AnchiBERT: A Pre-Trained Model for Ancient Chinese Language Understanding and Generation

Huishuang Tian[1], Kexin Yang[1], Dayiheng Liu[1], Jiancheng Lv[1(✉)]

[1]College of Computer Science, Sichuan University, 610065, Chengdu, China
tianhuishuang@stu.scu.edu.cn
lvjiancheng@scu.edu.cn

Abstract

Ancient Chinese is the essence of Chinese culture. There are several natural language processing tasks of ancient Chinese domain, such as ancient-modern Chinese translation, poem generation, and couplet generation. Previous studies usually use the supervised models which deeply rely on parallel data. However, it is difficult to obtain large-scale parallel data of ancient Chinese. In order to make full use of the more easily available monolingual ancient Chinese corpora, we release AnchiBERT, a pre-trained language model based on the architecture of BERT, which is trained on large-scale ancient Chinese corpora. We evaluate AnchiBERT on both language understanding and generation tasks, including poem classification, ancient-modern Chinese translation, poem generation, and couplet generation. The experimental results show that AnchiBERT outperforms BERT as well as the non-pretrained models and achieves state-of-the-art results in all cases.

1 Introduction

Ancient Chinese is the written language in ancient China, which has been used for thousands of years. There are large amounts of unlabeled monolingual ancient Chinese text in various forms, such as ancient Chinese articles, poems, and couplets. Investigating ancient Chinese is a meaningful and essential domain. Previous studies have made several attempts on it. For example, Liu et al. (2020) train a Transformer model to translate ancient Chinese into modern Chinese. Yan et al. (2016) and Yuan et al. (2019) apply an RNN-based model with attention mechanism to generate Chinese couplets. Yi et al. (2017a) generate ancient Chinese poems with RNN encoder-decoder framework. These ancient Chinese tasks often employ supervised models, which deeply rely on the scale of parallel datasets.

Ancient Chinese	文武争驰，君臣无事，可以尽豫游之乐。
Modern Chinese	文臣武将争先恐后前来效力，国君和大臣没有大事烦扰，国君就可以尽情享受安逸的生活。
English	Civil servants and military generals work hard, then the monarch and ministers could enjoy a comfortable life without any disturbance.

Figure 1: Linguistic characteristics shift between modern Chinese and ancient Chinese.

However, those datasets are costly and difficult to obtain due to the requirement for expert annotation.

In the absence of parallel data, previous studies have proposed pre-trained language models to utilize the large-scale unlabeled corpora to further improve the model performance on NLP tasks, such as ELMo (Peters et al., 2018), GPT (Radford et al., 2018), and BERT (Devlin et al., 2019). These pre-trained models learn universal language representations from large-scale corpora with self-supervised objectives, and then are fine-tuned on downstream tasks. However, these models are trained on general-domain text which has linguistic characteristics shift from ancient Chinese text. The shift between modern Chinese and ancient Chinese is shown in figure 1.

Therefore, we propose AnchiBERT, a pre-trained language model based on the architecture of BERT, which is trained on the large-scale ancient Chinese corpora. We evaluate the performance of AnchiBERT on both language understanding and generation tasks. Our contributions are as follows:

- To our best knowledge, we propose a first pre-trained language model in ancient Chinese domain, which is trained on the large-scale ancient Chinese corpora we build.
- We evaluate the performance of AnchiBERT

on four ancient Chinese downstream tasks, including both language understanding and language generation tasks. AnchiBERT achieves new state-of-the-art results in all tasks which verify the effectiveness of pre-training strategy in ancient Chinese domain.

- We propose a complete pipeline to apply pre-trained model into several ancient Chinese domain tasks. We will release our code, pre-trained model, and corpora[1] to facilitate the further research on ancient Chinese domain tasks.

2 Related Works

2.1 Pre-Trained Representations in General

Pre-training is an effective strategy which is widely used for NLP tasks in recent years. As static representations, Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) are the early word-level methods to learn language representations. As dynamic representations, ELMo (Peters et al., 2018) provides the contextual representations based on a bidirectional language model. ELMo is pre-trained on huge text corpus and can learn better contextualized word embeddings for downstream tasks. GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) propose pre-trained Transformer-based model to learn universal language representations by fine-tuning on large-scale corpora. Compared to GPT, BERT is trained on masked token prediction and next sentence prediction task, which extracts bidirectional information instead of unidirectional. Moreover, recent studies propose new pre-trained models, such as XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020), which bring improvements on downstream tasks.

