CS224N 2020 Winter P4

1. NMT Seq2Seq

1.g Use of masks

- what effect the masks have on the entire attention computation:

  In attention computation, $e_{t,i} = (h_t^{dec})^T W_{attProj} h_i^{enc}$ will produce a T dimentional vector, with the ith component corresponds to the ith word in a sentence. The mask vector 'enc_masks' will put a '1' for each padded position, and it's embedding e_t is set to -inf where enc_masks has 1. It has the effect of having an attention 0 for each padded position.

- why it is necessary to use the masks in this way:

  Apart from achieving 0 attention for the padded position, the necessarty of using masks in this way is primarily because we use batch gradient descent during training, and by making each sentence the same length with padded words and masks the training can be done by GPU parallel computation. (Another thoughts of improving parallelization is to sample sentences with similar lengths into batches, hence reducing the number of padded position and improves the efficiency of training.)

1.i BLEU output

Corpus BLEU: 35.706406706846536

1.j Attention comparison

- advantage and disadvantage of dot product attention compared to multiplicative attention
  Dot product attention $e_t = (h_t^{dec})^T h^{enc}$ has the a much smaller amount of parameters because it has no attention projection matrix $W_{attProj}$.
  The limitation of Dot product attention is that embeddings in both encoder and decoder need to have the same dimension, which is not required in multiplication attention.

- advantage and disadvantage of additive attention compared to multiplicative attention
  Additive attention has better performance for larger dimensions; the hyper parameter can be adjusted in order to avoid high dimension computation, for example the dimension of v^T can be chosen a smaller number so as to achieve a better performance.

  In the case of smaller hidden state dimensionality, additive attention has a much more complex training model because it not only brings non-linearity into attention computation, but also has more training parameter.

2. Analyzing NMT

2.a

1. Here’s another favorite of my favorites, “The Starry Night”.
    - Problem: grammar issue, failed to use substitute word 'one'.
    - Reason: low-resource language pairs, the given target corpus isn't big enough to come up with a good language model
    - Solution: we can increase the size of the target language corpus so as to increase the quality of the target language model; or use a separate language model task and apply the learned target embeddings to the decoder;

1. You know what I do is write for children, and in fact, I’m probably the author for children, more reading in the U.S.
    - Problem: failed to generate grammarly correct long complex sentence.
    - Reason: Problem with maintaining context over long text. Another problem of not having a good language model.
    - Solution: we can increase the size of the target language corpus so as to increase the quality of the target language model; or use a separate language model task and apply the learned target embeddings to the decoder;

1. A friend of mine did that – Richard $<\mathrm{unk}>$.

    - Problem: failed to translate OOV words.
    - Reason: OOV words are missing from the embedding matrix.
    - Solution: In this particular case, the OOV problem with people's names can be fixed or mitigated by an NER task then simply copy the OOV name to the destination. Another possible solution is to learn OOV word embeddings from dictionary definitions, such as dict2vec.

1. You just have to go back to the apple to see it as an epiphany.
    - Problem: apple
    - Reason: domain mismatch , “manzana” has multiple meanings, it can be translated to either 'block' or 'mean depending on context.
    - Solution: add more language pairs containing 'manzana' -> 'block' in the corpus.

1. She saved my life by letting me go to the bathroom in the women's room.
    - Problem: women.
    - Reason: model limitations, translation has bias. In Spanish a lot of the words have different gender forms; but in English most of the words doesn't have different gender forms.
    - Solution: One way of mitigating the issue is by using larger corpus during training, or putting a gender modifier in front such as 'female teachers' as in this case.

1. That’s over 100,000 acres.
    - Problem: acres.
    - Reason: model limitations, Common sense error. Failed to deal with measurement system conversion

- Solution: Include more language pairs containing measurement system conversions in the corpus.

2.b

1. Error 1
   - src: Tambin necesita -- necesita dignidad, amor y placer. Y es nuestro trabajo proporcionar esas cosas.
   - ref: It also needs -- it needs dignity, love and pleasure, and it's our job to hand those things out.
   - nmt: It also needs -- needs dignity , love and pleasure , and it 's our job to provide those things .
   - error: Subjest missing: 'it needs dignity'
   - reason: Lack of language fluency caused by a good language model
   - solution: Increase the size of corpus, or use a pretrained language model.

