table 8: Statistic of Sample Length of GLNER dataset.

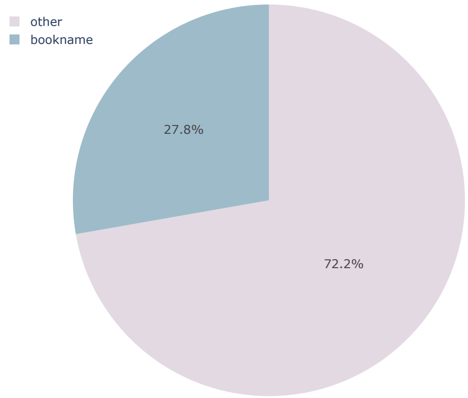

Figure 3: Percentage of labels to be predicted of GLNER dataset.

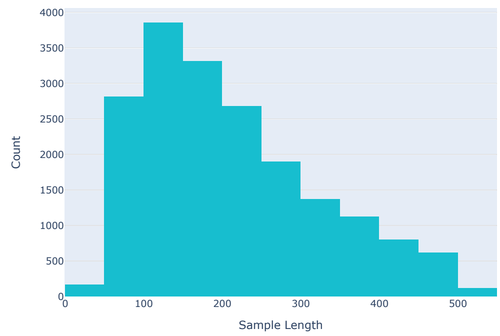

Figure 4: Statistic of Sample Length of GLNER dataset.

and "label", and every label is represented as start index, end index and category style. Statistics are show in Figure 3 Figure 4 and Table 8.

{
"text": "谢绛　三月戊戌，知礼仪院、兵部员外、知制诰谢绛知邓州。十一月己酉，卒。欧文。长编：绛按召信臣故迹，距城三里，壅湍水，注钳庐陂，溉田，请复修之。可，罢州人岁役。",
"label": [[0, 2, "other"], [21, 23, "other"], [24, 26, "other"], [35, 36, "other"], [38, 40, "bookname"], [41, 42, "other"], [43, 46, "other"], [59, 62, "other"]]
}

{
"text": "六月己未，郑居中等上哲宗御集。壬戌，景灵宫建禧祖殿室。复广、惠、康、贺州旧铸夹锡钱监。辛未，湖南路提点刑狱陈义夫奏邵阳县贼平。",
"label": [[5, 8, "other"], [10, 14, "bookname"], [18, 21, "other"], [22, 24, "other"], [28, 29, "other"], [30, 31, "other"], [32, 33, "other"], [34, 36, "other"], [46, 49, "other"], [53, 56, "other"], [57, 61, "other"]]
}

B.3 TLC

This dataset is in TSV format as shown bellow. The three segments of a sample are Period label, Dynasty label and source text respectively. Statistics are show in Figure 5, Figure 6 Figure 7 and Table 9.

近古 [SEP] 元明 [SEP] 主治风痹，筋骨不仁，功与脂同。补虚羸。
Indications of wind arthralgia, numbness of the muscles and bones, its power is the same as fat. Make up for weakness.
中古 [SEP] 魏晋南北朝 [SEP] 东观汉记曰：羌什长巩便。然更盖其种也。尚书曰：歼厥渠魁。既已袭而馆其县。左氏传曰：凡师轻曰袭。杜预曰：掩其不备。子以眇尔之身，介乎重围之里；率寡弱之众，据十雉之城。
Dongguan Han Ji said: Gong Bian, the leader of the Qiang people. But it is another cover. The book of Shang said: Destroy the head of the thief. Has attacked Fianxian and stayed in a hotel. Zuo's biography said: "Any army with light baggage is called 袭." Du Yu said: Attacking who is unprepared. With a small body, you are in the center of the encirclement, leading the weak, and defending the city of ten feet.

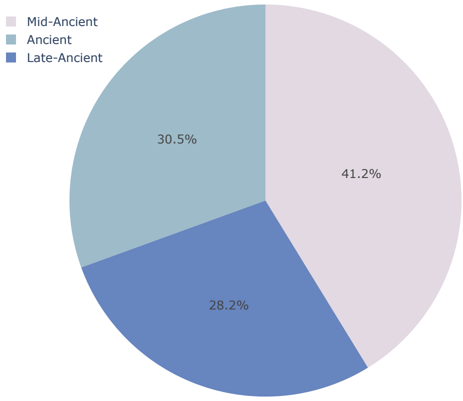

Figure 5: Percentage of Period labels.

