1 Neural Machine Translation with RNNs

(g)

Attention scores will be minus infinity in positions where enc_masks equals to 1, i.e. attention scores for padding words will be minus infinity and these words wont be considered in the attention output, because weights for padding words will be $e^{-\infty} = 0$. It is necessary to use the masks in this way because padding words don't really exist, so they shouldn't affect the attention output.

(j)

- Dot product attention:
  Advantage - computes asymptotically faster then two other attention mechanisms and doesn't have any additional parameters.
  Disadvantage - there is a restriction that vectors should have same dimensionality.
- Multiplicative attention:
  Advantage - in contrast to dot product attention has a learnable parameters.
  Disadvantage - computes asymptotically slower than dot product attention.
- Additive attention:
  Advantage - Has a nonlinearity which possibly allows the model to learn more complex functions.
  Disadvantage - the heaviest computations, comparing to two other attention techniques.

2 Analyzing NMT Systems

(a) i. Models translation doesn't sound naturally: "another favorite of my favorites". The reason for that mistake might be that model didn't learn that word "one"can be used not as a number. One of possible solutions is to get more data, where "one"is used in a similar context.

ii. Descriptive phrase "more reading in the U.S"is located right after "children changing the meaning of the sentences from "most read author in the U.S"to "children who read more in U.S than in other countries". Probably, this syntax structure is too complicated. For that problem solutions might be to get parallel corpus of more complex data.

iii. Problem - unknown word. That last name was deleted from the data as a rare word. Solution - just copy it from the source sentence to the target sentence.

iv.

v.

vi.

(b) i.

Sentence in Spanish - por 3 aos,
Reference English translation - She did it for three years.

NMT model's English translation - For three years.
Error - Person and cation is missing.
Reason - literal translation.
Solution - Add more similar, short sentences like this to the dataset.
ii.
Sentence in Spanish - Pueden ver a la izquierda un pequeo bote.
Reference English translation - You can see on the left side a small boat.
NMT model's English translation - You can see the left a little boat.
Error - two nouns in a row: "the left"and "a little boat".
Reason - "left"can be both boat adjective and the direction where the boat is.
Solution - add more examples where "left"is used as a direction.
(c) i.
c_1:
1gram = $\{the, love, can, alwayes, do\}$
$max_{i=1,...,k} count_{r_i}(\text{the}) = 0, count_c(\text{the}) = 1, min = 0$
$max_{i=1,...,k} count_{r_i}(\text{love}) = 1, count_c(\text{love}) = 1, min = 1$
$max_{i=1,...,k} count_{r_i}(\text{can}) = 1, count_c(\text{can}) = 1, min = 1$
$max_{i=1,...,k} count_{r_i}(\text{alwayes}) = 1, count_c(\text{alwayes}) = 1, min = 1$
$max_{i=1,...,k} count_{r_i}(\text{do}) = 0, count_c(\text{do}) = 1, min = 0$
numerator = $sum_{1gram \in c_1} min(..) = 3$
denominator = $sum_{1gram \in c_1} count_c(1gram) = 5$
$p_1 = \frac{3}{5}$
2gram = {the love, love can, can always, always do}
$max_{i=1,...,k} count_{r_i}(\text{the love}) = 0, count_c(\text{the love}) = 1, min = 0$
$max_{i=1,...,k} count_{r_i}(\text{love can}) = 1, count_c(\text{love can}) = 1, min = 1$
$max_{i=1,...,k} count_{r_i}(\text{can always}) = 1, count_c(\text{can always}) = 1, min = 1$
$max_{i=1,...,k} count_{r_i}(\text{always do}) = 0, count_c(\text{always do}) = 1, min = 0$
numerator = $sum_{2gram \in c_1} min(..) = 2$
denominator = $sum_{2gram \in c_1} count_c(2gram) = 4$
$p_2 = \frac{2}{4}$
$c = 5$, $r^* = 5$, $BP = 1$
$BLEU = 1 * \exp(\frac{1}{2} * \log(\frac{3}{5}) + \frac{1}{2} * \log(\frac{1}{2})) = 0.547723$
c_2:
1gram = {love, can, make, anything, possible}
$max_{i=1,...,k} count_{r_i}(\text{love}) = 1, count_c(\text{love}) = 1, min = 1$
$max_{i=1,...,k} count_{r_i}(\text{can}) = 1, count_c(\text{can}) = 1, min = 1$
$max_{i=1,...,k} count_{r_i}(\text{make}) = 0, count_c(\text{make}) = 1, min = 0$
$max_{i=1,...,k} count_{r_i}(\text{anything}) = 1, count_c(\text{anything}) = 1, min = 1$
$max_{i=1,...,k} count_{r_i}(\text{possible}) = 1, count_c(\text{possible}) = 1, min = 1$
numerator = $sum_{1gram \in c_2} min(..) = 4$
denominator = $sum_{1gram \in c_2} count_c(1gram) = 5$
$p_1 = \frac{4}{5}$
2gram = {love can, can make, make anything, anything possible}
$max_{i=1,...,k} count_{r_i}(\text{love can}) = 1, count_c(\text{love can}) = 1, min = 1$
$max_{i=1,...,k} count_{r_i}(\text{can make}) = 0, count_c(\text{can make}) = 1, min = 0$
$max_{i=1,...,k} count_{r_i}(\text{make anything}) = 0, count_c(\text{make anything}) = 1, min = 0$

