Assignment 5

Gengcong Yan - 1009903

ELEC-E5510 - Speech Recognition

December 4, 2021

1 Question 1

1.1 A

$pairs\_batch\_test\_clean : WER = 0.24$
$pairs\_batch\_test\_other : WER = 0.37$

1.2 B

In the paper[1] where the dataset come from, it introduces the difference of recording quality and accents vary test performances. The model was trained based on higher recording quality and US English accents. when the test dataset suffer the lower recording quality and various accents, The performance can't be better on the clean dataset.

2 Question 2

2.1 A

$pairs\_batch\_test\_long : WER = 0.35.$
Longer utterances lower the performance of model.

2.2 B

Longer utterances means more information in longer sentences. The words in the sentences become more distant from each other, but are still linked to each other by information overall. Base on attention mechanism, we need to consider information from further back in sentences to predict current words, which increases inaccuracy and decreases performance.

3 Question 3

3.1 A

Greedy decoding is the most straightforward approach, which is fast and easy to understand. Each step we just select the word with highest probability. This is a greedily action sometimes output low quality words. In some cases with longer sequences. The approach may easily get stuck on particular word and form dead cycles in prediction, which means repeatedly assign some special words in a cycle. It significantly affects prediction results.

3.2 B

Beam search decoder can be applied in training process. Compared to consider only one word in greedy decoding, now we can keep some words in a single step. Taking more words into consideration generate more hypothesis. The score of words are base on the sum log probability of these hypothesis. It shares somewhat similar ideas with N-gram in that they both include more words to increase the accuracy of the model.

4 Question 4

4.1 A

In LSTM, when sample sequences are too long, there is a risk of losing information due to long distances between states and long time dependencies in training. The feature vectors in BLSTM contain information not only before the sequence but also after the sequence. So BLSTM requires a bi-direction flow of information and computation, which increases the model training time notably.

4.2 B

Transformer[2] model may be applied in the task. Abandoning the LSTM network structure, Transformers base on the encoder-decoder idea and the attention mechanism for model construction. Instead of relying on the hidden state of the past to capture dependencies on previous words, transformer processes a sentence as a whole, rather than word by word. So there is no risk of losing past information due to long dependencies in LSTM.

References

[1] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.