CS224n Assignment 4 - NMT
Written Answers

Peng Ningxin (A0242091N)
pengn@u.nus.edu

April 2022

1 Neural Machine Translation with RNNs

(1-g)

i. The masks act as indicators of whether a token is a padding token or not. When calculating attention, we could use `masked_fill()` to efficiently set the attention value of tokens with `pad_mask == 1` to `-inf`. When applying softmax to this vector, the softmax values of tokens with `-inf` attention values would be `0`, thus being "ignored" by the Attention mechanism.

ii. It is essential and intuitive to ignore the padding tokens in the Attention mechanism, because these tokens have no real value and could even introduce unnecessary disruption to the attention calculation. Even worse, the token could be predicted as `<pad>` in the decoding stage. The implementation of `pad_mask` is flexible and efficient when we need to pad a specific token in the sequence.

(1-h)

The corpus BLEU score is **13.22**.

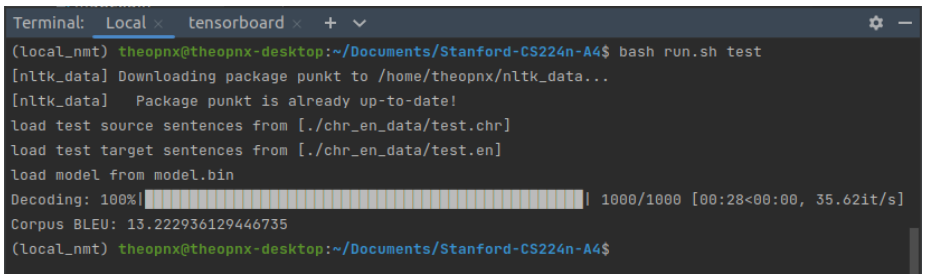

Figure 1: NMT chr2en - Experiment Result

(1-i)

i. Dot-product vs Multiplicative

- Advantage: Less parameters to train, less time and memory cost.
- Disadvantage: Less flexibility, due to the lack of a trainable weight matrix.

ii. Additive vs Multiplicative

- Advantage: The scale of the additive results are more controllable, given that it's easier to obtain huge values when applying multiplications than applying additions.
- Disadvantage: Compared to multiply operations, the addition operation takes more time, as Python has specified optimization for matrix multiplications.

2 Analyzing NMT Systems

(2-a)

As Cherokee is a polysynthetic language, the user could create new complex words based on different combination of morphemes, which could easily lead to Out-Of-Vocabulary (OOV) problems when using word-level embedding. Leveraging subword-level learning could not only alleviate OOV problems, but also make sure that words with similar spellings would have embeddings that are close with each other in the vector space.

(2-b)

When applying subword-level learning such as BPE, the subwords that show up frequently would be first extracted. Therefore, the "*ts-*" prefix, which is a common prefix in Cherokee, has a high possibility to be extracted as a subword-level token, and could be shared by all the words with "*ts-*" prefix. What's more, using transliterated Cherokee could allow us to obtain the raw morphemes that could not be further divided, while in original Cherokee a single character could contain multiple morphemes.

(2-c)

The basic idea of multilingual training is to find the generic representation of the language tokens among all languages, i.e, if the tokens are semantically close to each other, they should also share the similar representation in the hidden vector space. Therefore, if we jointly train the multilingual corpus in the same time, the high-resource corpus could help the model to learn the generic language representations that could also be used in low-resource languages.

(2-d)

i.
- The model translates "*her*" to "*it*", which is likely due to that the two words have similar spellings in Cherokee. Possible way to fix this problem is to introduce finer subword-learning mechanism.

- The model translated "*her*" to "*he*", which was supposed to be corresponded with "*she*" in previous contexts. This could be due to that the attention mechanism failed to notice the "*she*" token in the context. Possible way to fix this problem is to adjust the size of attention hidden size, or to introduce the temperature hyperparameter in softmax to flatten the distribution.

ii.
- The word "little" is repeated for multiple times, which is a common drawback of Beam Search decoding strategy. When applying beam search with beam size n, the model selects the token with **TOP** overall probability among candidates and keeps the top n beams for the next step. The repetition problem could occur if the probability of certain token is too large and the model keeps selecting it for multiple iterations. Ways to alleviate this problem includes: 1) Introduce penalties for repeated words such that prevent the model from selecting the same token for multiple times; 2) Introduce sampling methods to randomly select words from candidates with top-k probabilities rather than selecting the token with maximum probability.

iii.
- The model generates sequences that is too generic. This could be due to overregularization. We could try to fix the problem by reducing the scale of regularization.

(2-e)

i. Line 724

- **Ref:** *Grace to you and peace from God our Father and the Lord Jesus Christ.*
- **Hyp:** *Grace to you and peace from God our Father and the Lord Jesus Christ.*

Obviously, this is a perfectly good case. However, this sequence did occur verbatim in the training set – for three times. (Line 2962, 3930, 7734)

ii. Line 236

- **Ref:** *She went back into the house and put on her dress that looked like a watermelon and the hat that matched.*
- **Hyp:** *She went back into the house across the house and looked away.*

This error could be due to that the attention mechanism failed to focus on the information in the source sequence. We could try adjusting the hidden size of encoder or improve the attention framework to help the model focus more about the source sequences.

(2-f)

i.
- For c_1:
  $p_1 = 0.9231$, $p_2 = 0.8333$
  $len(c) = 13$, $len(r) = min(13, 14) = 13$, BP $= 1$
  BLEU score = 0.8771
- For c_2:
  $p_1 = 0.8462$, $p_2 = 0.75$

$len(c) = 13$, $len(r) = min(13, 14) = 13$, BP = 1
BLEU score = 0.7966

The first prediction is considered the better one according to the BLEU score, because it accurately covers more n-gram tokens such as *"in the"* and *"not comprehend"*.

ii. • For c_1: BLEU score = **0.7161**; For c_2: BLEU score = **0.7966**

If we just focus on the BLEU score, the second prediction is consider the better one, as it hits more n-gram tokens. However, if we look into the predictions by human eyes, we could notice that, comparing with the first one, the second prediction suffers from some grammatical incoherence. What's more, the second prediction also generates irrelevant words such as "trails".

iii. In machine translation task, we should be more concerned about whether the prediction retains the complete semantic information of source sentences, rather than judging whether it contains specific words/tokens. There are many situations where we could use multiple paraphrases or synonyms to express the same semantic meaning, which is exactly the point of using multiple references. Generally we could obtain more accurate and robust evaluation results using multiple references than using only one single reference.

iv. Advantages:

- Easy to compute, high efficiency, which could save the manpower cost.
- Irrelevant to language, which reduces the need of domain experts.

Disadvantages:

- No consideration for grammatical correctness.
- No consideration for synonyms or similar expressions.