Chinese Couplet Generation Proposal

Kuan-Yu Chiang
kuanyu@usc.edu

Qian Yin
qianyin@usc.edu

Joe Chen
zchen462@usc.edu

Shihao Lin
shihaol@usc.edu

Qizhen Jin
qizhenji@usc.edu

1 Project Domain and Goals

Poetry is a form of literature that is arranged for its meaning and rhythm, and it has been proven that it plays a significant role in language development and the advancement of history. Among the diversity of literature, Chinese couplet stands out with a special poetry form using pairs of orderly sentences. With the nurture of over 1000 years, the Chinese couplet has become a unique art and historical treasure, however, due to the complexity of semantic and grammatical rules, the creation of a suitable couplet is a formidable challenge.

The complexity of the semantic and grammatical rules of Chinese couplets tends to address greater difficulty to the data preprocessing and model design. More attention is needed in processing the Chinese language data. In this paper, we will formulate the language generation problem by utilizing the Nature Language Processing (NLP) techniques and neural networks. Natural Language Processing will be helpful in preprocessing the data, and the neural network structure model is effective in solving the text generation problem. With this application, we will be able to successfully generate the Chinese couplets, and we aim to extend the model to apply and generate other types of literature.

Our objective in this project is to improve the previous NLP work on the Chinese Couplet's generation. We aim to implement an attention-based Transformer model for the couplet generation. Different from the previous research, we will focus on designing a model that specifically fits the linguistic features of Chinese couplets. This includes adapting the model to meet the restricted requirements of pattern, vocabulary, and flatness for Chinese Couplets.

2 Related Work

Research in text generation first appeared in the 1970s (Goldman, 1974). And in the last decade, with the availability of large collections of datasets, automated conditional text generation rose to popularity. There have been several different approaches to the generation of Chinese poetry that have had success.

Each couplet is made up of two sentences, and given the first sentence, one can use Statistical Machine Translation model to generate the second sentence (Zhou et al., 2009). Because of the special structure of Chinese couplets, each character in the first sentence have a strong connection to the character in the same position in the second sentence. Therefore, Zhou, M. et al. (2009) used a phrase-based SMT to "translate" each character of the first sentence to a list of candidate characters that could act as its counterpart in the second sentence. Then, filter out candidates that do not satisfy linguistic constraints of couplets.

Xingxing Z. et al. (2014) drew inspiration from sequence to sequence learning using neural networks. They proposed a Recurrent Neural Networks model that generated quatrains based on some given keywords. The RNN learns the representation of different characters and the interaction between characters through a large collection of poems and generates lines in a poem based on the previous lines probabilistically. In contrast to the previous study by Zhou, M. et al (2009), Xingxing Z. et al. (2014) made no Markov assumptions about the dependency of the characters within a line.

In recent years, the development of an attention mechanism based Encoder-Decoder framework in the NLP area has made it possible for computers to understand the text better. Google researchers Ashish V. et al. (2017) proposed a new simple network architecture called the Transformer. Although the Transformer is solely based on the attention mechanism, it perfectly preserved the recurrence and convolutions in previous neural networks models.

J. Zhang, et al (2018) introduced an open-source Chinese couplets generation system called VV-Couplets. It is an attention-based sequence to sequence neural model that maps the first line of couplets to the second line, generating similar output text giving the input. Compared to previous work on Chinese couplets, they addressed the entity names of person and location specially in their model.

Although the Transformer model is widely used in the Natural Language Processing area, the existent framework is not fully explored for linguistic characteristics in Chinese. The vocabularies of Chinese couplets are anticipated and obsolete. Therefore, Wang Y, Zhang J, Zhang B, Jin Q (2021) further improve the Transformer's performance on Chinese couplets by intentionally adding POS taggings for special patterns in couplets, unregistered ancient words, and a polishing mechanism to improve the coherence.

3 Datasets

To make the results of our model more convincing, we perform experiments on the **Chinese Couplets Dataset** and **Chinese Poetry Dataset**.

3.1 Chinese Couplets Dataset

Chinese Couplets Dataset is available on `https://github.com/v-zich/couplet-clean-dataset`, which comes from an existing dataset in GitHub and contains around 740,000 couplets. Sensitive words in this dataset are deleted by searching the existing sensitive word vocabulary.

