rT5: A Retrieval-Augmented Pre-trained Model for Ancient Chinese Entity Description Generation

Mengting Hu[1], Xiaoqun Zhao[1], Jiaqi Wei[1], Jianfeng Wu[2], Xiaosu Sun[1], Zhengdan Li[1], Yike Wu[3,4(✉)], Yufei Sun[1], and Yuzhi Zhang[1]

[1] College of Software, Nankai University, Tianjin, China
mthu@nankai.edu.cn
[2] College of Computer Science, Nankai University, Tianjin, China
[3] School of Journalism and Communication, Nankai University, Tianjin, China
wuyike@nankai.edu.cn
[4] Convergence Media Research Center, Nankai University, Tianjin, China

Abstract. Ancient Chinese, the natural language of ancient China, serves as the key to understanding and propagating Chinese rich history and civilization. However, to facilitate comprehension and education, human experts previously need to write modern language descriptions for special entities, such as persons and locations, out of ancient Chinese texts. This process requires specialized knowledge and can be time-consuming. To address these challenges, we propose a new task called Ancient Chinese Entity Description Generation (ACEDG), which aims to automatically generate modern language descriptions for ancient entities. To address ACEDG, we propose two expert-annotated datasets, XunZi and MengZi, each containing ancient Chinese texts, and some of them have been annotated with entities and their descriptions by human experts. To leverage both labeled and unlabeled texts, we propose a retrieval-augmented pre-trained model called rT5. Specifically, a pseudo-parallel corpus is constructed using retrieval techniques to augment the pre-training stage. Subsequently, the pre-trained model is fine-tuned on our high-quality human-annotated entity-description corpus. Our experimental results, evaluated using various metrics, demonstrate the effectiveness of our method. By combining retrieval techniques and pre-training, our approach significantly advances the state-of-the-art performance in the ACEDG task compared with strong pre-trained models.

Keywords: Ancient Chinese · Entity Description Generation

1 Introduction

Throughout the extensive history of China, ancient Chinese texts are the essence of national thought and cultural spirit. They serve as a bridge connecting history

This research is supported by the youth program of National Science Fund of Tianjin, China (Grant No. 22JCQNJC01340), the Fundamental Research Funds for the Central University, Nankai University (Grant No. 63221028 and No. 63232114).

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14302, pp. 736–748, 2023.
https://doi.org/10.1007/978-3-031-44693-1_57

with modern culture and transmit valuable information to future generations. Notably, ancient Chinese differs from modern Chinese in various aspects, including vocabulary and syntax. To aid in comprehension and education, there are several natural language processing tasks focused on ancient Chinese, such as named entity recognition [1], poem generation [2], and couplet generation [3].

Different from these works, we deal with ancient Chinese from the entity description view. That is to say, human experts previously need to annotate ancient Chinese entities with modern Chinese descriptions, which is very time-consuming and relies heavily on experts' knowledge. To save human effort, we propose a new task called Ancient Chinese Entity Description Generation (ACEDG), aiming to automatically generate modern language descriptions for ancient entities. As shown in Fig. 1, "申徒狄" (Shentu Di) is a person entity in the ancient Chinese sentence. ACEDG aims to generate a modern Chinese description considering the contexts. In this example, the entity description presents a brief biography of this person.

Ancient Chinese	故怀负石而赴河，是行之难为者也，而申徒狄能之。 Therefore, it is difficult to go to the river with stones, but the Shentu Di can.
Entity-Description	商朝末年官吏。亦作申屠狄。因不忍见商纣王无道，谏而未被采纳，负石投河而死。 Officials in the last years of the Shang Dynasty. Also known as Shentu Di. Unable to bear to see that King Zhou of the Shang Dynasty had no way, his admonition was not accepted, and he threw himself into the river and died.
Retrieval Result	或称司徒狄。商朝人。狄向君王强烈谏言而不被采纳，其宁愿赴河也不愿意背叛自己的国家。 Or Stuti. People of the Shang Dynasty. Di made a strong recommendation to the king but was not accepted. He would rather go to the river than betray his country.

