CCPM: A Chinese Classical Poetry Matching Dataset

Wenhao Li[1,2], **Fanchao Qi**[1,2], **Maosong Sun**[1,2,3]*, **Xiaoyuan Yi**[1,2], **Jiarui Zhang**[2,4]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Beijing National Research Center for Information Science and Technology
[3]Institute for Artificial Intelligence, Tsinghua University, Beijing, China
[4]Department of Electronic Engineering, Tsinghua University, Beijing, China
{wh-li20, qfc17, yi-xy16, zhangjr18}@mails.tsinghua.edu.cn,
sms@tsinghua.edu.cn

Abstract

Poetry is one of the most important art forms of human language. Recently many studies have focused on incorporating some linguistic features of poetry, such as style or sentiment, into its understanding or generation system. However, there is no one focus on understanding or evaluating the semantic of poetry. Therefore, we propose a novel task to assess the level of the model's semantic understanding in poetry, poem matching. This task requires the model to select one poem among four according to the corresponding translation of the correct poem in modern Chinese. To construct this dataset, we first obtain a set of bilingual parallel data of classical Chinese Poetry and modern Chinese. Then we retrieve similar poems with the poems in the bilingual pairs in the poetry corpus as confusion choices. We named our dataset Chinese Classical Poetry Matching Dataset (CCPM) and released it on https://github.com/THUNLP-AIPoet/CCPM. We hope this dataset can further enhance the study of incorporating deep semantics into the automatic understanding and generation system of classical Chinese poetry.We also preliminarily run two variants of BERT(Devlin et al., 2019) on this dataset as the baseline for this dataset.

1 Introduction

Language is one of the most crucial forms of human intelligence. Among all the genres of human language, poetry is a distinctive artistic genre with exquisite expression, rich content, and diverse styles. In the long history of humankind, poetry shows profound impacts across different countries, nationalities, and cultures.

Poetry has various distinguishing characteristics from other genres, including powerful emotion, explicit language style, and rich content expressed in an abstractive manner. These characteristics differentiate the automatic processing of poetry from the processing of other genres by a large margin. As a result, there have been many works focusing on some of their features of poetry such as style (Yang et al., 2018; Yutong et al., 2020) and sentiment (Chen et al., 2019). However, to our best knowledge, there was no work concentrating on the internal semantics of poems. There may be a possible reason. In poem writing, the poet often needs to compress plentiful meanings to the limited length of contents constrained by the genre. Therefore, the semantics presented in the poem is much fuzzier and more entangled among different segments than other genres. That leads to the difficulty of automatically analyze and evaluate poem semantics, which encourages more work in this area. We also run BERT (Devlin et al., 2019) on our dataset and got the highest accuracy of 84.96%.

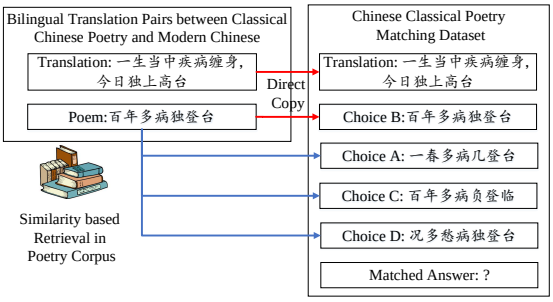

Figure 1: The major process of the dataset construction. We first collect bilingual translation pairs and retrieve the most similar candidates from the poetry corpus as confusing choices

Therefore, in this work, we propose a benchmark on the semantic of Chinese Classical Poems. More specifically, we design a novel task to quantify the semantic modeling ability around different models-poetry matching. This task is to test whether the model can discern the correct poem line with other

*Corresponding Author

similar lines given the corresponding translation of the correct line in the modern Chinese language.

