NLP – homework 4 **Ghazal Zamaninejad**

Part 1

g) In step function, we set values to -inf for each element which was padded (`# Set e_t to -inf where enc_masks has 1`). Applying softmax to these elements, results in 0, which it means no attention for them. It is necessary because padded tokens aren't important during training and were added in order to make all sequences have the same length.

h) model's corpus BLEU Score = `12.44119147888199`

```
1 !sh run.sh test
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
load test source sentences from [./chr_en_data/test.chr]
load test target sentences from [./chr_en_data/test.en]
load model from model.bin
Decoding: 100% 1000/1000 [00:39<00:00, 27.99it/s]
Corpus BLEU: 12.44119147888199
```

i)

i. dot product attention doesn't have learnable parameter in comparison to multiplicative attention. It could be both advantage and disadvantage; On one hand, its calculation is simpler because there's no need to train W. On the other hand, a learnable parameter can affect the result and might improve the training process.

ii. Additive attention uses activation function so it adds non-linearity to the calculations that it might be so useful. But it has more learnable parameters (weights) so its computation is more complex.

Part 2

a) As said, Cherokee is a polysynthetic language which it means words are composed of many morphemes (word parts that have independent meaning but may or may not be able to stand alone). These languages typically have long "sentence-words" so if we want the model be able to generate these words, subword-level embeddings should be used.

b) Both character-level and subword-level are smaller because less components should be stored. For example, in an English character-level embedding, we'll only have 26 characters which they don't have meaning. By combining them we can generate millions of meaningful words.

c) In multilingual training, we can study transfer, where insights gained through training on one language can be applied to the translation of other languages. We observe strong positive transfer towards low-resource languages, dramatically improving the translation quality of languages.

d)

i. Model was not able to predict 'a crown of daisies' because it predicted 'her hair' in the wrong position and it couldn't traceback. I think model has used a greedy approach and has chosen the most probable word in each position. We can solve the problem by using beam search.

ii. The model couldn't differentiate pronouns and it predicted 'it' instead of 'she'. This might be related to the sort of pronouns in Cherokee or the size of training data. By increasing the number of training samples, the problem might get resolved.

iii. Model couldn't predict the correct word and it tried to predicted the meaning of words instead of the exact word. By increasing the number of training samples, the problem might get resolved.

e)

i. line 34, "Oh, Charlotte," is produced which is exactly as desired so it is verbatim. It shows that MT model is trying to learn each word translation.

ii. line 46, 'And Jesus answered and said, We know not what the cup of the cup of the cup of the cup of the cup of the cup of the cup of the cup of the cup of me?' was produced while the reference translation was 'But Jesus answered and said, Ye know not what ye ask. Are ye able to drink the cup that I am about to drink? They say unto him, We are able.'. First part of the prediction was exactly same as the reference but after that it became wrong and started repeating. It seems when the sentence became too long it failed in predicting.

f)

i.

·) C1 P_1 صحت $= \min(\max(0,0),1) + \min(\max(1,1),1) + \min(\max(1,0),1)$

$+ \min(\max(1,0),1) + \min(\max(0,0),1) = 0+1+1+1+0 = 3$

P_1 مخرج $= 1+1+1+1+1 = 5 \Rightarrow P_1 = 3/5$

P_2 صحت $= \min(\max(0,0),1) + \min(\max(1,0),1) + \min(\max(1,0),1)$

$+ \min(\max(0,0),1) = 0+1+1+0 = 2$

P_2 مخرج $= 1+1+1+1 = 4 \Rightarrow P_2 = 1/2$

$len(c_1) = 5,\ len(r) = 4 \Rightarrow BP = 1$

$BLEU = 1 \times \exp(1/2 \ln 3/5 + 1/2 \ln 1/2) = \sqrt{3/5} \times \sqrt{1/2} = \sqrt{3/10} \simeq 0.54$

