CCPM: A Chinese Classical Poetry Match 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

tbd

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 and sentiment. 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.

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 by testing whether the model can discern the correct poem line with other 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 31M 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.

We believe 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 Dataset Construction

2.1 Bilingual Pair Extraction

2.2 Candidates Retrieval

2.3 Dataset Statistics

3 Experiments

References

*Corresponding Author