Fig. 1. An example of ACEDG. Our purpose is to generate modern Chinese descriptions from ancient entities. The retrieval result demonstrates that retrieved texts have the potential of providing extra knowledge to help generate.

The above example illustrates that interpreting ancient Chinese texts requires not only contextual information but also expert knowledge. However, acquiring such knowledge relies heavily on human experts with a solid background in history and literature. Additionally, accurately conveying the importance of the entity category presented in the original sentence poses another challenge. This is due to the nature of the language itself, as ancient Chinese is markedly distinct from modern Chinese in terms of sentence structure and vocabulary. Ancient Chinese sentences feature complex combinations of content and function words, alongside colloquial characters, making them more challenging to comprehend than modern Chinese.

To deal with the above challenges, we propose a retrieval-augmented pre-trained model called rT5. Concretely, we first construct pseudo-parallel corpora based on different retrieval algorithms. By comparing different evaluation

metrics, the corpus generated by the best retrieval algorithm is chosen as the pre-training data for the model. Subsequently, we utilize retrieval techniques to augment the pre-training phase and fine-tune the pre-trained model on our high-quality human-annotated entity-description corpus. To verify the effectiveness of our method, we conduct experiments on two expert-annotated datasets, XunZi and MengZi. Each dataset contains ancient Chinese texts, and some of them have been annotated with entities and their descriptions by human experts. Extensive experimental results suggest the effectiveness of our method. By combining retrieval techniques and pre-training with fine-tuning, our approach significantly advances the state-of-the-art performance in the ACEDG task compared with strong pre-trained models.

In summary, the contributions of this work are three-fold:

- We propose a new task called Ancient Chinese Entity Description Generation (ACEDG) and propose two expert-annotated datasets, XunZi and MengZi. To the best of our knowledge, though there are some works dealing with ancient Chinese, we are the first to focus on ancient Chinese entity description generation.
- To leverage both labeled and unlabeled texts, we propose a retrieval-augmented pre-trained model called rT5 for promoting ACEDG task.
- Our experimental results demonstrate the effectiveness of our method. By combining retrieval techniques and pre-training, our approach significantly advances the state-of-the-art performance in the ACEDG task compared with strong pre-trained models.

Table 1. The statistics of entity-description pairs for each type in the datasets.

category	Person	Location	Thing	Literature	Official	Institution	Time	Knowledge
XunZi	487	114	78	689	77	9	35	3822
MengZi	141	14	52	135	4	0	0	126

2 Dataset Construction

Two kinds of corpora are constructed in the experiments, including expert-annotated datasets and retrieval-based pre-trained corpus. The first one is high-quality ancient Chinese entities and corresponding entity descriptions. Due to its high cost, the scale is relevantly small. Then the second one is constructed based on retrieval techniques, which is cost-friendly.

Expert-Annotated Datasets. To address ACEDG, we propose two expert-annotated datasets, namely XunZi and MengZi, which consist of ancient Chinese texts that have been annotated with entities and their corresponding descriptions by experts, who are professors at the school of history. The manual annotation

of entities and their descriptions is conducted on the traditional Chinese Culture books of XunZi and MengZi, resulting in a total of 5311 XunZi entity-description pairs and 473 MengZi entity-description pairs. The entities are classified into eight types, including person, location, thing, literature, official, institution, time and knowledge. Table 1 presents the statistics of each entity type in the datasets.

Retrieval-Based Pre-trained Corpus. Since we only annotate a small fraction of texts in the original XunZi and MengZi books, there still left many unlabeled ancient Chinese texts. To leverage them, we first use the original XunZi book to pre-train a BERT model [4], called XunZiBERT. Then the pre-trained model is utilized to fine-tune for named entity recognition (NER), with F1 score of more than 93.5%. The NER model is further leveraged to annotate entities for two books, respectively. The pseudo-parallel corpus of ancient Chinese entity descriptions is constructed by retrieving entity information of ancient Chinese entities from more than 910,000 Wikipedia entries. Among them, the conversion of traditional Chinese characters to simplified characters is completed by Opencc. According to the principle of the retrieval enhancement model, dual encoders are used to encode the information of ancient Chinese text entities and Wikipedia respectively. The most relevant item is retrieved. The details are shown in Fig. 2.

