Time-Aware Ancient Chinese Text Translation and Inference

⋆**Ernie Chang**[文], ○**Yow-Ting Shiue**[文], ⋆**Hui-Syuan Yeh**, ⋆**Vera Demberg**
⋆Dept. of Language Science and Technology, Saarland University
○Dept. of Computer Science, University of Maryland, College Park
{cychang}@coli.uni-saarland.de, ytshiue@cs.umd.edu

Abstract

In this paper, we aim to address the challenges surrounding the translation of ancient Chinese text: (1) The *linguistic gap* due to the difference in eras results in translations that are poor in quality, and (2) most translations are missing the contextual information that is often very crucial to understanding the text. To this end, we improve upon past translation techniques by proposing the following: We reframe the task as a *multi-label prediction task* where the model predicts both the translation and its particular era. We observe that this helps to bridge the linguistic gap as chronological context is also used as auxiliary information. We validate our framework on a parallel corpus annotated with chronology information and show experimentally its efficacy in producing quality translation outputs. We release both the code and the data[1] for future research.

1 Introduction

The Chinese language inherits a lot of phrases from ancient time (Bao-chuan, 2008; Liu, 2019) and is spoken by roughly 1.3 billion native speakers. However, the language's ancient variant (or *ancient Chinese*) is mastered by a few and proved to be a bottleneck in understanding the essence of the Chinese culture. Building a translation system from the ancient Chinese to the modern text thus serves a few important purposes: (I) The ancient Chinese is considered as an essential part of the curriculum in all of the Chinese-speaking regions[2], so an ancient Chinese translation system can be used to bolster the immediate understanding of ancient texts. (II) Further, the translation system can help to settle the linguistic debate with regard to the era of origin of an independent segment of text. This is especially useful for the identification of a discovered artifacts where carbon dating cannot pinpoint the exact era, but where their linguistic features can formulate a clear-cut dynasty or time period.

However, it is not without challenge in constructing such translation systems. One primary obstacle lies in the extensive timeline where ancient texts can be derived – one segment of ancient text can come from the Pre-Qin (先秦) era, and another coming from the Song dynasty (宋朝), which are roughly about 700 years apart. This gap witnessed a drastic evolution of linguistic properties where the usage of phrases became imbued with different meanings. Besides, different eras often consist of various amounts of available data, and thus the same translation model training will be exposed to data imbalance, which complicates the design of the translation systems and limits their generalizability. On the other hand, past attempts at building such translation systems yield poor performance that renders them practically unusable as-is in the practical settings (Zhang et al., 2018; Liu et al., 2019) – these efforts are still largely limited as parallel data is scarce for some eras.

Recent advances in machine translation and text style transfer/generation utilize semi-supervised techniques to tackle similar challenges by aligning latent representations from different styles for the low resource scenarios (Shen et al., 2017; Hu et al., 2017; Rao and Tetreault, 2018; Prabhumoye et al., 2018; Jin et al., 2019; Chang et al., 2020, 2021c). To this end, we aim to bridge this gap that makes the following contributions:

- We showed that having ancient Chinese text of all eras in a single corpus is not ideal as they are difficult to model jointly as a single distribution, and that the additional *chrono-*

[文]These authors contributed equally.

[1]https://github.com/orina1123/time-aware-ancient-text-translation

[2]This includes the mainland China, Taiwan, Singapore, Malaysia, etc.

Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021, pages 1–6

logical context helps to improve translation of ancient Chinese to modern Chinese sentences.

- For future research in this direction, we release our code and parallel data consisting of annotated *chronological identifiers* which allow to infer the approximate era of the written text in the practical settings.

2 Background

At a fine-grained view, the notion of "ancient Chinese" may not be considered a single language with a static word-meaning mapping. Therefore, we direct our efforts toward three particular eras: Pre-Qin (先秦), Han (汉), and Song (宋) to verify the hypothesis that the chronology of a text directly influences the word meaning and model performance. In particular, *Pre-Qin* and *Han* are closer chronologically, so we expect their model performances to be closer than that between *Pre-Qin* and *Song*, as was shown in other ancient text translation (Park et al., 2020).

One reason for this difference is the use of polysemous single-character words, which are highly ambiguous. Some words begin to lose meanings over time. For example, in ancient Chinese, the word 看('kàn') has many meanings such as "to visit" and "to listen", in addition to the major modern meaning, "to look".

As the language evolved, vocabulary changed and lexical semantic shift took place, creating diachronic semantic gaps that may introduce subtle differences in the understanding of the text. For instance, the earliest known meaning of "看" is "to look into the distance". The meaning of "to look at something closely" emerged during the Han period and eventually became the prominent meaning of this verb in modern Chinese. In sum, the language change across time suggests a modeling approach that is aware of when the text was written.