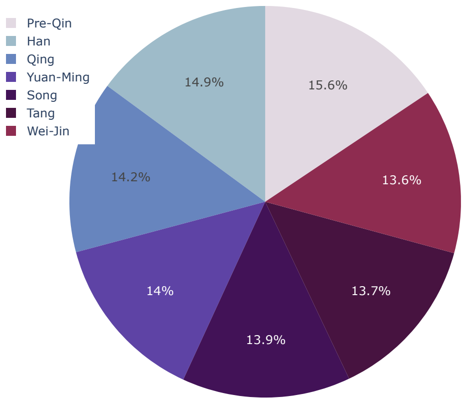

Figure 6: Percentage of Dynasty labels.

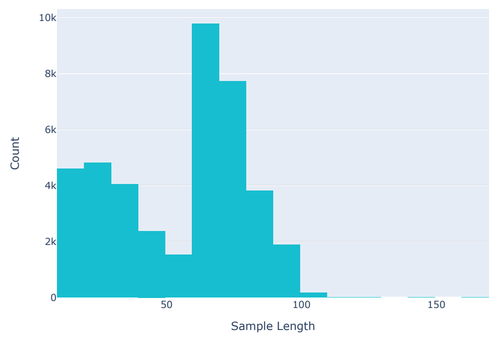

Figure 7: Statistic of Sample Length of TLC dataset.

远古 [SEP] 先秦 [SEP] 齐晏桓子卒，晏婴粗斩，苴、带、杖，菅屦，食鬻，居倚庐，寝苫、枕草。
When father died, Yan Ying wore coarse cloth mourning clothes, made filial piety clothes, belts and walking sticks of coarse linen, wore shoes made of thatch, ate thatch, ate thatch, lived in a leaning hut, and slept on a straw mattress.

Total Samples	40788
Mean Sample Length	54
Min Sample Length	11
Max Sample Length	166

Table 9: Statistic of Sample Length of TLC dataset.

Total Samples	200000
Mean Sample Length	92
Min Sample Length	5
Max Sample Length	257

Table 10: Statistic of Sample Length of GJC dataset.

B.4 GJC

This dataset is in text–category format as shown bellow. Statistics are show in Figure 8 Figure 9 and Table 10.

然则世所谓雅乐者，未必如古，而教坊所奏，岂尽为淫声哉？” [SEP] 艺藏
However, the elegant music in the society is not necessarily the same as in ancient times, but is the music played by Jiaofang all debauched music?
有丧必求牧师殓，独自入房把门掩。[SEP] 子藏
If there is a funeral, you must find the priest to be buried, entering the room alone and close the door.
“梦幻空花，何劳把捉？得失是非，一时放却。” [SEP] 佛藏
"Dreaming of empty flowers, how to take the handle? Do not care about the right and wrong, and put it back at once. ""
羲，乃天皇伏羲氏也。齐驱，即并驾。元始，万有万无之祖号。比肩，并立之义。是足上文比喻也。学者慎毋住相，是即舜何人也，予何人也云尔。[SEP] 道藏
Xi is also called the Emperor Fuxi. “齐驱” means the two marched side by side. “元始”, the ancestor of all things and nothing. "比肩" means to stand side by side. It is enough to describe the above. A scholar must not be too pretentious. This person is what kind of person Shun is, and what kind of person am I.

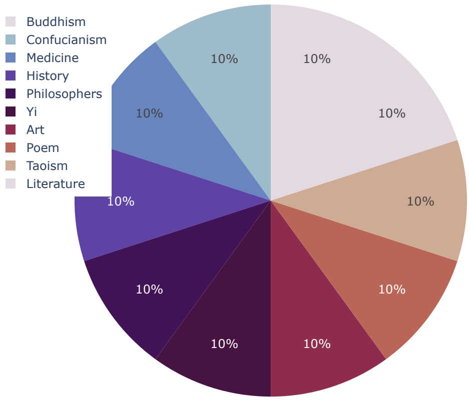

Figure 8: Percentage of labels to be predicted of GJC dataset.