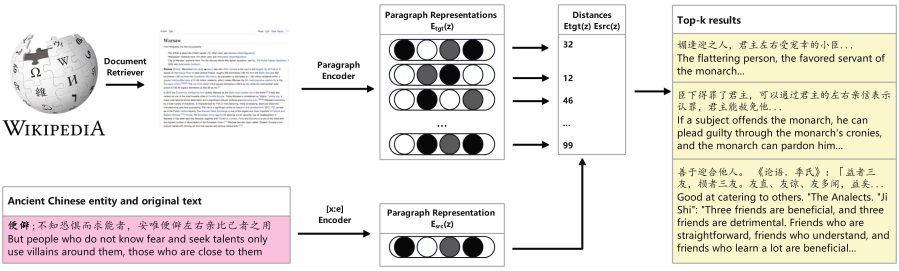

Fig. 2. An illustration of retrieval process. We first retrieve data from Wikipedia, divide it into paragraphs of approximately equal length, and encode each paragraph separately with the target encoder. Each time a pair of entity and entity descriptions is generated, the entity is encoded, and the most suitable paragraph is used as the entity description by dot multiplication.

3 Methodology

3.1 Problem Formulation and Overview

In this paper, we define a practical problem, i.e. ACEDG, and tackle this task by leveraging external knowledge from retrieval. In practice, it is difficult to describe an ancient entity only by its context. The reason is that a good description

requires additional expertise and knowledge background. As the example shown in Fig. 1, retrieval result, i.e. *"Or Stuti. People of the Shang Dynasty. Di made a strong recommendation to the king but was not accepted. He would rather go to the river than betray his country."*, contains rich knowledge for improving the quality of description. Therefore, in addition to the original text context information in the ancient text, we introduce pre-training and retrieval methods. A pseudo-parallel corpus is constructed through retrieval enhancement, which is further used to pre-train the sequence-to-sequence learning model.

Formally, given a sentence $\boldsymbol{x}$ in ancient Chinese and an entity $\boldsymbol{e}$ in this sentence, ACEDG task aims to generate the entity's description $\boldsymbol{y}$ in modern Chinese. Our method enhances ACEDG with retrieved knowledge $\boldsymbol{k}$, which mainly comprises the following three stages:

- **Knowledge Retrieval:** The ancient Chinese sentence $\boldsymbol{x}$ and entity $\boldsymbol{e}$ are leveraged to retrieve modern Chinese knowledge $\boldsymbol{k}$, which is regarded as the pseudo-parallel pairs for the next pre-training stage.
- **Pseudo-Parallel Corpus Pre-Training:** ACEDG relies heavily on human experts' historical knowledge. Though pseudo-parallel pairs are noisy, they still provide rich references for ACEDG. Thus, the retrieved sequences are utilized to pre-train a generation model.
- **Knowledge Enhanced Fine-Tuning:** Finally, the pre-trained model is further fine-tuned on expert-annotated datasets, i.e. XunZi and MengZi in a knowledge-enhanced sequence-to-sequence learning manner.

Table 2. Evaluation results for various retrieval methods on the testing set of XunZi, in terms of BLEU (%), Rouge-1 (%), Rouge-2 (%), Rouge-L (%), and Meteor (%).