3 Task Formulation

We assume two nonparallel datasets A and M of sentences in *Ancient Chinese (zh-a)* and *Modern Chinese (zh-m)* respectively.A parallel dataset P that contains the pairs of sentences in both variants of text is also present. The sizes of the three datasets are denoted as $|A|$, $|M|$ and $|P|$, respectively. As the nonparallel data is abundant but the parallel data is limited, size $|A|, |M| \gg |P|$. The main objective is to convert the input ancient Chinese text a to its modern variant m. This task is akin to style transfer, or if the text are drastically different, machine translation. In this paper, we are only concerned with the direction from zh-a to zh-m. Additionally, we include the prediction of the chronological period of the ancient text as an auxiliary task.

4 Proposed Framework

Our framework translates the given ancient Chinese text (§4.1) while providing additional chronological context information (§4.2) (see Table 1). We train the *translation model* in a semi-supervised manner such that cheap and easy-to-obtain modern Chinese text can be utilized in the training process. To better select from the pool of generated candidates in a time-aware way, we use the multi-label prediction model as both the *reranker* and the *chronology predictor*.The predicted chronological period also provides users with crucial context for understanding the ancient text.

4.1 Semi-Supervised Translation Model

Our sequence-to-sequence model is based on the Transformer (Vaswani et al., 2017) encoder-decoder architecture. Given an input, the encoder first converts it into an intermediate vector, and then the decoder takes the intermediate representation as input to generate a target output. In what follows, we describe the training objectives that allows the translation model to utilize augmented monolingual data.

Semi-Supervised Objectives. Inspired by the previous work on CycleGANs (Zhu et al., 2017) and dual learning (He et al., 2016; Chang et al., 2021a,b), our method trains the initial model in both forward and backward directions, and defines a semi-supervised optimization objective that combines direct supervision ($L_{supervised}$) and a language model loss (L_{lm}) over the parallel data P, and two monolingual corpora A and M:

$$L = L_{supervised}(P) + L_{lm}(A) + L_{lm}(M)$$

where $L_{supervised}(P)$ utilizes the aligned sentence pairs in P to perform domain alignment, ensuring that the representation of the ancient Chinese text can be semantically aligned with its modern variant. Moreover, *the semi-supervised training allows us to augment monolingual modern Chinese* for language modeling. Empirically, we found that this

	Text	Chronological Period
Source (Ancient Chinese)	孟子曰：道在尔而求诸远，事在易而求之难。	
Reference	孟子说：道路在近旁而偏要向远处去寻求，事情本来很容易而偏要向难处下手。 (Menzie said: "The right path is just beside but people take far away ones instead; things are easy but people handle them with difficult ways.")	`pre-qin`
System (Modern Chinese)	孟子说：道理在于尔而求得远方，事情在于易而求得难 。	`pre-qin`
Source (Ancient Chinese)	秦昭王召见，与语，大说之，拜为客卿。	
Reference	秦昭王便召见了蔡泽，跟他谈话后，很喜欢他，授给他客卿职位。 (The King of Qin summoned Mr. Ze Cai and, after talking to him, liked him and gave him a government official position for foreigners.)	`han`
System (Modern Chinese)	秦昭王召见他，与他谈话，非常高兴，拜他为客卿。	`han`
Source (Ancient Chinese)	太子曰：吾君老矣，非骊姬，寝不安，食不甘。	
Reference	太子说：我父亲年老了，没有骊姬将睡不稳、食无味。 (The Prince said: "My father is old. Without this girl, Li, he cannot sleep well or eat well.")	`han`
System (Modern Chinese)	太子说：我国君已经老了，不是骊姬的姬妾，吃不甘。	`han`

Table 1: Examples of system output consisting of the *ancient Chinese source*, *modern Chinese reference* and the *chronological period prediction*.

benefits the forward translation from zh-a to zh-m and proves to be a viable way for improving the system.

4.2 Multi-Label Prediction

Further, we improve upon the translation model via the use of the *chronology inference* and *translation reranking* via the dual-purpose *multi-label prediction model*. Specifically, we pretrain a modern Chinese language model then fine-tune this model in a task-specific manner to help predicting the chronological period and using it to also rank the translation model's predictions.

Chronology Inference. To do so, we first pretrain a large-scale language model on the monolingual modern Chinese corpus following objectives in Radford et al. (2019) for GPT-2. This enables the model to be familiarized with the language semantics where some of which are transferrable to the ancient text. Next, we continue to train the GPT-2 model to perform conditional task-specific generation by maximizing the joint probability $p_{\text{GPT-2}}(a, m, c)$, where a is the ancient Chinese text, m is the modern Chinese text, and c represents the contextual information as the chronological period of the ancient text. Specifically, for each sentence pair, the ancient Chinese tokens w_i^a, the modern Chinese tokens w_j^m, and the chronological period are concatenated into "[zh_a] $w_1^a \cdots w_{|a|}^a$ [zh_m] $w_1^m \cdots w_{|m|}^m$ [chron] c", and the model is trained to maximize the probability of this sequence.