Meanwhile, we also established the dataset due to this task. We first collected 31K bilingual parallel data between Chinese Classical Poems and the modern Chinese language. Then we cleaned the data and retrieved the most similar poem lines in our poetry corpus for each poem. The major process of constructing this dataset is showed in 1

We hope this dataset can further enhance the research on semantic in poetry. It can benefit the semantic understanding of the poetry analysis models. It also bridges the semantic of daily Chinese with uncommonly used poetic language, providing a chance for poetry generation models to better understands the users' intend and to improve the semantic relevance between the user input and the generated poem.

Therefore, the contributions of this work lie in:

- Proposed a new task to match the translation of Chinese Classical Poems on the modern Chinese language to its original lines;
- Released a dataset on this task to futhur evluate and improve the semantic understanding of both autmatic analysis and automatic generation model of Chinese Classical Poems.

2 Related Works

There have been several datasets in Classical Chinese Poetry released to the public. The first fine-grained dataset, to our best knowledge, is released by Chen et al. (2019). They annotated a manually-labeled Fine-grained Sentiment Poetry Corpus including 5,000 Chinese quatrains and released it on GitHub. They also released a Chinese poetry dataset CCPC, which contains 151,835 unlabelled Chinese quatrains. Moreover, Yutong et al. (2020) released a dataset with 3940 quatrains with themes and 1917 with sentiments automatically annotated by template-based methods. They also constructed a knowledge graph on Classical Chinese Poetry using the Apriori algorithm and human-defined scheme system. To our best knowledge, the most similar work to ours is (Liu et al., 2019). They also collected a dataset of the parallel bilingual pairs between ancient Chinese and modern Chinese from the web. They also used string matching algorithms to align the lines in parallel pairs. However, they focus on using this data to train a neural machine translation model to tackle the poetry translation poem. Instead, we aim to construct a matching dataset based on those bilingual pairs to assess the semantic understanding ability of the models.

3 Dataset Construction

3.1 Bilingual Pair Extraction

The extraction of bilingual data mainly consists of three subprocesses: raw data acquiring, line segmenting, and format filtering.

Raw data acquiring We collected 6K paragraphs in ancient Chinese and their corresponding translation in modern Chinese language from the web to construct this dataset. They include classical Chinese poetry and other literature in the ancient Chinese language.

Line segmenting As stated in the introduction, we want to build our dataset on the level of poem lines. Therefore, we need to do this segmentation. We split some instances into lines by the line separation in the raw web contents and punctuations. For some other instances, we can only split them into the units of two lines. Considering that this unit can also have consecutive semantics, we also keep this part of examples in our dataset in the form of two-line instances.

Format filtering We only want to keep classical Chinese poems in our dataset, so we need to filter the others. Since the 5-character line (5 yan) poem and the 7-character line is most commonly seen in classical Chinese poetry by a large margin, we only keep lines in these two formats in our dataset. We achieve this by specifying a length constraint on the two consecutive lines.

After these steps, we extracted 27,218 bilingual pairs of classical Chinese poetry and their translation into modern Chinese for our dataset. For more detailed statistics, the reader can refer to Chapter 2.3.

3.2 Candidates Retrieval

After obtaining the bilingual pairs, the next step is to extract the negative choices according to the correct answer. To select more confusing opposing candidates, we need to seek more similar poem lines with the ground truth answer. Therefore, we used our pre-trained model on poetry, BERT-CCPoem (Guo et al., 2020), to calculate the similarity between sentences. This model is trained on an

	Training	Validation	Test	Total
5-yan 1-line	10,166	1,270	1,270	12,706
7-yan 1-line	9,392	1,173	1,173	11,738
5-yan 2-lines	1,025	128	128	1,281
7-yan 2-lines	1,195	149	149	1,493
Total	21,778	2,720	2,720	27,218

Table 1: Detailed Dataset Statistics: *5-yan* and *7-yan* refer to 5 or 7 Chinese characters in one poem line, and *1-line* and *2-lines* refer to 1 or 2 lines in one sample.

(almost) full collection of Chinese classical poems, CCPC-Full v1.0, consisting of 926,024 classical poems with 8,933,162 sentences.