2.2 Domain-Specific pre-trained Models

Several studies propose pre-trained models which adapt to specific domains or tasks. BioBERT (Lee et al., 2020) is trained on large-scale biomedical text for biomedical domain tasks. SciBERT (Beltagy et al., 2019) is trained for scientific domain tasks on biomedical and computer science text, using its own vocabulary (SCIVOCAB). ClinicalBERT (Alsentzer et al., 2019) is proposed due to the need for specialized clinical pre-trained model and is applied to clinical tasks. In addition, recent studies also release monolingual pre-trained models for a specific language besides English. FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2020) are trained for French. BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020) are trained for Dutch. AraBERT (Antoun et al., 2020) is trained for Arabic language.

2.3 Ancient Chinese Domain Tasks

Ancient Chinese domain tasks include translating ancient Chinese into modern Chinese, generating poems, generating couplets, and so on. For translation, Liu et al. (2020) translate ancient Chinese into modern Chinese with a Transformer model. For poem generation, several studies are based on templates and rules (Tosa et al., 2008; Wu et al., 2009; Manurung et al., 2012). With the development of deep learning, some approaches generate poems with an encoder-decoder framework (Wang et al., 2016; Yi et al., 2017b; Liu et al., 2018). Moreover, many new model methods are applied to poem generation, such as reinforcement learning (Yi et al., 2018) and variational autoencoder (Yang et al., 2018). For couplet generation, Jiang and Zhou (2008) use a statistical machine translation approach. Yan et al. (2016) and Yuan et al. (2019) apply an RNN-based model with attention mechanism to generate couplets. However, these tasks use limited annotated data and leave the large-scale unlabeled ancient Chinese text behind. We utilize the unlabeled data to train AnchiBERT, a pre-trained model which adapts to ancient Chinese domain. AnchiBERT achieves SOTA results in all downstream tasks.

3 Method

3.1 Model Architecture

AnchiBERT exactly follows the same architecture as BERT (Devlin et al., 2019), using a multi-layer Transformer (Vaswani et al., 2017). AnchiBERT uses the configuration of BERT-base, with 12 layers, the hidden size of 768, and 12 attention heads. The total number of model parameters is about 102M.

3.2 Pre-Training Data

The ancient Chinese corpora used for training AnchiBERT are listed in Table 1. The corpora consist of articles, poems and couplets which are written in ancient Chinese, resulting in the corpora size of