Quality Estimation for Reranking. At inference time, we append each of the chronology labels to the translation outputs, then allow the multi-label prediction model to predict their qualities. Specifically, the fine-tuned LM computes the negative log loss on each of the triplets (a, m', c') from the upstream *translation model* by appending *exhaustively* all possible *chronology labels* c' to the end of the generated sequence m' following the same format as above and selecting the best.

5 Dataset Construction

We obtain parallel ancient-modern Chinese sentence pairs, and nonparallel ancient (zh-a) and modern Chinese (zh-m) sentences from two sources (Liu et al., 2019; Shang et al., 2019). Table 2 summarizes the data we used for the experiments.

Chronology Annotation. In this paper, we focus on translating ancient prose. There are a total of 28,807 ancient Chinese prose sentences. We annotate each of these sentences with the Chinese historical period (dynasty) in which it was written. Specifically, we consider three chronology labels: `pre-qin` (先秦), `han` (汉), and `song` (宋). The annotation is based on the source of the sentences, i.e., which ancient book the sentences are taken from. The total number of annotated sentences for each period is 1,244, 20,460, and 7,103 respectively. This annotation scheme can be adopted for a larger set of periods when ancient text of a wider time span is available.

Parallel Data. For the sentences with chronology annotation, we randomly assign 10% sentences to the development set and test set respectively. The

		# sentences	# characters
Nonparallel	zh-a	269,409	4M
	zh-m	77,687	826K
Parallel	Train	27,807	(524K, 797K)
	Dev	2,880	(59K, 88K)
	Test	2,880	(60K, 90K)

Table 2: Statistics of the dataset. For each part of the dataset, the number of sentences and the (source, target) number of characters are shown.

	BLEU			
Training Objectives	All	pre-qin	han	song
$L_{supervised}$ (Liu et al., 2019)	19.59	14.41	**20.02**	19.13
$L_{supervised} + L_{lm}(M)$	23.05	15.97	**23.32**	23.17
$L_{supervised} + L_{lm}(M) + L_{lm}(A)$	23.15	14.15	23.34	**23.72**
+ share decoder embeddings	24.38	15.70	24.52	**24.99**
+ time-aware reranking	**24.51**	15.50	24.62	**25.24**

Table 3: Ancient to modern Chinese translation performance. BLEU scores are calculated with 1 to 4 character n-grams.

remaining sentences are used as training data. We further supplement the parallel training data with 4,760 sentences from ancient Chinese poems, each also with a modern Chinese translation. The final training, development and test set statistics and be found in Table 2.

Nonparallel Data. We extend the source-side data by including 269,409 more ancient poem sentences without translation. For extending target-side data, we add 77,687 sentences from modern lyrics, following Shang et al. (2019). The details of nonparallel data are also shown in Table 2.

6 Experimental Settings

We tokenized both ancient and modern Chinese text by splitting characters. The vocabulary sizes are 4,824 and 4,600 respectively. We built our model upon the Fairseq toolkit[3]. The architecture is Transformer with about 54M parameters, which largely follows the configuration of Liu et al. (2019). Translations were generated with beam size 5, and we consider top 5 candidates for reranking. For the *multi-label prediction model*, we adapted existing code[4] to build a GPT-2 Language Model reranker with approximately 82M parameters. First, we pre-trained the model with 1.2 GB of Chinese Wikipedia text. Then, we fine-tuned the pre-trained model with the chronologically-annotated training data. For each ancient-modern sentence pair with chronology information, we form a text-period query string with the scheme described in §4.2. We select the final model according to perplexity computed on the development set.

[3] https://github.com/pytorch/fairseq

[4] https://github.com/Morizeyao/GPT2-Chinese

Period (# test)	Precision	Recall	F1
pre-qin (117)	0.05	0.53	0.09
han (2043)	0.85	0.57	0.68
song (720)	0.85	0.27	0.41
Accuracy			0.49
Macro avg.	0.58	0.45	0.39
Weighted avg.	0.82	0.49	0.59

