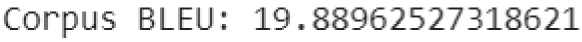
Hoorieh Sabzevari – 98412004

NLP – A4 (written)

1) Neural Machine Translation with RNNs

(g) The masks produced by the generate_sent_masks() function are used to prevent attention from being computed over the pad tokens in the input sequence. This is done by setting the attention scores corresponding to pad tokens to negative infinity, which results in those attention weights becoming zero after a softmax operation. It is necessary to use the masks in this way because including pad tokens in the attention computation would introduce noise and potentially degrade the quality of the model's predictions.

(h) 

```
Corpus BLEU: 19.88962527318621
```

(i)

i. The advantage of dot product attention compared to multiplicative attention is that it is faster to compute, making it more efficient for large matrices. However, a disadvantage of dot product attention is that it can have issues with scaling, which can lead to numerical instability or vanishing/exploding gradients during training.

ii. One advantage of additive attention over multiplicative attention is that it allows for more flexibility in capturing complex interactions between the query and key vectors. However, one disadvantage of additive attention is that it can be computationally more expensive than multiplicative attention due to the additional matrix multiplication involved in its calculation.

2) Analyzing NMT Systems

(a) Adding a 1D Convolutional layer after the embedding layer and before passing the embeddings into the bidirectional encoder can help our NMT system by capturing local dependencies among neighboring words, reducing dimensionality of input embeddings, introducing non-linearity and improving its ability to capture complex patterns and relationships between words in the input sequence.

(b)

i. 1. The error in the NMT translation is that it translated "贼人" as "the culprit" instead of "the culprits".

2. The reason could be due to the syntactic ambiguity in the source sentence, where "贼人" can refer to either a singular or plural entity.

3. To improve our model's performance, we have two options. We can either increase its complexity or prolong its training duration by running it for more epochs. However, if these measures fail to yield satisfactory results, we may need to gather additional data containing a sufficient number of singular and plural word pairs in order to further train our model.

ii. 1. The NMT system produced a repeated phrase "resources have been exhausted" in the translation.

2. The original sentence contains two separate pieces of information about the situation - lack of space and depletion of resources. It is possible that the NMT model failed to properly differentiate between these two concepts and instead produced a repetitive output.

3. To improve the model's ability to produce accurate translations, we can modify the attention mechanism, which is a crucial component of NMT systems that helps the model focus on relevant words.

To address the model limitation, we can provide more diverse training data covering a wide range of topics and styles of text. This will help the model learn language nuances, improving its ability to translate new sentences accurately.

iii. 1. The error in the NMT translation is that it mistranslates about "today's day" instead of "national mourning day."

2. It's possible that the model did not have enough training data or exposure to phrases like this word, leading to a lack of familiarity with the term.

3. One option could be providing the model with more training data that includes similar phrases and terms. we could adjust the attention mechanism so that the model focuses more on individual words and phrases within the sentence. Another potential solution would be tweaking the size of the hidden layers to help the model better capture the nuances of the language.

iv. 1. The error in the NMT translation is that it gives the opposite meaning of the original sentence.

2. One possible reason why the model made this error could be due to a specific linguistic construct in the source language. The phrase "唔做唔错" is

a common Chinese idiom that uses a double negative to express a positive meaning. However, this type of construction can be challenging for an NMT system to handle correctly, as it requires understanding the underlying meaning and context of the phrase.

3. One possible way to fix the observed error is to provide additional training data that includes examples of similar idioms with double negatives. o fix the error, we can add training data with similar idioms, modify the attention mechanism, adjust hidden layer size, or experiment with different hyperparameters.

(c) i.

c_1:

$$p_1 = \frac{4}{9} \qquad p_2 = \frac{3}{8}$$

c_2:

$$p_1 = \frac{6}{6} \qquad p_2 = \frac{3}{5}$$

$len(c_1) = 9$, $len(c_2) = 6$, $len(r_1) = 11$, $len(r_2) = 6$

$BP(c_1) = \exp\left(1 - \frac{11}{9}\right) = \exp\frac{-2}{9}$

$BP(c_2) = 1$

$BLEU(c_1) = \exp\left(\frac{-2}{9}\right) .\exp\left(\frac{1}{2}\frac{4}{9} + \frac{1}{2}\frac{3}{8}\right) = 1.2$

$BLEU(c_2) = \exp\left(\frac{1}{2} + \frac{1}{2}\frac{3}{5}\right) = 2.2$

The BLEU score of c_2 is higher than c_1 But the c_1 works better than c_2.

ii.

c_1:

$$p_1 = \frac{4}{9} \qquad p_2 = \frac{3}{8}$$

c_2:

$$p_1 = \frac{3}{6} \qquad p_2 = \frac{1}{5}$$

$BP(c_1) = 1$

$BP(c_2) = 1$

$BLEU(c_1) = \exp\left(\frac{1}{2}\frac{4}{9} + \frac{1}{2}\frac{3}{8}\right) = 1.5$

$BLEU(c_2) = \exp\left(\frac{1}{2}\frac{3}{6} + \frac{1}{2}\frac{1}{5}\right) = 1.4$

iii. Evaluating NMT systems with only one reference can be problematic because it may not fully capture the range of acceptable translations. This can lead to a biased evaluation of the system's performance. BLEU score addresses this issue by taking into account multiple reference translations and comparing them to the candidate translation, resulting in a more comprehensive evaluation of the quality of the translation.

iv. Advantages of BLEU is its objectivity and scalability. BLEU provides an automated, consistent, and unbiased score that can be applied to a large number of translations with relative ease. Also it’s language-independent so we don’t need to care about the language we use.

Disadvantages of BLEU include its limitations in capturing the full range of translation quality and its relative insensitivity to semantic and structural errors. Additionally, BLEU scores tend to favor translations that are closer to the reference translations, which may not always be the best translations from a human perspective.