[1]The dataset and model will be available at https://github.com/xxxxxx

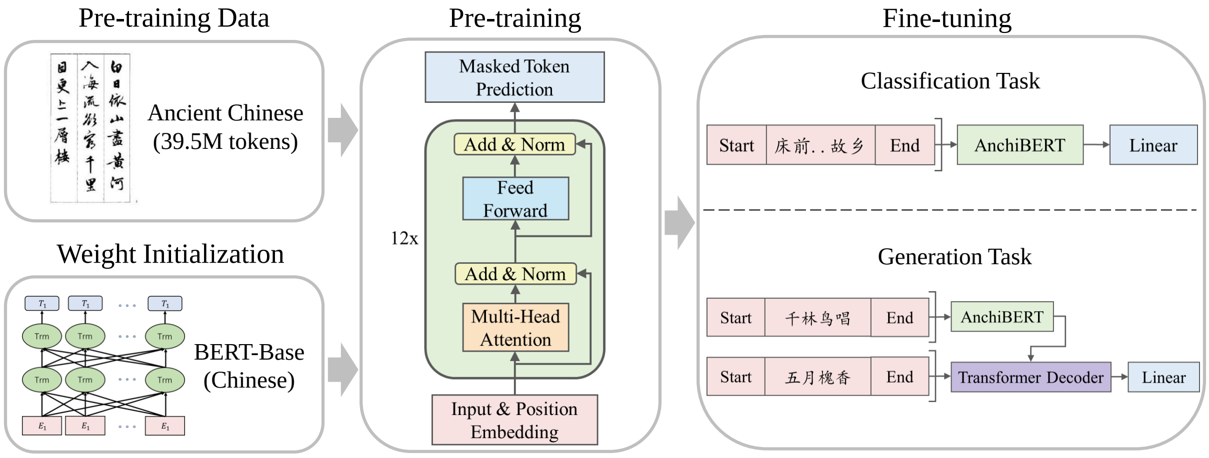

Figure 2: Overview of pre-training and fine-tuning process of AnchiBERT.

Corpus Type	Number of Tokens
Ancient Chinese Article	16.9M
Ancient Chinese Poetry	6.7M
Ancient Chinese Couplet	15.9M

Table 1: Pre-training data used for AnchiBERT.

39.5M ancient Chinese tokens. Most of our ancient Chinese corpora are written in dynasties of ancient China by many celebrities (about 1000BC-200BC).

We preprocess the raw data crawled from the Internet[2]. We first clean the data by removing the useless symbols. Then we convert traditional Chinese characters into simplified characters. Finally, we remove the titles of articles and poems and only leave the bodies.

3.3 Pre-Training AnchiBERT

Instead of training from scratch, AnchiBERT continues pre-training based on the BERT-base (Chinese) model[3] on our ancient Chinese corpora, as shown in Figure 2. We use masked token prediction task (MLM) to train AnchiBERT. Following Devlin et al. (2019), given a text sequence $x = \{x_1, x_2, ..., x_n\}$ as input, we randomly mask 15% of the tokens from x. During pre-training, 80% of those selected tokens are replaced with [MASK] token, 10% are replaced with a random token, and 10% are unchanged. The training objective is to predict the masked tokens with cross entropy loss. We do not use next sentence prediction (NSP) task because previous work shows this objective does not improve downstream task performance (Liu et al., 2019).

Following Devlin et al. (2019), we optimize the MLM loss using Adam (Kingma and Ba, 2015) with a learning rate of 1e-4 and weight decay of 0.01. Due to the limited memory of GPU we train the model with batch size of 15. The maximum sentence length is set to 512 tokens.

We adopt the original tokenization script[4] and tokenize text based on the granularity of Chinese character, where a Chinese character denotes a token. We use the originally released vocabulary in BERT-base (Chinese).

3.4 Fine-Tuning AnchiBERT

For ancient Chinese understanding task, we apply a classification layer atop AnchiBERT. For ancient Chinese generation tasks, we use a Transformer-based encoder-decoder framework, which employs AnchiBERT as encoder and uses a transformer decoder with random initialization parameters. Details can be found in § 4.2.

4 Experiments

In this section, we first describe the pre-training details of AnchiBERT, and then introduce the task objective, dataset, settings, baselines, and metrics of each downstream task.

4.1 AnchiBERT Pre-training

AnchiBERT continues pre-training from BERT-base (Chinese) on our ancient Chinese corpora

[2]Part of the ancient Chinese text comes from the website http://www.gushiwen.org and http://wyw.5156edu.com.

[3]https://github.com/huggingface/transformers

[4]https://github.com/huggingface/transformers/blob/master/src/transformers/tokenization_bert.py

Task	Data(train/dev/test)
PTC	2.8K/0.2K/0.2K
AMCT	1.0M/125.7K/100.6K
CPG	0.22M/5.4K/5.4K
CCG	0.77M/4.0K/4.0K

Table 2: Train/dev/test dataset sizes (number of pairs) of each task.

rather than from scratch. AnchiBERT follows the same configuration as BERT-base. Details of model configuration and pre-training data are in § 3.1 and § 3.2 respectively.

During training, we set the maximum sentence length of 512 tokens to train the model on masked token prediction task with Adam optimizer. The batch size is 15 and training steps are 250K. We use 3 RTX 2080ti GPUs for training. AnchiBERT training takes about 3 days. Our code is implemented based on the Pytorch-Transformers library released by huggingface[5] (Wolf et al., 2019).