Methods	BLEU-1	BLEU-2	Rouge-1	Rouge-2	Rouge-L	Meteor
BM25	8.17	0.39	9.05	0.57	7.63	3.9
Cosine	6.38	0.11	6.14	0.06	7.63	3.9
IDF	6.44	0.07	6.43	0.06	5.26	2.39
Jaccard	**10.22**	0.21	13.05	0.35	9.66	5.15
mContriever ($\boldsymbol{e}$)	4.79	0.45	6.52	0.06	5.32	2.43
mContriever ($[\boldsymbol{x}; \boldsymbol{e}]$)	7.29	**1.63**	**13.26**	**2.18**	**10.27**	**9.33**

3.2 Knowledge Retrieval

To obtain external modern Chinese knowledge, we retrieve the sentences most related to ancient Chinese entities from Wikipedia's Modern Chinese corpus Z. Concretely, the retrieval model mContriever [5] is adopted. Given an input sentence $\boldsymbol{x}$ and one of its entities $\boldsymbol{e}$ in ancient Chinese, the retrieval model first

selects a number of possibly helpful sentences $\{\boldsymbol{k_i}\}_{i=1}^{M}$ from Z, where $M \ll |Z|$, according to a relevance function f.

$$f([\boldsymbol{x};\boldsymbol{e}],\boldsymbol{k}) = E_{src}([\boldsymbol{x};\boldsymbol{e}])^T E_{tgt}(\boldsymbol{k}) \tag{1}$$

where $[;]$ indicates concatenation. E_{src} and E_{tgt} are the source and target sentence encoders that map $[\boldsymbol{x};\boldsymbol{e}]$ and $\boldsymbol{k}$ to d-dimensional vectors respectively.

To explore the performance of multiple retrieval methods, we evaluate the retrieval results on the testing set of XunZi. In other words, the retrieval sentence is compared with the expert-annotated description. The results are shown in Table 2. It can be seen that using $[\boldsymbol{x};\boldsymbol{e}]$ as a query can achieve the overall best performance, which significantly outperforms only using ancient entity $\boldsymbol{e}$. In addition, compared with traditional methods, mContriever ($[\boldsymbol{x};\boldsymbol{e}]$) also presents superiority. The main reason is that mContriever uses dense representations, which can effectively solve the out-of-vocabulary (OOV) problem. Based on these results, we choose mContriever ($[\boldsymbol{x};\boldsymbol{e}]$) for the following stages.

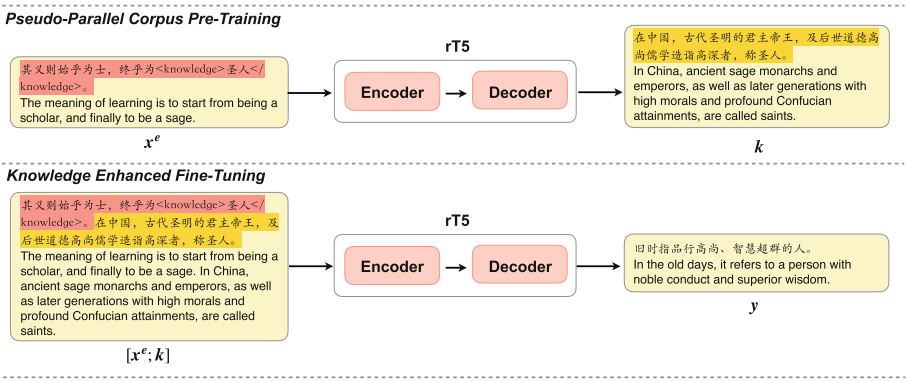

Fig. 3. An illustration of pre-training and fine-tuning processes.

3.3 Pseudo-Parallel Corpus Pre-training

In light of the retrieved knowledge $\boldsymbol{k}$, we then pre-train our model rT5, which has the same encoder-decoder structure as the previous T5 [6] and its related model mT5 [7]. Initially, since ACEDG aims to generate descriptions for entities in the ancient Chinese sentence, the target entity should be highlighted. Therefore, we pre-process the input sequence by inserting special tokens that can emphasize the entity type. As the example shown in Fig. 3, given the sentence "为士,终乎为圣人" (*"The meaning of learning is to start from being a scholar, and finally to be a sage."*), we highlight the entity "圣人" (*"sage"*) with its type and become "< *knowledge* > 圣人 < */knowledge* >". In this way, the original input sentence $\boldsymbol{x}$ is formulated into $\boldsymbol{x^e}$.