1. Error 2
   - src: Le encontramos un lugar, la internamos, y la cuidamos y nos encargamos de su familia, porque era necesario,
   - ref: We found her one, we got her there, and we took care of her and watched over her family, because it was necessary.
   - nmt: We found a place , the $<\mathrm{unk}>$, and the $<\mathrm{unk}>$, and we take care of her family , because necessary .
   - error:
     A. OOV word problem;
     B. missing subject: because 'it was' necessary.
   - reason:
     A. OOV words are missing from the embedding matrix.
     B. Lack of language model
   - solution:
     A. Use dict2vec;
     B. Increase language model quality by pretrained language models.

2.c BLEU score

1. BLEU

- c1: the love can always do;

1 gram	Count r1	Count r2	max(ri)	min(1-gram)
the	0	0	0	0
love	1	1	1	1
can	1	0	1	1
always	1	0	1	1
do	0	0	0	0

This gives $p_1 = \frac{\sum \min(\max \text{Count}_{ri}, \text{Count}_c)}{\sum \text{Count}_c} = \frac{3}{5} = 0.6$

2 gram	Count r1	Count r2	max(ri)	min(2-gram)
the love	0	0	0	0
love can	1	0	1	1
can always	1	0	1	1
always do	0	0	0	0

This gives $p_2 = \frac{\sum \min(\max \text{Count}_{ri}, \text{Count}_c)}{\sum \text{Count}_c} = \frac{2}{4} = 0.5$
Since r1 and r2 are of the same lenth difference to c, use r2 in this case to calculate $\text{BP} = 1$;
Finally the blue for c1 is $\text{BLEU}_{c1} = e^{0.5 * \log 0.6 + 0.5 * \log 0.5} = 0.5477$

- c2: love can make anything possible;

1 gram	Count r1	Count r2	max(ri)	min(1-gram)
love	1	1	1	1
can	1	0	1	1

1 gram	Count r1	Count r2	max(ri)	min(1-gram)
make	0	0	0	0
anything	0	1	1	1
possible	0	1	1	1

This gives $p_1 = \frac{\sum \min(\max \text{Count}_{ri}, \text{Count}_c)}{\sum \text{Count}_c} = \frac{4}{5} = 0.8$

2 gram	Count r1	Count r2	max(ri)	min(2-gram)
love can	1	0	1	1
can make	0	0	0	0
make anything	0	0	0	0
anything possible	0	1	1	1

This gives $p_2 = \frac{\sum \min(\max \text{Count}_{ri}, \text{Count}_c)}{\sum \text{Count}_c} = \frac{2}{4} = 0.5$

Same as above, use r2 in this case to calculate $\text{BP} = 1$;

Finally the blue for c1 is $\text{BLEU}_{c1} = e^{0.5*\log 0.8 + 0.5*\log 0.5} = 0.6325$

2. BLEU w.r.t. r1 only

- c1:
  $p_1 = \frac{\sum \min(\max \text{Count}_{ri}, \text{Count}_c)}{\sum \text{Count}_c} = \frac{3}{5} = 0.6$
  $p_2 = \frac{\sum \min(\max \text{Count}_{ri}, \text{Count}_c)}{\sum \text{Count}_c} = \frac{2}{4} = 0.5$
  Use len(r1) for BP: $\text{BP} = 1 - \exp(1 - 6/5) = 0.8187$
  BLUE: $\text{BLEU}_{c1} = 0.8187 * e^{0.5 * \log 0.6 + 0.5 * \log 0.5} = 0.4484$
- c2:
  $p_1 = \frac{\sum \min(\max \text{Count}_{ri}, \text{Count}_c)}{\sum \text{Count}_c} = \frac{2}{5} = 0.4$
  $p_2 = \frac{\sum \min(\max \text{Count}_{ri}, \text{Count}_c)}{\sum \text{Count}_c} = \frac{1}{4} = 0.25$
  Use len(r1) for BP: $\text{BP} = 1 - \exp(1 - 6/5) = 0.8187$
  BLUE: $\text{BLEU}_{c1} = 0.8187 * e^{0.5 * \log 0.4 + 0.5 * \log 0.25} = 0.2589$

From the BLEU score, c1 is a better translation in this case.

3. Problem with single reference translation

The problem with only a single reference translation is that it may cause poor BLEU score even though the translation is very good. There are many different good translation out there, so a good way to increase n-gram overlap is by using several reference translations.

4. Advantages and disadvantages of BLEU

Advantages:

1. Easy to understand and cost efficient with minimal human labor;
2. Can be used on an on-going basis during system development to test changes

Disadvantages:

1. Hard to distinguish well between subtle differences.
2. A good translation can get a poor BLEU score because it has low n-gram overlap with the human translation