4.2 AnchiBERT Fine-tuning

4.2.1 Poem Topic Classification (PTC)

Given a poem, the objective of Poem Topic Classification (PTC) task is to obtain the corresponding literary topic. We fine-tune and evaluate AnchiBERT on a publicly released dataset[6]. The dataset contains 3.2K four-line classical Chinese poems combined with titles and keywords, and each poem has one annotated literary topic (e.g., farewell poem, warfare poem). Details of data splits are shown in table 2.

For training settings, we feed the final hidden vector corresponding to [CLS] token into a classification layer to obtain the topic label, as figure 2 shows. The input is the body of a poem and output is the corresponding topic label. We apply a batch size of 24 and use Adam optimizer with a learning rate of 5e-5. The dropout rate is always 0.1. The number of training epoch is around 5.

We compare our AnchiBERT with the following baselines:

1. Std-Transformer: Std-Transformer is a standard Transformer encoder following the same architecture and configuration as official BERT-base (Chinese), such as the number of layers and hidden size. The vocabulary is the same as well. However, the training weights are randomly initialized instead of pre-trained.

2. BERT-Base: We choose the pre-trained weights of official version BERT-base (Chinese) (Devlin et al., 2019) to initialize BERT-Base. We adopt the original vocabulary.

For automatic evaluation metric, we evaluate models on classification accuracy.

4.2.2 Ancient-Modern Chinese Translation (AMCT)

Ancient-Modern Chinese Translation (AMCT) task translates ancient Chinese sentences into modern Chinese, because ancient Chinese is difficult for modern people to understand. We conduct experiments on ancient-modern Chinese dataset (Liu et al., 2020). This dataset contains 1.2M aligned ancient-modern Chinese sentence pairs, with ancient Chinese sentence as input and modern Chinese as target.

For training settings, this task is based on encoder-decoder framework. As figure 2 shows, we initialize the encoder with AnchiBERT and use a Transformer-based decoder, which is randomly initialized. Following the framework of Transformer, our decoder generates text conditioned on encoder hidden representations through multi-head attention. The training objective is to minimize the negative log likelihood of the generated text.

The training batch size and the layer number of decoder is 30 and 4, respectively. We use the same optimizer as Transformer, with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1\text{e-}9$ and a linear warmup over 4000 steps. The dropout rate is 0.1. We choose the best number of epoch on the Dev set.

We compare our AnchiBERT with the following baselines:

1. Transformer-A: Transformer-A (Liu et al., 2020) is a Transformer model with augmented data of ancient-modern Chinese pairs.

2. Std-Transformer: Std-Transformer follows the framework of Transformer, with an encoder identical to Std-Transformer in § 4.2.1 and a randomly initialized decoder.

3. BERT-Base: BERT-Base follows the framework of Transformer, with an encoder identical to BERT-Base in § 4.2.1 and a randomly initialized decoder.

[5] https://github.com/huggingface/transformers

[6] https://github.com/shuizhonghaitong/classification_GAT/tree/master/data

For automatic evaluation metric, we adapt BLEU evaluation (Papineni et al., 2002) which compares the quality of generated sentences with the ground truth. We apply BLEU-4 in this task.

We also include human evaluation for generation tasks because the above automatic evaluation metric has some flaws. For example, given an ancient Chinese sentence, there is only one ground truth. But in fact there are more than one appropriate ways of expression for modern Chinese. Thus we follow the evaluation standards in (Yan et al., 2016), and invite 10 evaluators to rank the generations in two aspects: syntactic and semantic. As for syntactic, evaluators evaluate whether the composition of translated modern Chinese is complete. As for semantic, evaluators consider whether the generated sentences are coherent and fluent. The score is assigned with 0 and 1, with 1 meaning good.

4.2.3 Chinese Poem Generation (CPG)

In Chinese Poem Generation (CPG) task, we implement two experimental settings. The first task is to generate the last two lines of a poem from the first two lines (2-2), the second task is to generate the last three lines from the first line (1-3). These four lines of a poem should match each other by following the syntactic and semantic rules in ancient Chinese poems. We use another publicly available poetry dataset[7] for experiment, which contains 0.23M four-line classical Chinese poems.