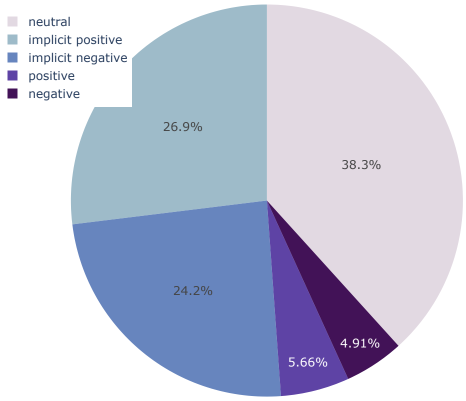

Figure 10: Percentage of labels to be predicted of FSPC dataset.

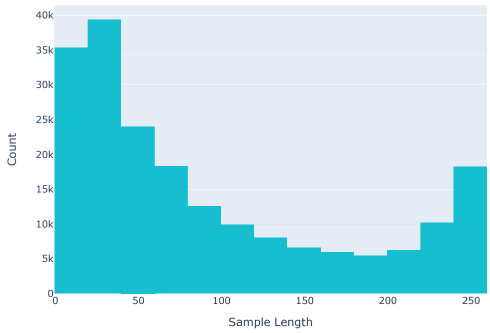

Figure 9: Statistic of Sample Length of GJC dataset.

B.5 FSPC

This dataset is in JSON format as shown below. The sentiment labels are of five specifications which shift from negative to positive. Statistics are show in Figure 10 Figure 11.

{
"poet": "范仲淹",
"poem": "静映寒林晚未芳|人人欲看寿阳妆|玉颜须傍韶春笑|莫斗严风与恶霜",
Quietly reflecting the cold, the forest is late and not fragrant | Everyone wants to see Shouyang makeup | Jade face must be close to spring smile | Do not fight with strong wind and cold frost.
"dynasty": "宋",
"sentiments": {
"holistic": "implicit positive",
"line1": "implicit positive",
"line2": "neutral",
"line3": "implicit positive",
"line4": "neutral"
},
"title": "和提刑赵学士探梅三绝"
},
{
"poet": "王维",
"poem": "独在异乡为异客|每逢佳节倍思亲|遥知兄弟登高处|遍插茱萸少一人",
Being alone and a stranger in a foreign land | I miss my relatives every time during the festival | I know my brothers climb a high place from afar | Everyone plant cornel all over the place but me
"dynasty": "唐",
"sentiments": {
"holistic": "implicit negative",
"line1": "implicit negative",
"line2": "implicit negative",
"line3": "neutral",
"line4": "implicit negative"
},

B.6 Xuci

This dataset is in TSV format. Statistics are show in Figure 12 Figure 13 and Table 11.

使夫邪污之气无由得接焉。[SEP] 复驾言兮焉求。[SEP] 10, 10 [SEP] 4, 4 [SEP] f
so that there is no way for those evil and filthy atmospheres to reach them. [SEP]What am I driving for?
上官令民送牛羊之陕西。[SEP] 久之，举于朝。[SEP] 7, 7 [SEP] 1, 1 [SEP] f
The superior commander sent cattle and

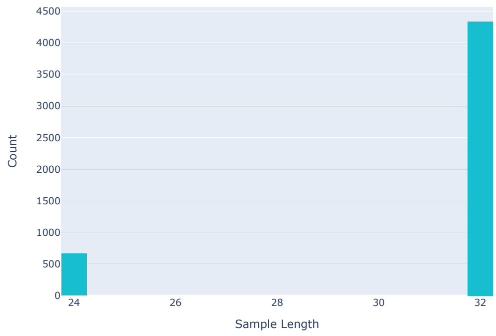

Figure 11: Statistic of Sample Length of FSPC dataset. Note that this dataset only contains five-character quatrains and seven-character quatrains.

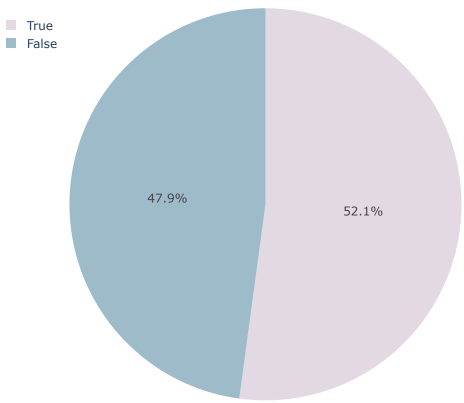

Figure 12: Percentage of labels to be predicted of Xuci dataset.