C2 P_1 صحت $= \min(\max(1,1),1) + \min(\max(1,0),1) + \min(\max(0,0),1) +$

$+ \min(\max(0,1),1) + \min(\max(0,1),1) = 1+1+0+1+1 = 4$

P_1 مخرج $= 1+1+1+1+1 = 5 \Rightarrow P_1 = 4/5$

P_2 صحت $= \min(\max(1,0),1) + \min(\max(0,0),1) + \min(\max(0,0),1) +$

$+ \min(\max(0,1),1) = 1+0+0+1 = 2$

P_2 مخرج $= 1+1+1+1 = 4 \Rightarrow P_2 = 1/2$

$len(c_2) = 5,\ len(r) = 4 \Rightarrow BP = 1$

$BLEU = 1 \times \exp(1/2 \ln 4/5 + 1/2 \ln 1/2) = \sqrt{4/5} \times \sqrt{1/2} = \sqrt{2/5} \simeq 0.63$

$BLEU(c_2) > BLEU(c_1)$

I agree second translation is better.

ii.

(ii) C_1 P_1 صورت $= \min(\max(0),1) + \min(\max(1),1) + \min(\max(1),1) + \min(\max(1),1) + \min(\max(1),1) + \min(\max(0),1) = 0+1+1+1+0 = 3$

P_1 مخرج $= 1+1+1+1+1 = 5 \Rightarrow P_1 = 3/5$

P_2 صورت $= \min(\max(0),1) + \min(\max(1),1) + \min(\max(1),1) + \min(\max(0),1) = 0+1+1+0 = 2$

P_2 مخرج $= 1+1+1+1 = 4 \Rightarrow P_2 = 1/2$

$len(c_1) = 5,\ len(r) = 6 \Rightarrow BP = \exp(1 - \frac{6}{5}) = \exp(\frac{-1}{5})$

$BLEU = \exp(\frac{-1}{5}) \cdot \exp(\frac{1}{2}\ln\frac{3}{5} + \frac{1}{2}\ln\frac{1}{2}) = \exp(\frac{-1}{5}) \cdot \sqrt{\frac{3}{5}\times\frac{1}{2}} = 0.44$

C_2 P_1 صورت $= \min(\max(1),1) + \min(\max(1),1) + \min(\max(0),1) + \min(\max(0),1) + \min(\max(0),1) = 1+1+0+0+0 = 2$

P_1 مخرج $= 1+1+1+1+1 = 5 \Rightarrow P_1 = 2/5$

P_2 صورت $= \min(\max(1),1) + \min(\max(0),1) + \min(\max(0),1) + \min(\max(0),1) = 1+0+0+0 = 1$

P_2 مخرج $= 1+1+1+1 = 4 \Rightarrow P_2 = 1/4$

$len(c_2) = 5,\ len(r) = 6 \Rightarrow BP = \exp(1 - \frac{6}{5}) = \exp(\frac{-1}{5})$

$BLEU = e^{\frac{-1}{5}} \cdot \exp(\frac{1}{2}\ln\frac{2}{5} + \frac{1}{2}\ln\frac{1}{4}) = e^{\frac{-1}{5}} \cdot \sqrt{\frac{2}{5}\times\frac{1}{4}} \simeq 0.25$

$BLEU(c_1) > BLEU(c_2)$

I don't agree that first translation is better.

iii. As we seen in the 2 previous questions, this might be problematic. Although second prediction was better, in (ii) that we considered only one reference, it accepted first translation as the best one. Moreover, there might be more than one way to translate a sentence because each language's vocabulary is wide.

iv. advantages:

- It is quick to calculate and easy to understand.
- It is language-independent making it straightforward to apply to NLP models.

disadvantages:

- It does not consider the meaning of words. It is perfectly acceptable to a human to use a different word with the same meaning eg. Use "watchman" instead of "guard". But Bleu Score considers that an incorrect word.
- It looks only for exact word matches. Sometimes a variant of the same word can be used eg. "rain" and "raining", but Bleu Score counts that as an error.