For training settings, this task uses the same encoder-decoder framework and loss function as AMCT described in § 4.2.2. We apply a batch size of 80 and a 2-layer randomly initialized decoder. We use the same optimizer as AMCT in § 4.2.2. We choose the best number of epoch on the Dev set.

We compare our AnchiBERT with the following baselines:

1. Std-Transformer: Std-Transformer follows the framework of Transformer, with an encoder identical to Std-Transformer in § 4.2.1 and a randomly initialized decoder.

2. BERT-Base: BERT-Base follows the framework of Transformer, with an encoder identical to BERT-Base in § 4.2.1 and a randomly initialized decoder.

For automatic evaluation metric, we use BLEU-4 in this task. Meanwhile, we follow the human metric in § 4.2.2 to evaluate the generated poems in syntactic and semantic. Especially, for syntactic, evaluators consider whether the generated poem sentences conform to the length and rhyming rules.

4.2.4 Chinese Couplet Generation (CCG)

Chinese Couplet Generation (CCG) task generates the second sentence (namely a subsequent clause) of couplet, given the first sentence (namely an antecedent clause) of couplet. We conduct this experiment on a publicly available couplet dataset[8], which contains 0.77M couplet pairs.

For training settings, we use the same model architecture and loss function described in § 4.2.2. The batch size is 80 and the layer number of decoder is 4. We use the same optimizer in § 4.2.2 and fine-tune for around 60 epochs.

We compare our AnchiBERT with the following baselines:

1. RNN-based Models: We first implement the basic LSTM and Seq2Seq model. We also include SeqGAN model (Yu et al., 2017), which applies reinforcement learning into Generative Adversarial Net (GAN) to solve the problems in generating discrete sequence tokens. Furthermore, NCM (Yan et al., 2016) is an RNN-based Seq2Seq model incorporating the attention mechanism. NCM also includes a polishing schema, which generates a draft first and then refines the wordings.

2. Std-Transformer: Std-Transformer follows the framework of Transformer, with an encoder identical to Std-Transformer in § 4.2.1 and a randomly initialized decoder.

3. BERT-Base: BERT-Base follows the framework of Transformer, with an encoder identical to BERT-Base in § 4.2.1 and a randomly initialized decoder.

For automatic evaluation metric, because the generated couplet sentences are often less than 10 tokens, we use BLEU-2 in CCG task. Meanwhile, we use the human evaluation metric in § 4.2.2 to evaluate couplet in syntactic and semantic. For syntactic, the generated subsequent clauses should conform to the length and pattern rules.

[7] https://github.com/chinese-poetry/chinese-poetry

[8] https://github.com/wb14123/couplet-dataset

Model	AMCT		CPG (2-2)		CPG (1-3)		CCG		Average
	Syntactic	Semantic	Syntactic	Semantic	Syntactic	Semantic	Syntactic	Semantic	
Std-Transformer	0.63	0.58	0.69	0.60	0.63	0.52	0.61	0.59	0.61
BERT-Base	0.69	0.61	0.72	0.64	0.67	0.54	0.63	0.62	0.64
AnchiBERT	**0.71**	**0.62**	**0.73**	**0.65**	**0.69**	**0.55**	**0.65**	**0.63**	**0.65**

Table 3: Human evaluation results of generation tasks.

Task	Model	BLEU-4
AMCT	Transformer-A	27.16
	Std-Transformer	27.80
	BERT-Base	28.89
	AnchiBERT	**31.22**
CPG (2-2)	Std-Transformer	27.47
	BERT-Base	29.82
	AnchiBERT	**30.08**
CPG (1-3)	Std-Transformer[9]	19.52
	BERT-Base	21.63
	AnchiBERT	**22.10**

Table 4: Evaluation results on AMCT and CPG tasks. For CPG task, we implement two experimental settings, including generating the last two sentences from the first two sentences (2-2) and generating the last three sentences from the first sentence (1-3).

Task	Model	BLEU-2
CCG	LSTM	10.18
	Seq2Seq	19.46
	SeqGAN	10.23
	NCM	20.55
	Std-Transformer	27.14
	BERT-Base	33.01
	AnchiBERT	**33.37**

Table 5: Evaluation results on CCG task, we apply BLEU-2 as evaluation metric.