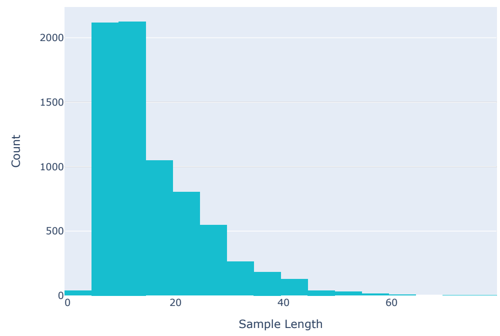

Figure 13: Statistic of Sample Length of Xuci dataset.

sheep to Shaanxi. [SEP] After a long time, he was recommended to the court. 容与乎阳林，流眄乎洛川。[SEP] 她也曾近乎撒娇似地问过他。[SEP] 2, 2 [SEP] 4, 4 [SEP] t
However, he calmly left the sun and looked at Luochuan with vast water. [SEP] She had also asked him almost coquettishly.

B.7 IRC

This dataset is in JSON format. Every sample consist of four fields which are "idiom", "options", "label" and "origin". The ground truth "label" is best fit of the four options. Statistics are show in Figure 14 Figure 15 and Table 12.

Total Samples	7350
Mean Sample Length	16
Min Sample Length	3
Max Sample Length	79

Table 11: Statistic of Sample Length of Xuci dataset.

{ "idiom": "眼去眉来",
eye to eyebrow
"options": [
"火烧到眉毛。比喻事到眼前，非常急迫。",
The fire burned to the eyebrows. The metaphor is very urgent.
"形容事情已到眼前，情势十分紧迫。",
Describe the matter has come to the front, the situation is very urgent.
"原指眼前见到的。后形容用眉眼传情。",
It meant what was seen. After describing the use of eyebrows teasing.
"形容眉眼含情示意的神态。"
Describe the expression of the eyebrows showing affection.
], "label": 2, "origin": "落日苍茫，风才定，片帆无力。还记得眉来眼去，水光山色。"
The setting sun is vast, the wind is fixed, and the sails are weak. I still remember the frowning, the water and the mountains. },

B.8 WYWMT

This dataset is in sentence pair TSV format. Samples are represented as "source" and "reference" segment which are separated by "tab". Statistics are show in Figure 16 and Table 14.

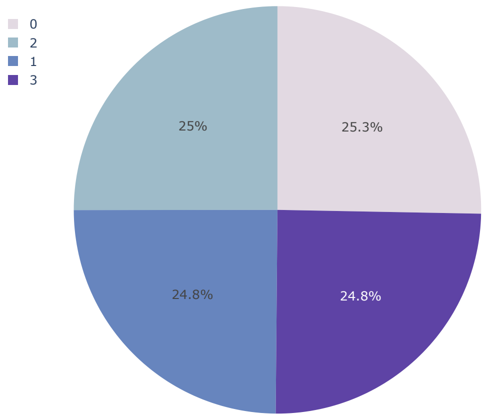

Figure 14: Percentage of labels to be predicted of IRC dataset.

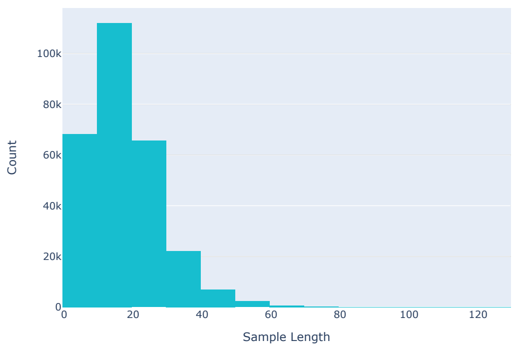

Figure 15: Statistic of Sample Length of IRC dataset.

共四里，又越一冈脊而下，其脊高不及高井之半，而实为西北来过脊以趋清秀者也。[SEP] 共四里，又越一道冈脊后下走，这个冈脊高处不到高井的一半，但实际上是从西北前来趋向清秀山的延伸而过的山脊。

After a total of four miles, we went down after another ridge. This ridge was less than half of the height of Gaojing, but it was actually a ridge extending from the northwest towards Qingxiu Mountain.

