HW3-1 report

R06944043
陳偉

1. Accuracy of the models on the testing data

Here I just list the results of one signal model, and no results ensemble.

model with attention:

- POS: 0.8724
- Rhyme: 0.9327
- Number of segments: 0.9997

model without attention:

- POS: 0.7972
- Rhyme: 0.8808
- Number of segments: 0.9930

2. Your model structure

My model structure is quite simple, a basic seq2seq, in order to diminish the time and memory in training and testing process. I also try beam search (of k = 5), the results of beam search on controlling is worse than results of only attention, but it do have generated sequences with higher LM score. By rule based method, like generating n answers per input, then, picking the one with highest score will obtain better results.

Model structure:
Encoder: 1 layer GRU (dim = 128), bidirectional.
Decoder: 1 layer GRU (dim = 128)

3. Experiments, such as

- **Parameter tuning**
  - My parameters summarization:
    - Large embeddings with 2048 dimensions achieved the best results, but only by a small margin. Even small embeddings with 128 dimensions seem to have sufficient capacity to capture most of the necessary semantic information.
    - LSTM Cells consistently outperformed GRU Cells.
    - Bidirectional encoders with 2 to 4 layers per- formed best. Deeper encoders were significantly more unstable to train, but show potential if they can be optimized well.
    - Deep 4-layer decoders slightly outperformed shallower decoders.
    - Attention yielded the overall best results.
    - Beam search yielded lower language perplexity.

- Larger vocabulary size get less “<unk>” in output.

- My final parameters:
  - epoch = 4
  - vocab_size = 50000
  - teacher_forcing_ratio=0.5
  - decoder output dropout = 0.2
  - Optimizer: Adam, clip_grad_norm max_grad_norm=5

- **Different kinds of attentions**

  I just use Luong global attention method.