5 Results

The experiment results are shown in tables above. Generally, we find that AnchiBERT outperforms BERT-Base as well as the non-pretrained models on all ancient Chinese domain tasks. AnchiBERT also achieves new SOTA results in all cases.

5.1 Automatic Evaluation Results

The accuracy (the higher the better) is shown in table 6 and BLEU (the higher the better) results are shown in table 4 and table 5 respectively.

Poem Topic Classification Table 6 shows AnchiBERT achieves the SOTA result in Poem Topic Classification task. AnchiBERT improves accuracy by 6.99 over BERT-Base and 12.34 over Std-Transformer. Because the scale of this task dataset is very small, the result illustrates pre-training, especially domain-specific pre-training can significantly improve performance on low-resource task.

Model	Accuracy (%)
Std-Transformer	69.96
BERT-Base	75.31
AnchiBERT	**82.30**

Table 6: Results on Poem Topic Classification task. Accuracies (%) are reported.

Ancient-Modern Chinese Translation Table 4 shows AnchiBERT outperforms all the baseline models in Ancient-Modern Chinese Translation task. AnchiBERT raises the BLEU score by 2.33 points over BERT-Base and 3.42 over Std-Transformer, which demonstrates the effectiveness of domain-specific pre-training in language generation task.

Chinese Poem Generation AnchiBERT improves performance over two variants (BERT-Base and Std-Transformer) in both experimental settings. In CPG (2-2), AnchiBERT reaches a slightly higher score by 0.26 than BERT-Base and +2.62 than Std-Transformer. In CPG (1-3), AnchiBERT reaches +0.47 over BERT-Base and +2.58 over Std-Transformer.

Chinese Couplet Generation Table 5 shows AnchiBERT outperforms all of the non-pretrained baseline models and two variants (+0.36 over BERT-Base and +6.23 over Std-Transformer). Note that the task-specific model NCM performs better than general model Std-Transformer, which demonstrates the need for task-specific model architectures. However, the pre-trained models

[9] The performance of Std-Transformer (12 layers of encoder) is extremely poor for CPG (1-3), so we train a randomly initialized Transformer (6 layers of encoder) for this experimental setting and present the best result.

Ancient-Modern Chinese Translation	
Ancient Sentence	听诸生诵书；听已，辄默记。(After listening to other students reading books, he always carefully memorizes what they read.)
Ground Truth	去听学生念书。听完以后，总是默默地记住。(After listening to other students reading books, he always carefully memorizes what they read.)
Std-Trans	听由众儒生诵读书，都默记下来。(He listens to other students reading books and memorizes what they read.)
Bert-Base	听任诸儒生诵读书，就默默记在心里。(He listens to other students reading books, then memorizes what they read.)
AnchiBERT	听学生们诵读书。听完之后，就默默地记在心里。(After listening to other students reading books, he always carefully memorizes what they read.)

Chinese Couplet Generation	
Antecedent Clause	千林鸟唱，四季风和，喜看万木连云翠。(Birds sing in the forest. It's sunny in all seasons. People are happy to see green woods and white clouds.)
Ground Truth	五月槐香，三河水碧，更引九龙破壁飞。(The fragrance of sophora japonicae in May and the blue water in the rivers lead to nine dragons flying.)
Std-Trans	五岭松青，五湖水碧，更引五湖映月明。(Pine trees are green on mountains, and five lakes are green. These lead to moon reflected in the lake.)
Bert-Base	百里花开，一湖水暖，更引九龙破壁飞。(Hundreds of flowers blossom. A lake is warm. These lead to nine dragons flying.)
AnchiBERT	五月荷香，三秋桂馥，更引九州带露红。(The fragrance of lotus flowers in May and Osmanthus in autumn leads to the land flourishing.)