We use the hidden state of the `[CLS]` token as the semantic representation of the whole sentence and use the cosine similarity of two lines' representation as the similarity metric between them. We also adopt the Locality Sensitive Hashing (LSH) algorithm(Arya et al., 1998) to speed up retrieving the most similar sentences. The top candidates retrieved are re-ranked by the weighted average of the BERT similarity and the Longest Common Sequence (LCS) between each candidate and the ground truth poem line like in Equation 1. More details in this re-ranking algorithm are available in (Guo, 2020).

$$\begin{aligned}\mathrm{Sim}(A,B) = 0.4 * \mathrm{LCS}(A,B) \\ + 0.6 * \mathrm{EMB}(A,B)\end{aligned} \quad (1)$$

After choosing the most similar candidates, we select the most similar candidate and randomly sample two choices separately from the top 2-5 and the top 6-10 similar candidates. We use these three as the negative choice and shuffle them with our ground truth answer.

3.3 Dataset Statistics

We obtain 27,218 translation-candidates pairs and split them into training, validation, and test set. The detailed statistics are in Table 1.

4 Experiments

In this section, we evaluate some popular NLP models on the CCPM dataset.

Evaluation Metric We use accuracy, namely the percentage of the test samples for which the evaluated model predicts the correct answer, as the evaluation metric. The higher the accuracy is, the model performs better.

Model	Accuracy
BERT-Cls	84.96
BERT-Match	82.60

Table 2: The accuracy results of the evaluated model on the test set of CCPM.

Evaluated Models We choose the popular pre-trained language model BERT (Devlin et al., 2019) as the sentence encoder and design the following two models:

(1) BERT-Cls. Similar to previous work (Devlin et al., 2019), it concatenates the given translation with each candidate poem line (with an additional separator token) and feeds the concatenation into BERT. Then it feeds the hidden state of the `[CLS]` token into a linear layer and makes a binary classification: match or not.

(2) BERT-Match, which regards the task as a sentence match problem. Specifically, it uses two BERTs to encode the translation and candidate poem lines, respectively, obtaining the embeddings of translation and poem lines. Then it calculates the cosine similarity between the embeddings of translation and each poem line. The candidate poem that is most similar to the translation is selected as the final answer.

Implementation Details For both models, we choose `bert-base-chinese` from the Transformer library (Wolf et al., 2020) as the sentence encoder. During fine-tuning, we use the Adam optimizer (Kingma and Ba, 2015), with an initial learning rate $2e-5$ that decreases linearly and train the models for 4 epochs.

Results The accuracy results of the two models are shown in Table 2. We can see that the two models perform similarly and both achieve an acceptable accuracy. BERT-Cls outperforms BERT-Match, presumably because BERT-Cls achieves more interaction between the translation and a poem line with the self-attention mechanism of Transformer (Vaswani et al., 2017) of BERT.

5 Conclusion

This paper proposed a novel task-poem matching to assess the semantic understanding ability in classical Chinese poetry. Further, we constructed the dataset of this task by collecting the parallel data of the classical poems and their translation and

retrieving similar poems in the poetry corpus as confusion choices. We also ran two variants of the BERT model and compared the result.

In the future, we will further refine this dataset in three ways. Firstly, we will collect more parallel data to increase the volume of this dataset. Moreover, we will explore more confusing ways to construct negative choices. Thirdly, we will test more commonly used NLP models on our benchmark.

References

Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu. 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4925–4931. AAAI Press.

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

Zhipeng Guo. 2020. A study on the ranking and recommendation of chinese classical poetry. Master's thesis, Tsinghua University.

Zhipeng Guo, Jinyi Hu, and Maosong Sun. 2020. Bert-ccpoem. https://github.com/THUNLP-AIPoet/BERT-CCPoem.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Proceedings of NeurIPS*.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP*.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3960–3969.

Liu Yutong, Wu Bin, and Bai Ting. 2020. The construction and analysis of classical chinese poetry knowledge graph. *Journal of Computer Research and Development*, 57(6):1252.