$max_{i=1,\dots,k} count_{r_i}(\text{anything possible}) = 1, count_c(\text{anything possible}) = 1, min = 1$
numerator $= sum_{2gram \in c_2} min(..) = 2$
denominator $= sum_{2gram \in c_2} count_c(2gram) = 4$
$p_2 = \frac{1}{2}$
$c = 5$, $r^* = 5$, $BP = 1$
$BLEU = 1 * \exp(\frac{1}{2} * \log(\frac{4}{5}) + \frac{1}{2} * \log(\frac{1}{2})) = 0.632456$
According to the BLUE Score, second translation is better, I agree with that.

ii.

c_1:

$$p_1 = \frac{\sum_{1gram \in \{\text{the, love, can, always, do}\}} min(count_{r_1}(1gram), count_c(1gram))}{\sum_{1gram \in \{\text{the, love, can, always, do}\}} count_c(1gram)} = \frac{0+1+1+1+0}{1+1+1+1+1} =$$

$$= \frac{3}{5}$$

$$p_2 = \frac{\sum_{2gram \in \{\text{the love, love can, can always, always do}\}} min(count_{r_1}(2gram), count_c(2gram))}{\sum_{\{2gram \in \text{the love, love can, can always, always do}\}} count_c(2gram)} =$$

$$= \frac{0+1+1+0}{1+1+1+1} = \frac{1}{2}$$

$BLEU = 1 * \exp(\frac{1}{2} * \log(\frac{3}{5}) + \frac{1}{2} * \log(\frac{1}{2})) = 0.547723$

c_2:

$$p_1 = \frac{\sum_{1gram \in \{\text{love, can, make, anything, possible}\}} min(count_{r_1}(1gram), count_c(1gram))}{\sum_{1gram \in \{\text{love, can, make, anything, possible}\}} count_c(1gram)} = \frac{1+1+0+0+0}{1+1+1+1+1} =$$

$$= \frac{2}{5}$$

$$p_2 = \frac{\sum_{2gram \in \{\text{love can, can make, make anything, anything possible}\}} min(count_{r_1}(2gram), count_c(2gram))}{\sum_{\{2gram \in \text{love can, can make, make anything, anything possible}\}} count_c(2gram)} =$$

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

$BLEU = 1 * \exp(\frac{1}{2} * \log(\frac{2}{5}) + \frac{1}{2} * \log(\frac{1}{4})) = 0.316228$
Now, first translation receives higher BLUE score, but I don't agree that that translation is better.

iii.

Often, there are more that one possible correct translation and single reference that we have may differ from the quality translation of NMT system, so BLUE score just won't be representative and won't correlate with the given task.

iv.

Advantages:
1) Being automatic metric, BLEU can be computed effectively on the big corpus.
2) You can compute it by ourself, there is no need to ask or hire experts from other fields.
Disadvantages:
1) It doesn't correlate with final task perfectly.
2) It requires multiple reference translations for better quality.