(a)

Chinese Poem Generation (2-2)	
First Two Lines	空有丹心贯碧霄，泮冰亡国不崇朝。(My sincerity is in vain. The country will eventually perish.)
Ground Truth	小臣万死无遗慨，曾见天家十八朝。(I have no regret even if I die, because I have seen a prosperous dynasty.)
Std-Trans	小臣欲献天王赐，天子无私二十朝。(I want to send a gift to the emperor, so that the emperor will always be in power.)
Bert-Base	太平无限风流事，曾见君王万世朝。(People live happily in the world, and I have seen a prosperous dynasty.)
AnchiBERT	万家有母无人识，曾向天家几度朝。(life is falling apart in many families, and I have seen a prosperous dynasty.)

Chinese Poem Generation (1-3)	
First Line	云重时飞雪 (It's cloudy and snowy now and then.)
Ground Truth	春迟未见梅。山园多芋栗，夜话且燔煨。(Spring comes late, and the plum trees are not in bloom. There are taros in the yard, and people are talking and baking together.)
Std-Trans	风高已度春。何当见尧舜，为作太平民。(The wind is strong and spring is coming. When can I see Yao and Shun ? So that I can live in peace and prosperity.)
Bert-Base	春寒未见梅。相思还怅望，旅食不须催。(It's cold in spring and the plum trees are not in bloom. Missing friends makes me sad, and I am unable to eat any more.)
AnchiBERT	春迟未见梅。道人多失计，夜夜听松风。(Spring comes late, and the plum trees are not in bloom. I am frustrated, listening to the wind blowing pine trees every night.)

(b)

Figure 3: Sample comparison of generation tasks. 'Std-Trans' in the figure is short for Std-Transformer.

(AnchiBERT and BERT-Base) outperform NCM. This illustrates that sometimes simple pre-trained model is better than complex model architectures.

Our goal of proposing AnchiBERT is to confirm the performance of pre-training strategy in ancient Chinese domain. As we expect, all pre-trained models (AnchiBERT and BERT-Base) perform better than non-pretrained baselines. Meanwhile, AnchiBERT achieves new SOTA results on all ancient Chinese domain tasks.

5.2 Human Evaluation Results

Table 3 reports the human evaluation results on generation tasks. We only compare with BERT variants (Std-Transformer and BERT-Base) because we focus on the effectiveness of domain-specific pre-training. For each experiment, we collect 20 generations respectively. We invite 10 evaluators who are proficient in Chinese literature.

In general, the average results demonstrate our model AnchiBERT outperforms all variants. The syntactic scores of our pre-trained AnchiBERT show that although no templates or rules (such as rhythm and length for poem) are set in the AnchiBERT model explicitly, the model can automatically generate text conforming to these grammatical rules. The semantic scores indicate that AnchiBERT learns semantic rules during pre-training, so in downstream tasks AnchiBERT can generate more coherent text across sentences. Note that BERT-Base achieves similar scores with AnchiBERT, which demonstrates pre-training on general-domain text is efficient as well.

5.3 Samples Analysis

Figure 3 shows some samples of ancient Chinese translation, poem generation and couplet generation. In the generation tasks, we observe that the inability of Std-Transformer to learn language representation leads to the lack of coherence in generated sentences. BERT-Base learns representation from modern Chinese corpus, so it performs slightly worse for ancient Chinese. AnchiBERT is able to generate ancient Chinese sentences which is coherent and meaningful.

For example, in Ancient-Modern Chinese Translation task, ancient sentence '听已' (after listening) is translated into '听完以后' (after listening). However, Std-Transformer and BERT-Base ignore this sentence, whereas AnchiBERT makes the translation. In Chinese Poem Generation (2-2), the original ground truth describes the patriotism of the author. However, the generated sentences of Std-Transformer do not have this meaning. Meanwhile, the first generated sentence of BERT-Base describes the life of ordinary people, which has a semantic shift from the ground truth. AnchiBERT generates sentences which express the heavy atmosphere and the expectations for a prosperous dynasty and fit the poem topic well.

5.4 Discussion

We observe that pre-training is an effective strategy in ancient Chinese domain, not only in language understanding but also in language generation tasks. On automatic evaluation, AnchiBERT performs better than BERT-base in all ancient Chinese domain tasks, and significantly outperforms the non-pretrained models. Human evaluators also think that AnchiBERT is able to generate text which follows grammatical rules better and is more fluent for people to read.