Table 4: Performance of Chronology Inference

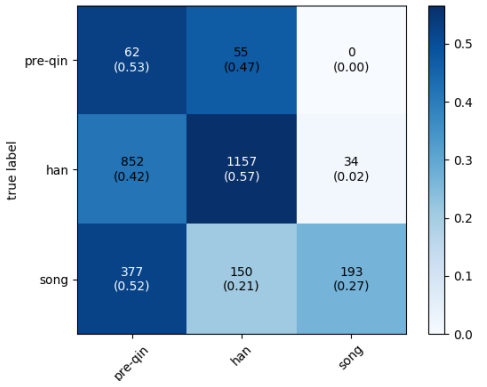

Figure 1: Confusion matrix for Chronology Inference

7 Main Results

Overall, we observe from Table 3 and 4 that the use of the multi-label prediction model not only allows for better context than pure translation, but also helps to boost the general performance on the translation tasks. Moreover, *translations of ancient text chronologically closer to modern Chinese (han and song) tend to yield better performances, as the semantic gaps are generally smaller*. We also demonstrate that the semi-supervised training which avail of the additional nonparallel text helps to improve the translation model even further. Specifically, zh-m nonparallel data enhances the decoder's ability to generate modern Chinese, while zh-a nonparallel data may help the encoder to maintain crucial semantic information. We achieved a BLEU score of 23.15 in this setting. As the source and target side vocabularies have a large overlap, we experimented with sharing decoder embeddings and got +1.23 BLEU improvement, which may also serve as an evidence that there are still ancient components in modern Chinese. Finally, reranking further boosted the BLEU score to 24.51.

Error Analysis. We perform *human evaluation* on 100 randomly sampled output instances and ob-

serve them to be high in *adequacy* and *fluency*, 4.06 and 3.68 respectively, on a scale of 0-5. This was done by averaging the fluency and adequacy ratings of three domain experts. Further, we also observe that the chronology of text impacts the model performance as in Table 3. Leveraging zh-m nonparallel data is most helpful for translating text from the `song` period, which is much closer to modern Chinese compared to the text from the other two periods. Further, from Figure 1 we observe that the chronology inference depends very much on the *data scarcity* and the *closeness* of chronological periods. On the Chinese historical timeline, `han` is very close to `pre-qin`, but `han` and `song` are more separated. Another source of difficulty is that ancient Chinese writings tend to quote a considerable amount of text written in previous time periods. For example, a history book written in the `song` period may inherit narratives written in `pre-qin` and `han` for the history before `han`. As a result, it is challenging to perform chronology inference based solely on the linguistic properties of individual sentences. Nevertheless, chronological inference can still provide useful signals for the translation model to better capture semantic differences across time.

8 Conclusion

In this paper, we present a framework that translates ancient Chinese texts into its modern correspondence in low resource scenarios with very little parallel data and a larger set of nonparallel sentences without ancient-modern alignment information. We display the importance and usefulness of chronology inference as an auxiliary task that hints at potential diachronic semantic gaps. We hope to extend this research to further model additional contextual information about each era.

Acknowledgements

This research was funded in part by the German Research Foundation (DFG) as part of SFB 248 "Foundations of Perspicuous Software Systems". We sincerely thank the anonymous reviewers for their insightful comments that helped us to improve this paper.

References

LI Bao-chuan. 2008. Illustrations of antique meanings of the chinese phrases. *Journal of Radio & TV University (Philosophy & Social Sciences)*, 3.

Ernie Chang, Jeriah Caplinger, Alex Marin, Xiaoyu Shen, and Vera Demberg. 2020. Dart: A lightweight quality-suggestive data-to-text annotation tool. *arXiv preprint arXiv:2010.04141*.

Ernie Chang, Vera Demberg, and Alex Marin. 2021a. Jointly improving language understanding and generation with quality-weighted weak supervision of automatic labeling. *Proceedings of EACL 2021*.

Ernie Chang, Xiaoyu Shen, Dawei Zhu, Vera Demberg, and Hui. Su. 2021b. Neural data-to-text generation with lm-based text augmentation. *Proceedings of EACL 2021*.

Ernie Chang, Hui-Syuan Yeh, and Vera Demberg. 2021c. Does the order of training samples matter? improving neural data-to-text generation with curriculum learning. *Proceedings of EACL 2021*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. Unsupervised text style transfer via iterative matching and translation. *arXiv preprint arXiv:1901.11333*.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2019. Ancient–modern chinese translation with a new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–13.

Yu Liu. 2019. 现代汉语常用文言虚词的语块教学 [formulaic language instruction of classical grammatical words in modern chinese]. *Chinese as a Second Language. The journal of the Chinese Language Teachers Association, USA*, 54(2):122–144.

Chanjun Park, Chanhee Lee, Yeongwook Yang, and Heuiseok Lim. 2020. Ancient korean neural machine translation. *IEEE Access*, 8:116617–116625.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140. Association for Computational Linguistics.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4937–4946, Hong Kong, China. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zhiyuan Zhang, Wei Li, and Xu Sun. 2018. Automatic transferring between ancient chinese and contemporary chinese. *arXiv preprint arXiv:1803.01557*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.