Then we pre-train rT5 with pseudo-parallel pairs in a sequence-to-sequence learning manner. Assuming the parameter is θ, the overall pre-training objective is to model the conditional probability $p_\theta(\boldsymbol{k}|\boldsymbol{x}^e)$. Concretely, at the t-th time step, the decoder output $\boldsymbol{k}_t$ is computed with the entity-highlighted input $\boldsymbol{x}$ and the previous outputs $\boldsymbol{k}_{<t}$.

$$p_\theta(\boldsymbol{k}_t|\boldsymbol{x}^e, \boldsymbol{k}_{<t}) = \text{softmax}(W^{\mathrm{T}}\boldsymbol{k}_{<t}) \tag{2}$$

where W maps $\boldsymbol{k}_{<t}$ into a vector, which can represent the probability distribution over the whole vocabulary set.

Then rT5 is pre-trained with minimizing the cross-entropy loss,

$$\mathcal{L}(\boldsymbol{x}^e, \boldsymbol{k}) = -\sum_{t=1}^{n} \log p_\theta(\boldsymbol{k}_t|\boldsymbol{x}^e, \boldsymbol{k}_{<t}) \tag{3}$$

where n is the length of the external knowledge $\boldsymbol{k}$.

3.4 Knowledge Enhanced Fine-Tuning

Finally, we fine-tune the model with the expert-annotated XunZi dataset. To leverage external knowledge in this stage, modern Chinese knowledge $\boldsymbol{k}$ is simply concatenated with formatted ancient Chinese sentence $\boldsymbol{x}^e$ as the input. rT5 is further fine-tuned by minimizing the cross-entropy loss.

$$\mathcal{L}([\boldsymbol{x}^e; \boldsymbol{k}], \boldsymbol{y}) = -\sum_{t=1}^{m} \log p_\theta(\boldsymbol{y}_t|[\boldsymbol{x}^e; \boldsymbol{k}], \boldsymbol{y}_{<t}) \tag{4}$$

where m is the length of the ground-truth target sequence $\boldsymbol{y}$.

4 Experiments

4.1 Experimental Setup

Datasets. We conducted experiments on the expert-annotated datasets proposed in Sect. 2. We divide XunZi dataset into training, valid and testing sets at the ratio of 6:1:3. The statistics of datasets are shown in Table 3.

Table 3. Dataset statistics. Since the scale of MengZi is too small, all data is regarded as the testing set to conduct an out-of-domain evaluation.

Dataset	All	Train	Valid	Test
XunZi	5311	3187	530	1594
MengZi	473	-	-	473

Evaluation Metrics. Both automatic evaluation metrics and human evaluation are used to access the performance of models. The automatic evaluation metrics include BLEU [8], Rouge [9] and Meteor [10]. Human Evaluation includes fluency, semantic consistency, and meaningfulness of the texts.

Implementation Details. To complete our experiment, we propose a retrieval-augmented pre-trained model called rT5. We first construct pseudo-parallel corpora based on full-text data from the book of XunZi. Divide the entire text into sentences, resulting in 8249 sentences. They are adopted to construct the pseudo-parallel corpus.

The model used in the experiment adopts a unified parameter configuration for fairness, the learning rate of the Adam optimizer is $3e^{-4}$, and the number of iterations of each cycle is 100. The maximum sequence length of the input (the entity length will not be very long, but the length of the entity description information is limited) is set to 128.

Compared Methods. To make an extensive evaluation, we choose the following strong baseline methods: 1) generation-based model: **BART** [11] and **GPT** [12]; 2) auto-encoder model, such as **UNILM** [13], which treats generation task as a sentence completion problem. Following UNILM, we also extend **RoBerta** [14] to generate sequences; 3) to extract entities, we pre-train a BERT for NER, called **XunZiBERT**, which can also be leveraged for ACEDG.