6 Conclusion

In this paper, we release AnchiBERT, the first pre-trained language model in ancient Chinese domain to the best of our knowledge. AnchiBERT is based on BERT and trained on ancient Chinese corpora. We evaluate AnchiBERT on downstream language understanding and generation tasks, which achieves state-of-the-art performance.

There are some directions for future research. First, find a suitable learning objective during pre-training in ancient Chinese domain. Then, find more ancient Chinese data and construct an ancient Chinese domain vocabulary to train AnchiBERT.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China under Contract 2017YFB1002201, in part by the National Natural Science Fund for Distinguished Young Scholar under Grant 61625204, and in part by the State Key Program of the National Science Foundation of China under Grant 61836006.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem M. Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *CoRR*, abs/2003.00104.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *CoRR*, abs/2001.06286.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Long Jiang and Ming Zhou. 2008. Generating chinese couplets using a statistical MT approach. In *COLING 2008, 22nd International Conference on Computational Linguistics, Proceedings of the Conference, 18-22 August 2008, Manchester, UK*, pages 377–384.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2479–2490. European Language Resources Association.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Dayiheng Liu, Quan Guo, Wubo Li, and Jiancheng Lv. 2018. A multi-modal chinese poetry generation model. In *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pages 1–8. IEEE.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2020. Ancient-modern chinese translation with a new large training dataset. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(1):6:1–6:13.

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. *CoRR*, abs/1907.11692.

Ruli Manurung, Graeme Ritchie, and Henry Thompson. 2012. Using genetic algorithms to create meaningful poetic text. *J. Exp. Theor. Artif. Intell.*, 24(1):43–64.

Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7203–7219. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Naoko Tosa, Hideto Obara, and Michihiko Minoh. 2008. Hitch haiku: An interactive supporting system for composing haiku poem. In *Entertainment Computing - ICEC 2008, 7th International Conference, Pittsburgh, PA, USA, September 25-27, 2008. Proceedings*, volume 5309 of *Lecture Notes in Computer Science*, pages 209–216. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch BERT model. *CoRR*, abs/1912.09582.

Qixin Wang, Tianyi Luo, and Dong Wang. 2016. Can machine generate traditional chinese poetry? A feigenbaum test. In *Advances in Brain Inspired Cognitive Systems - 8th International Conference, BICS 2016, Beijing, China, November 28-30, 2016, Proceedings*, volume 10023 of *Lecture Notes in Computer Science*, pages 34–46.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Xiao-feng Wu, Naoko Tosa, and Ryohei Nakatsu. 2009. New hitch haiku: An interactive renku poem composition supporting tool applied for sightseeing navigation system. In *Entertainment Computing - ICEC 2009, 8th International Conference, Paris, France, September 3-5, 2009. Proceedings*, volume 5709 of *Lecture Notes in Computer Science*, pages 191–196. Springer.

Rui Yan, Cheng-Te Li, Xiaohua Hu, and Ming Zhang. 2016. Chinese couplet generation with neural network structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Xiaopeng Yang, Xiaowen Lin, Shunda Suo, and Ming Li. 2018. Generating thematic chinese poetry using conditional variational autoencoders with hybrid decoders. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4539–4545. ijcai.org.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. 2017a. Generating chinese classical poems with RNN encoder-decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data - 16th China National Conference, CCL 2017, - and - 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings*, volume 10565 of *Lecture Notes in Computer Science*, pages 211–223. Springer.

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. 2017b. Generating chinese classical poems with RNN encoder-decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data - 16th China National Conference, CCL 2017, - and - 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings*, volume 10565 of *Lecture Notes in Computer Science*, pages 211–223. Springer.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018. Automatic poetry generation with mutual reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3143–3153. Association for Computational Linguistics.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2852–2858. AAAI Press.

Shengqiong Yuan, Luo Zhong, Lin Li, and Rui Zhang. 2019. Automatic generation of chinese couplets with attention based encoder-decoder model. In *2nd IEEE Conference on Multimedia Information Processing and Retrieval, MIPR 2019, San Jose, CA, USA, March 28-30, 2019*, pages 65–70. IEEE.