读性理书时，则杂以诗文各集，以歧其趋。[SEP] 在读性理书的时候，又掺杂写诗文，走了岔路。

When I read books about ethics, I mixed it with writing poetry, so I went to a wrong road.

	Idiom	Origin	1	2	3	4
Total	46471					
Mean	-	19	18	18	16	16
Min	4	5	3	3	4	4
Max	16	121	76	80	75	76

Table 12: Statistic of Sample Length of IRC dataset.

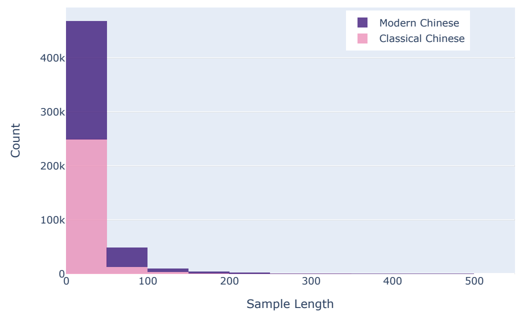

Figure 16: Statistic of Classical Sample Length of WYWMT dataset.

	Classical	Modern
Total Samples	46471	
Mean Sample Length	23	37
Min Sample Length	5	5
Max Sample Length	381	508

Table 14: Statistic of Sample Length of WYWMT dataset.

C Details of Models Evaluated

In this section, we present the details of pre-trained language models we used, including guwenbert-base, guwenbert-large, guwenbert-base-fs, roberta-classical-chinese-base-char, roberta-classical-chinese-large-char, SikuBERT, SikuRoBERTa, DeBERTa-base and RoBERTa-wwm-ext. As shown in 13, the masking, scale, corpus, vocabulary and parameter initialization are different in each pre-trained language model.

D Hyper-parameters for fine-tuning

As shown in Table 15, we present the hyper-parameters applied in fine-tuning. For different scale of pre-trained language model, we set different learning rates. In large scale, we set learning rates with 5e-6, 8e-6, 9e-16 and 1e-5. In base scale, we set learning rates from 1e-5 to 5e-5. We set warmup to 0.1, maximum epochs to 10. For Adam, we set ϵ to 1e-6, β_1 and β_2 to 0.9 and 0.999 respectively. Meanwhile, we use linear for LR decay and set weight decay to 0.01.

Model	Masking	Scale	Corpus	Optimizer	Vocabulary	Init.
guwenbert-base	WWM	base	DaiZhiGe	AdamW	23292	RoBERTa-wwm-ext
guwenbert-large	WWM	large	DaiZhiGe	AdamW	23292	RoBERTa-wwm-ext
guwenbert-base-fs	WWM	base	DaiZhiGe	AdamW	23292	Scrach Classical
roberta-classical-chinese-base-char	Mask	base	DaiZhiGe	AdamW	26318	guwenbert-base
roberta-calssical-chinese-large-char	Mask	large	DaiZhiGe	AdamW	26318	guwenbert-large
SikuBERT	Mask	base	Sikuquanshu	AdamW	29791	Scrach Classical
SikuRoBERTa	Mask	base	Sikuquanshu	AdamW	29791	Scrach Classical
DeBERTa-base	n-gram	base	DaiZhiGe	AdamW	22669	Scrach Classical
RoBERTa-wwm-ext	WWM	base	Chinese Corpus	AdamW	21128	Scrach Modern

Table 13: Parameters for pretraining of collected models.

Hyper-parameter	Large scale	Base scale
Dropout	{0,0.1,0.15}	{0,0.1,0.15}
Warmup	0.1	0.1
Learning Rates	{5e-6, 8e-6, 9e-6, 1e-5}	{1e-5 to 5e-5}
Batch Size	{16,32,48,64}	{16,32,48,64}
Weight Decay	0.01	0.01
Maximum Epochs	10	10
LR Decay	Linear	Linear
Adam ϵ	1e-6	1e-6
Adam β_1	0.9	0.9
Adam β_2	0.999	0.999
Gradient Clipping	1.0	1.0

Table 15: Hyper-parameters for fine-tuning.