3.2 Chinese Poetry Dataset

Chinese Poetry Dataset is available on `https://github.com/hlthu/Chinese-Poetry-Dataset`, which comes from an existing dataset in GitHub and contains about 55,000 Tang poems. From a grammatical point of view, the third and fourth sentences basically conform to the grammar of couplets in Tang poetry. Therefore, these sentences can be used as training couplets. The number of our datasets will increase by around 40,000.

3.3 Data Preprocessing

Word embedding is a standard method for processing text and sequences. Word embedding solves the shortcomings of one-hot encoding-high dimensionality and irrelevance (Mikolov et al., 2013). Moreover, the spatial distance of the vector can reflect the semantic relevance between words. Generally speaking, embedding is a dictionary that maps integer indices to dense vectors. It receives integers as input, looks up these integers in the dictionary, and returns the associated vector. In our model, we use word embedding to vectorize all words for preprocessing.

4 Technical Challenge

Developing a model for the Chinese Couplet generation does not have a closed form solution. Meanwhile, writing a well matched Chinese Couplet for a given sentence is a challenging task for humans. In short, Chinese Couplet generation can be considered as a sequence-to-sequence question. Many researchers have tried to conquer this problem with LSTM, RNN, and transformer related models. In this project, we aim to improve the existing algorithms from two directions: embedding the Chinese Tokenization and a new decording mechanism.

4.1 Chinese Tokenization

In Chinese, each character can be considered as a word. But most of the words in the vocabulary consist of more than one character. There are more than three thousand common characters in Chinese. Hence, Chinese Tokenization itself is a challenging problem in the Chinese-NLP domain. Most of the existing research on Chinese Couplet generation does not embed Chinese Word tokenization step into their models. Only consider single character embedding, the model will inevitably lose some Information related to the connection between characters. Therefore, we want to test whether the embedment of the Chinese tokenization can help to fill up the information gap.

4.2 Decoding Mechanism

With the inspiration of the polish-up mechanism and bidirectional encoder model, we aim to develop a bidirectional decoder structure, which has not been seen on the existing works. In a regular sequence-to-sequence model, the output is generated one by one from the beginning of the sentence and provides the information for the later prediction.

4.3 Evaluation Metrics

To evaluate the quality of the generated couplets, we would mainly use three well-known automatic machine evaluation metrics: Preplexity score, Bilingual Evaluation Understudy (BLEU), and Rouge Score. For language models, preplexity is the most common use metric. A lower preplexity score indicates the model can generate a fluent couplet. A high BLEU or Rough score shows the model is capable of generating a couplet that is similar to the gold standard.

References

Neil M. Goldman 1974. *Computer Generation of Natural Language from a Deep Conceptual Base*. Ph.D. thesis, Standford University Stanford AI Laboratory Memo AIM-247 or CS Technical Report CS-74-461

Zhou, M., Jiang, L.,and He, J. 2009. *Generating Chinese couplets and quatrain using a statistical approach* Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Vol. 1 (pp. 43-52).

Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg; Dean, Jeffrey 2013. *Distributed Representations of Words and Phrases and their Compositionality* Advances in Neural Information Processing Systems, Vol. 26, 2013

Zhang, Xingxing, and Mirella Lapata. 2014. *Chinese poetry generation with recurrent neural networks.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014

Zhang, Xingxing, and Mirella Lapata. 2014. *Chinese poetry generation with recurrent neural networks.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin 2017. *Attention is All you Need.* NIPS 2017: 5998-6008.

Jiyuan Zhang, Zheling Zhang, Shiyue Zhang and Dong Wang 2018. *VV-Couplet: An open source Chinese couplet generation system* Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018, pp. 1756-1760, doi: 10.23919/APSIPA.2018.8659474.

Yufeng Wang, Jiang Zhang, Bo Zhang, and Qun Jin 2021. *Research and Implementation of Chinese Couplet Generation System With Attention Based Transformer Mechanism* IEEE transactions on computational social systems. Published online 2021:1-9. doi:10.1109/TCSS.2021.3072153.