Iran University of
Science &Technology

Natural Language Processing (CS22N)

Student Name:

Babak Behkamkia

Instructor Name:

Dr. Sauleh Eetemadi

Spring 2023

1. Neural Machine Translation with RNNs

g) First of all, we need to pad the sequences to make them the same length because otherwise we are not able to feed them into NMT model. After padding, we need to determine which tokens belong to the sequence and which are padding tokens. Thus, we need attention mask to identify padding tokens and do not compute attention for these tokens.

We put 1s for padding tokens in the attention mask, and when we want to compute the attention values, we fill these 1s with –inf because we use Softmax function to compute alpha values. Thus, the corresponding alpha values of these tokens turn to 0s because Softmax computes the exp function over the input elements.

h) BLEU score: 19.67851416

i)

i.

Advantage:

It is computationally more efficient than multiplicative attention.

Disadvantage:

It is sensitive to the scale of input vectors. If the input vectors have large magnitudes, the resulting attention weights can become very large as well, leading to instability and potential numerical issues during training.

ii.

Advantage:

It provides greater flexibility and captures complex relationships.

Disadvantage:

It is not computational efficient and comes at the cost of increased computational complexity.

2. Analyzing NMT Systems

a) By using a Conv1D layer we could extract local features in a sentence instead of just looking to a single word. Since the Mandarin Chinese characters have different meanings when they are combined. Thus, it will be helpful to distinguish the features related to more complex Chinese words

b)

i. **Source Sentence**: 贼人其后被警方拘捕及被判处盗窃罪名成立。

Reference Translation: the culprits were subsequently arrested and convicted.

NMT Translation: the culprit was subsequently arrested and sentenced to theft.

1. Lack in detection of plurality because Chinese language does not use an special sign for plurality.
2. The number of pair plural and singular sentences that are in the training dataset/ Embedding of the singular and plural form of the words are very similar/ model underfitting.
3. First, we make our model more complex or run it for more epochs. If it doesn't work, we could collect more data which consists a reasonable number of singular and plural pairs of words.

ii. **Source Sentence**: **几乎已**经没有地方容纳这些人, 资源已经用尽。

Reference Translation: *there is almost no space to accommodate these people, and resources*

have run out.
NMT Translation: *the resources have been exhausted and resources have been exhausted.*

1. Generating a sentence multiple times
2. Overfitting/ paying more attention to a part of sentence.
3. Using more complex attention mechanism/ changing hyper parameters/ adding a bias to words that used before in the same sentence.

iii. **Source Sentence**: 当局已经宣布今天是国殇日。

Reference Translation: *authorities have announced a national mourning today.*
NMT Translation: *the administration has announced today's day.*

1. Lack of correct word selection
2. It shows that out model struggle when it should predict some combination of words. The reason can be the number of training epochs, the complexity of the model, or it might occur due to the lack of these combinations in training data.
3. Using pre-trained models/ using more complex model. For example, adding some additional layers to dense layers/ training model for more epochs/ improving the training dataset.

iv. **Source Sentence:** 俗语有云:“唔做唔错”。

Reference Translation: *" act not, err not ", so a saying goes.*
NMT Translation: *as the saying goes, " it's not wrong. "*

1. The NMT model could not translate the proverb well.
2. Proverbs differ from normal text. If there is not enough proverb samples in the training data, it is understandable that an error like this happen.
3. Adding proverb instances to the dataset.

c)

i. 1-gram tables:

	There	Is	A	Need	For	Adequate	And	Predictable	Resources
c1	1	1	1	1	1	1	1	1	1
r1	0	0	0	0	0	0	1	1	1
r2	0	0	0	0	0	1	1	1	1

	Resources	Be	Sufficient	And	Predictable	To
c2	1	1	1	1	1	1
r1	1	2	1	1	1	2
r2	1	0	0	1	1	0

2-gram tables:

	There is	Is A	A Need	Need for	For adequate	Adequate and	And predictable	Predictable resources
c1	1	1	1	1	1	1	1	1
r1	0	0	0	0	0	0	0	0
r2	0	0	0	0	0	1	1	1

	Resources be	Be sufficient	Sufficient and	And Predictable	Predictable to
c2	1	1	1	1	1
r1	0	1	1	0	0
r2	0	0	0	1	0

$$P_n = \frac{\sum_{ngram \in C} \min\left(\max_{i=1,\dots,h} Count_{r_i}(ngram), Count_c(ngram)\right)}{\sum_{ngram \in C} Count_c(ngram)}$$

C_1 :

$$P_1 = \frac{0+0+0+0+0+1+1+1+1}{9} = \frac{4}{9}$$

$$P_2 = \frac{0+0+0+0+0+1+1+1}{8} = \frac{3}{8}$$

C_2 :

$$P_1 = \frac{1+1+1+1+1+1}{6} = \frac{6}{6}$$

$$P_2 = \frac{0+1+1+1+0}{5} = \frac{3}{5}$$

$$BP = \begin{cases} 1 & len(c) \geq len(r) \\ \exp\left(1 - \frac{len(r)}{len(c)}\right) & \text{otheWise} \end{cases}$$