4.2 In-Domain Evaluation

Since rT5 is pre-trained on the XunZi book, we first evaluate its performance on the expert-annotated XunZi. The results are shown in Table 4.

Firstly, it can be observed that compared with strong pre-trained language models, rT5 achieves consistent improvements. Among the five models in the first part of Table 4, BART performs the best. Nevertheless, rT5 significantly outperforms BART. Secondly, compared with the naive retrieval results, i.e. mContriever ($\boldsymbol{e}$) and mContriever ($[\boldsymbol{x}; \boldsymbol{e}]$), rT5 also gains significantly. This further illustrates the superiority of our model. Finally, we can see that removing pre-training causes a consistent performance decrease, even reducing to the half scores of rT5. It is worth noting that rT5 outperforms rT5 w/o pre-train by +8.27% on BLEU-1 score, +12.84% on Rouge-1 score, and +3.73% on Meteor score. This also validates that pre-training with pseudo-parallel corpus is crucial.

4.3 Out-of-Domain Evaluation

To further validate the effectiveness of our proposed rT5 model, we test it using the entire MengZi dataset and obtain the results shown in Table 5. It is worth noting that rT5 is pre-trained on XunZi books. All baseline models and rT5 are fine-tuned on expert-annotated XunZi. After that, we directly evaluate rT5 and all baseline methods on the expert-annotated MengZi. The reason for selecting MengZi lies in its distinct philosophical perspective and unique set of entities and explanations, which effectively introduces an out-of-domain challenge.

Table 4. In-domain evaluation results compared with baseline methods. The best scores of each column are marked in bold.

Methods	XunZi					
	BLEU-1	BLEU-2	Rouge-1	Rouge-2	Rouge-L	Meteor
XunZiBERT	3.21	0.16	9.91	0.55	9.35	2.34
GPT	4.73	0.37	10.49	0.61	9.83	3.39
RoBerta	4.97	0.32	11.82	0.74	10.20	5.38
UNILM	5.37	0.35	13.16	0.71	11.36	6.49
BART	7.84	0.88	14.70	0.83	13.79	7.67
mContriever ($\boldsymbol{e}$)	4.79	0.45	6.52	0.06	5.32	2.43
mContriever ($[\boldsymbol{x};\boldsymbol{e}]$)	7.29	1.63	13.26	2.18	10.27	9.33
rT5	**17.32**	**3.78**	**31.78**	**2.59**	**30.68**	**13.24**
w/o pre-train	9.05	1.02	18.94	0.97	18.74	9.51

Table 5. Out-of-domain evaluation results compared with baseline methods. The best scores of each column are marked in bold.

Methods	XunZi→MengZi					
	BLEU	BLEU-2	Rouge-1	Rouge-2	Rouge-l	Meteor
XunZiBERT	4.1	0.13	7.62	0.39	7.26	1.97
GPT	4.7	0.31	7.92	0.42	7.31	2.36
RoBerta	5.8	0.27	8.47	0.51	8.15	3.86
UNILM	6.3	0.28	8.94	0.47	8.75	4.35
BART	6.9	0.74	9.75	0.52	9.68	5.39
rT5	**16.52**	**3.13**	**29.51**	**2.16**	**28.18**	**9.56**
w/o pre-train	8.8	0.95	11.63	0.86	11.63	6.16

From Table 5, we can observe that rT5 achieves the best performance compared with strong baselines. Among the five baseline models, BART performs best. Then compared with it, rT5 still obtains significant improvements. By removing pre-training, the performance also consistently declines. These evaluation results demonstrate that our model achieves good performance in the out-of-domain setting, which signifies its ability to adapt to different domains within the context of ancient Chinese literature. This finding not only validates the robustness and versatility of the rT5 model, but also highlights its potential for understanding and interpreting various ancient Chinese texts.

