Chinese Couplet Generation Status Report

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 Tasks Performed

1.1 Data Exploration and Preprocessing

To date, We had explored the dataset we proposed in the proposal. The Chinese Couplets Dataset contains 770,491 training data with various length and proper syntax. Due to the unique syntax of Chinese Couplet, additional attention is required in validating our dataset. We had successfully implemented the Embedding weighted by utilizing the pretrained weight from Google BERT-base-Chinese.

1.2 Model

The implementation of model had not been started yet, however, we had designed our model architecture. We plan to use the Bilingual Evaluation Understudy (BLEU) and Rouge Score as our model Evaluation metrics.

2 Risk and Challenges

We are currently behind our schedule of the original plan.

2.1 Data

Due to the complexity of the Chinese language, some issues might arise.

- It is likely that the BERT Embedding might not contains some Chinese characters in our training set or test set.
- Due to different syntax, additional attention to positional encoding might be needed in generating our dataset.

2.2 Model Implementation

To date, we have not implemented our model yet, we will encounter issues in implementing our model, and might need additional cloud help in model training.

3 Current Plans to Mitigate Risks

3.1 Team Collaboration

We are currently behind our schedule of original plan. To expedite the project implementation, we plan to set deadlines for milestones of the project, and assign tasks to each of the group members.

3.2 Data Preprocessing

To avoid losing information of unseen word in the BERT embedding, we can lower the restriction of the *Unknown* tag, and set the embedding to be trainable.

Data Preprocessing is essential in improving our neural network. We plan to use rule-based algorithm in generating the syntax information, and use it as a parameter in our model. Also, we can employ the Chinese part-of-speech (POS) tag as parameter to enhance the cohesion between words.

3.3 Model Implementation

We plan to use transformer and add syntax parameters we generated and part-of-speech (POS) tag to our dataset preprocessing.

In order to resolve the potential hardware limitation for model training, we can use the AWS platform to utilize the education credit from the course.

4 Significant Challenges that Changes Our Direction

Our model architecture did not vary significantly from our proposal, we add some more features, such as the syntax information and the part-of-speech (POS) tag to enhance our model.

We will evaluate different model architecture and might slightly change the model in the future as more implementations have been done.