$len(c_1) = 9$ $\quad$ $len(r_1) = 11$

$len(c_2) = 6$ $\quad$ $len(r_2) = 6$

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

$$BP(c_2) = 1$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{4} \lambda_n \log P_n\right)$$

$$BLEU(c_1) = \exp\left(-\frac{2}{9}\right) \cdot \exp\left(\frac{1}{2} \cdot \frac{4}{9} + \frac{1}{2} \cdot \frac{3}{8}\right) = 0.8 \cdot 1.5 = 1.2$$

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

Although C_1 is a better translation than C_2 , the BLEU score of C_2 is more than C_1 because the n-gram overlapping between C_2 and reference sentences are more than C_1 and reference sentences. This results show that the better BLEU score doesn’t show the better translation in all cases.

ii. 1-gram tables:

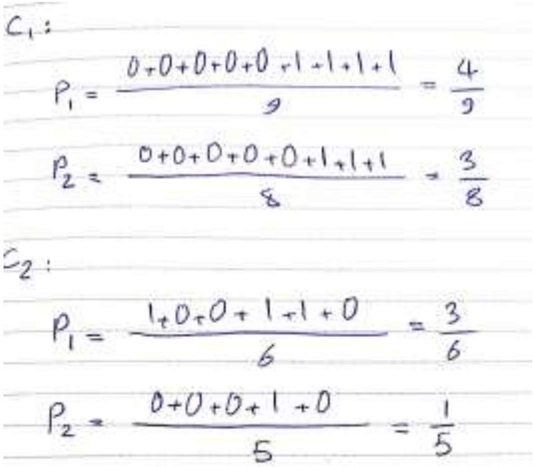
	There	Is	A	Need	For	Adequate	And	Predictable	Resources
c1	1	1	1	1	1	1	1	1	1
r2	0	0	0	0	0	1	1	1	1

	Resources	Be	Sufficient	And	Predictable	To
c2	1	1	1	1	1	1
r2	1	0	0	1	1	0

2-gram tables:

	There is	Is A	A Need	Need for	For adequate	Adequate and	And predictable	Predictable resources
c1	1	1	1	1	1	1	1	1
r2	0	0	0	0	0	1	1	1

	Resources be	Be sufficient	Sufficient and	And Predictable	Predictable to
c2	1	1	1	1	1
r2	0	0	0	1	0



C_1:

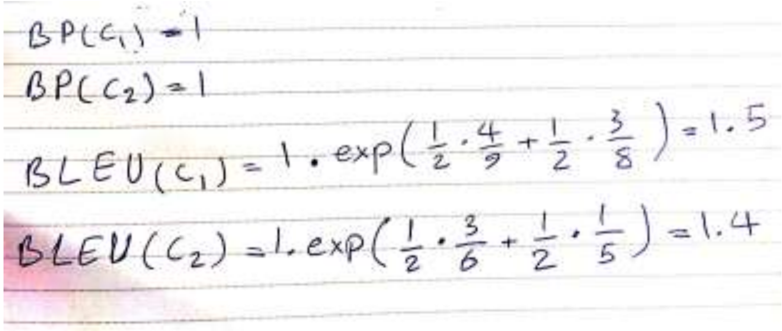
$$P_1 = \frac{0+0+0+0+0+1+1+1+1}{9} = \frac{4}{9}$$

$$P_2 = \frac{0+0+0+0+0+1+1+1}{8} = \frac{3}{8}$$

C_2:

$$P_1 = \frac{1+0+0+1+1+0}{6} = \frac{3}{6}$$

$$P_2 = \frac{0+0+0+1+0}{5} = \frac{1}{5}$$



$BP(C_1) = 1$

$BP(C_2) = 1$

$BLEU(C_1) = 1 \cdot \exp\left(\frac{1}{2} \cdot \frac{4}{9} + \frac{1}{2} \cdot \frac{3}{8}\right) = 1.5$

$BLEU(C_2) = 1 \cdot \exp\left(\frac{1}{2} \cdot \frac{3}{6} + \frac{1}{2} \cdot \frac{1}{5}\right) = 1.4$

iii. With just one reference sentence, BLEU score is not reliable because it might be noisy. Although it is the best evaluation method that we have. Having several reference sentences make this method more accurate because just one reference sentence may lead us to a misunderstanding.

iv.

Advantages:

1. This method increases the speed of evaluation. Evaluating all predictions by a human needs a large amount of time.
2. This method can be used for every language.

Disadvantages:

1. This method doesn't pay attention to the semantic of a sentence
2. Preparing reference sentences are requiring a large amount of human resource.