NLP Assignment 4

Parisa Yalsavar

May 1, 2021

1 Neural Machine Translation with RNNs

g

After computing attention scores and before applying softmax on them, these scores are masked using e_t . In constructing e_t, values of 1 in `enc_masks` are replaced with `-inf`, meaning we don't want to attent to these positions at all. This helps to ignore paddings and focus only on the real tokens of the sentence.

Masking is necessary because otherwise we would attend to tokens that are not real data and the result would become somehow corrupted.

h

```
[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 [01:55<00:00, 10.33it/s]
Corpus BLEU: 12.059856033661879
```

i

i.

Advantage: Simple computation compared to multiplicative attention. Due to its simplicity it's more memory efficient.

Disadvantage: Less flexible, because key and query must have the same dimension.

ii.

Advantage: Since each target hidden state and source hidden state have their own weight matrix and also a *tanh* is used, better results can be achieved.

Disadvantage: Has more computation complexity than multiplicative attention.

2 Analyzing NMT Systems

a

In polysynthetic languages words are composed of many morphemes. Breaking words into morphs (using subword-level embedding) can help reduce the embedding size compared to when a word-level embedding is used. Also in such languages working at subword level may increase model's inseparability from source language and thus increase performance.

b

For English there are about 470,000 unique words, while there is only 26 alphabet letters. For polysynthetic languages the number of morphemes is much less than number of words. Thus character-level and subword embeddings are often smaller than whole word embeddings.

c

As stated in multilingual training: Data skew across language-pairs creates an ideal scenario in which insights gained through training on one language can be applied to the translation of other languages. On one end of the distribution, there are high-resource languages like French, German and Spanish where there are billions of parallel examples, while on the other end, supervised data for low-resource languages such as Yoruba, Sindhi and Hawaiian, is limited to a few tens of thousands.

d

i.

- Reason: Either NMT cannot capture the dependency in *a crown of daisies in her hair* or it's decoding greedily which gives the translation *her hair* a higher score.
- Fix: Using Attention with LSTM can help better capture long-term dependencies. Greedy decoding can be replaced with beam search.

ii.

- Reason: As mentiond in Wikipedia Cherokee language pronouns in Cherokee does not represent gender, so the model may have acted in a greedy manner and chosen the token with the most score.
- Fix: Adding gender labels to the data can help this problem.

iii.

- Reason: Maybe there's a special word for *small fish* in Cherokee which translates to *LittleFish* in English or it's possible that *LittleFish* is a character's name.
- Fix: Using character-level models or copying the word from the source language can partially solve this problem.

e

a

Ground Truth: Verily I say unto you, This generation shall not pass away, till all things be accomplished.

Prediction: $< unk >$erily I say unto you, This generation shall not pass away, until all these things.

Model has got most of the translation right. This shows that model is good at translating simple sentences with common words.

b

Ground Truth: For I say unto you, Ye shall not see me henceforth, till ye shall say.

Prediction: For I say unto you, Ye shall not come unto you, until now come unto you.

In this example, at decoding model was not able to comprehend *see me henceforth, till ye shall say* and thus repeated *come unto you* instead. This can be a result of greedy decoding.

f

i.

Unigram	$\min(\max Count_{r_i}, Count_{c_1})$
the	0
love	1
can	1
always	1
do	0

Bigram	$\min(\max Count_{r_i}, Count_{c_1})$
the love	0
love can	1
can always	1
always do	0

Unigram	$\min(Count_{r_i}, Count_{c_2})$
love	1
can	1
make	0
anything	1
possible	1

Bigram	$\min(\max Count_{r_i}, Count_{c_2})$
love can	1
can make	0
make anything	0
anything possible	1

- For c_1
  - unigram: $p_1 = \frac{0+1+1+1+0}{5} = 0.6$
  - bigram: $p_2 = \frac{0+1+1+0}{4} = 0.5$
- For c_2
  - unigram: $p_1 = \frac{4}{5} = 0.8$
  - bigram: $p_2 = \frac{2}{4} = 0.5$

Because $len(c) = 5$ is greater than $len(r^*) = 4$ then $BP = 1$

$$BLEU_1 = BP \times \exp(0.5 \log 0.6 + 0.5 \log 0.5) = 0.5477$$

$$BLEU_2 = BP \times \exp(0.5 \log 0.8 + 0.5 \log 0.5) = 0.6324$$

The second translation *love can make anything possible*, has a higher BLEU score. I agree that c_2 is better than c_1 because c_1 doesn't really convey the meaning in the source sentence.

ii.

Unigram	$\min(Count_{r_1}, Count_{c_1})$
the	0
love	1
can	1
always	1
do	0

Bigram	$\min(Count_{r_1}, Count_{c_1})$
the love	0
love can	1
can always	1
always do	0

Unigram	$\min(Count_{r_1}, Count_{c_2})$
love	1
can	1
make	0
anything	0
possible	0

Bigram	$\min(Count_{r_1}, Count_{c_2})$
love can	1
can make	0
make anything	0
anything possible	0

- For c_1
  - unigram: $p_1 = \frac{3}{5} = 0.6$

- bigram: $p_2 = \frac{2}{4} = 0.5$

- For c_2
  - unigram: $p_1 = \frac{2}{5} = 0.4$
  - bigram: $p_2 = \frac{1}{4} = 0.25$

Since $len(c) = 5$, $len(r) = 6$ thus $BP = \exp(1 - (\frac{6}{5})) = 0.8187$.

$$BLEU_1 = BP \times \exp(0.5 \log 0.6 + 0.5 \log 0.5) = 0.4484$$

$$BLEU_2 = BP \times \exp(0.5 \log 0.4 + 0.5 \log 0.25) = 0.2589$$

According to BLEU score, c_1 is a better translation. I disagree with this.

iii.

Just like what happened in the previous section, using only one reference can cause problems. Translations that contain most of the n-grams in the reference will get a higher BLEU score, although there's no grantee that these sentences are good translations. For this reason there should be multiple references.

iv.

Advantages:

1. Faster to compute with easier implementation than human evaluation.
2. It's a language independent method. While human evaluation needs people that understand the source/target language.

Disadvantages:

1. Cannot measure the quality or sentence structure of the translation since it only measures n-gram overlaps.
2. It doesn't handle morphologically rich languages well.