4.4 Human Evaluation

In assessing the ACEDG task, human evaluation is generally considered more dependable and trustworthy than automated evaluation measures, due to the task's more nuanced, human-like nature. The assessment criteria consist of three

Table 6. Human evaluation results, including average scores and the standard deviation of three evaluators. The best scores of each row are marked in bold.

Model	GPT	BART	UNILM	rT5
Fluency	1.6 ± 0.5	2.5 ± 0.3	1.4 ± 0.4	**3.6 ± 0.3**
Consistency	1.2 ± 0.2	1.6 ± 0.4	1.2 ± 0.2	**2.3 ± 0.4**
Meaningfulness	1.4 ± 0.2	1.5 ± 0.4	1.1 ± 0.1	**3.4 ± 0.4**

distinct dimensions, each scored on a scale of 1 to 5. The fluency dimension appraises the grammatical smoothness of the produced text. The consistency dimension measures the degree to which the translated sentence preserves the original sentence's content. The meaningfulness dimension assesses whether the words in the output sentence convey substantive meanings. The ultimate human evaluation score is computed by taking the average of the scores assigned by all human evaluators. The results are shown in Table 6. It can be observed that among all three dimensions, rT5 generates entity descriptions with the best quality. The human evaluation further verifies the effectiveness of our approach.

ANCIENT CHINESE	<person>晋文公</person>攻原，裹十日粮，遂与大夫期十日。 Duke of Jinwen； Duke of Jinwen attacked the country of Yuan and took ten days' worth of food, so he agreed with the senior officials to withdraw his troops within ten days.
HUMAN EXPERT	春秋时晋国国君。献公子，名重耳。公元前636—前628年在位。因献公宠幸骊姬，立骊姬子奚齐为太子，他被驱逐，出奔在外十九年，经历狄、卫、齐、曹、宋、郑、楚、秦等国。惠公去世，怀公继立，不得人心。秦穆公乘机发兵护送他回国，得即君位。在位时重用随从流亡的狐偃、赵衰等人为卿，整顿内政，增强国力，图谋称霸。 Duke of Jinwen, a monarch of the State of Jin during the Spring and Autumn period. His name was Chong'er, a son of Duke of Jinxian. He was in power from 636 BC to 628 BC. Because of his father's favoritism towards Li Ji, and his decision to make her son, Xiqi, the crown prince, Chong'er was exiled and wandered in other states for nineteen years. After Duke of Jinhui's death and Duke of Jinhuai's unpopular reign, Duke of Qinmu saw an opportunity to send troops to escort him back to his state, where he ascended the throne. During his rule, he heavily relied on his followers who had accompanied him during his exile, such as Hu Yan and Zhao Shuai, for important positions. He reformed domestic affairs, enhanced national power, and aspired to hegemony.
UNILM	之谓天下之事也，是之谓天下之事也。 This is called the affairs of the world, this is called the affairs of the world.
BART	晋文公之舅，名狐偃，字犯。 Duke of Jinwen's uncle, named Hu Yan, style name is Fan.
GPT	下莫之敢当，和齐之兵。 None dare to be the first, and Qi's troops.
rT5	姬姓，晋氏，名重耳，晋献公之子，春秋时期著名的政治家，晋国国君，在位九年，整顿内政，图谋称霸。 From the Ji clan, of the Jin family, named Chong'er, he was the son of Duke of Jinxian. He was a famous politician during the Spring and Autumn period and a monarch of the State of Jin. He was in power for nine years, reformed domestic affairs, and aspired to hegemony.

Fig. 4. An illustration of the case study.

4.5 Case Study

To further evaluate the proposed model, a case study is conducted. The details are depicted in Fig. 4. These models are tested by generating an explanation for the historical entity "晋文公" *(Duke of Jinwen)*. For benchmarking, we use a human expert's explanation as a reference. Upon comparison, it can be observed that the rT5 model's output is the most accurate and comprehensive, closely matching the reference description provided by the human expert. The UNILM, BART, and GPT models' outputs are not entirely correct, lacking depth

and completeness compared with rT5. The superior performance of our rT5 model over existing sequence-to-sequence models demonstrates its potential in facilitating the understanding and analysis of ancient Chinese texts.

5 Related Works

NLP Applications for Ancient Chinese. Previous studies have gained great success by applying NLP techniques to ancient Chinese. To name a few, Chang *et al.* leverage local features for named entity recognition [15]. Li *et al.* [2] propose an approach that combines conditional variational autoencoder (CVAE) and adversarial training for Chinese poem generation. To better produce smooth poetry that fits the topic, Yang *et al.* [16] explore unsupervised machine translation (UMT) to generate classical Chinese poems from the vernacular, which allows the controlling over the semantics of generated poems. Other interesting works include couplet generation [3,17], a part of traditional Chinese culture and formatted as two sentences with symmetrical meanings. In this work, we focus on the entity description of ancient entities, which is meaningful for historical and cultural diffusion.

Retrieval-Based Generation. To utilize external knowledge, many retrieval-based generation works have been proposed. Guu *et al.* [18] propose to augment the pre-training of language model based on retrieval. Wang *et al.* [19] propose a new framework using retrieval methods to enhance the pre-training and fine-tuning of general knowledge generation. The prototype candidate sentence is retrieved by concept matching and used as an auxiliary input. In this paper, we use a general-purpose dense retriever based on the dual-encoder architecture of mContriever [5] to retrieve Wikipedia content according to the current context content, entity type, entity information. This aims to generate a large number of suitable pseudo-parallel corpus to enhance pre-training.

6 Conclusion

In this paper, we define a new task (i.e. ACEDG) and propose two expert-annotated datasets (i.e. XunZi and MengZi), for promoting Chinese culture and history. To tackle this task, a retrieval-augmented pre-trained model, i.e. rT5, is proposed. Specifically, we first adopt the retrieval technique for building pseudo pair of ancient Chinese entity and modern Chinese description. Then, the pseudo-parallel corpus is leveraged to pre-train our model, which incorporates external knowledge. Finally, rT5 is fine-tuned on the expert-annotated dataset with the help of external knowledge. Experimental results under various metrics show that our method can generate higher-quality entity descriptions. The rT5 model's success in this study highlights the potential of leveraging information retrieval-enhanced techniques for dataset construction and domain-specific performance. Future research can explore the application of these techniques in other domains and languages to further enhance the understanding of ancient Chinese.

References

1. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. IEEE Trans. Knowl. Data Eng. **34**(1), 50–70 (2020)
2. Li, J.: Generating classical Chinese poems via conditional variational autoencoder and adversarial training. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3890–3900. Association for Computational Linguistics, Brussels, Belgium, October–November 2018
3. Wang, Y., Zhang, J., Zhang, B., Jin, Q.: Research and implementation of Chinese couplet generation system with attention based transformer mechanism. IEEE Trans. Comput. Soc. Syst. (2021)
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Izacard, G., et al.: Unsupervised dense information retrieval with contrastive learning (2021)
6. Raffel, C.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
7. Xue, L.: mT5: a massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2020)
8. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, July 2002
9. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain, July 2004
10. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan, June 2005
11. Lewis, M.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
12. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
13. Dong, L.: Unified language model pre-training for natural language understanding and generation. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
14. Liu, Y.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
15. Chang, Y., Kong, L., Jia, K., Meng, Q.: Chinese named entity recognition method based on BERT. In: 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA), pp. 294–299 (2021)
16. Yang, Z., et al.: Generating classical Chinese poems from vernacular Chinese. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. In: Conference on Empirical Methods in Natural Language Processing, vol. 2019, p. 6155. NIH Public Access (2019)

17. Yuan, S., Zhong, L., Li, L., Zhang, R.: Automatic generation of Chinese couplets with attention based encoder-decoder model. In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 65–70. IEEE (2019)
18. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: International Conference on Machine Learning, pp. 3929–3938. PMLR (2020)
19. Wang, H.: Retrieval enhanced model for commonsense generation. arXiv preprint arXiv:2105.11174 (2021)