CS224n Assignment 4: Neural Machine Translation

David Lee

November 5, 2019

1. Neural Machine Translation with RNNs

(g) First explain (in around three sentences) what effect the masks have on the entire attention computation. Then explain (in one or two sentences) why it is necessary to use the masks in this way.

2. Analyzing NMT Systems

(a) For each example of a Spanish source sentence, reference English translation, and NMT English translation

i. **Reference Translation**: *So another one of my favorites, The Starry Night.*
NMT Translation: *Heres another favorite of my favorites, The Starry Night.*

- Error: *So another one* vs. *Heres another favorite*
- Reason: NMT might use the "greedy decoding" since the *favorites* is at the back of the sentence, but when it can't modified the generated *favorite*.
- Fix: Maybe use beam search (or exhaustive search) decoding will solve this problem.

ii. **Reference Translation**: *You know, what I do is write for children, and Im probably Americas most widely read childrens author, in fact.*
NMT Translation: *You know what I do is write for children, and in fact, Im probably the author for children, more reading in the U.S.*

- Error: *and Im probably Americas most widely read childrens author, in fact.* vs. *and in fact, Im probably the author for children, more reading in the U.S.*
- Reason: I think it's because of the "long-term dependency", he has mentioned *what I do is write for children* but the NMT repeated same express in a similar way: *author for children*
- Fix: Maybe use LSTM (GRU) or Attention to capture long-term dependencies

iii. **Reference Translation**: *A friend of mine did that Richard Bolingbroke.*
NMT Translation: *A friend of mine did that Richard ¡unk¿*

- Error: The classic out-of-vocabulary (OOV) problem on the word *Bolingbroke*.
- Reason: Because this word didn't show up in the training data (or pre-trained embedding).
- Fix: Maybe we can use "character-level" (smaller granularity) decoder to generate the output. (if we're not allow to modify the training data)

iv. **Reference Translation**: *Youve just got to go around the block to see it as an epiphany.*
NMT Translation: *You just have to go back to the apple to see it as a epiphany.*

- Error: Grammar errors (e.g. *have just got to go* vs. *just have to go*, *an epiphany* vs. *a eiphany*) and some word choice error (e.g. *around* vs. *back to*, *block* vs. *apple*)
- Reason: I think it is because the lack of the training data (or epoches) that it still hasn't learned the correct grammar and word.
- Fix: More training corpus and epoches.

v. **Reference Translation**: *She saved my life by letting me go to the bathroom in the teachers lounge.*
NMT Translation: *She saved my life by letting me go to the bathroom in the womens room.*

- Error: *in the teachers lounge* vs. *in the womens room*
- Reason: I think because the sentence begin with *She* so in the training data the woman is more likely to be in the women's room than in teachers' lounge.
- Fix: Fix the data bias in training data. Maybe use some data augmentation trick to make it possible for woman in any other places.

vi. **Reference Translation**: *Thats more than 250 thousand acres.*
NMT Translation: *Thats over 100,000 acres.*

- Error: 100,000 hecta'reas is equal to 250 thousand acres.
- Reason: NMT don't know anything about unit conversion. e.g. NTD $\rightleftharpoons$ USD
- Fix: Maybe nowaday we can only apply some rules on that like capture the units seperately and translate it individually.

(b) Please identify 2 different examples of errors that your model produced.

(c) Please consider this example:

Reference Translation r_1: love can always find a way
Reference Translation r_2: love makes anything possible
NMT Translation c_1: the love can always do
NMT Translation c_2: love can make anything possible

i. Compute the BLEU scores for c_1 and c_2. And answer which of the two NMT translations is considered the better translation according to the BLEU Score? Do you agree that it is the better translation?

- For c_1

- unigram: $p_1 = \frac{\min(\max(3,1),5)}{5} = 0.6$
- bigram: $p_2 = \frac{2}{4} = 0.5$

- For c_2
  - unigram: $p_1 = \frac{4}{5} = 0.8$
  - bigram: $p_2 = \frac{2}{4} = 0.5$

Because $c = 5$ is greater than $r^* = 4$ thus $BP = 1$

$$BLEU_1 = BP \times \exp(0.5 \log 0.6 + 0.5 \log 0.5) = 0.5477225575051662$$

$$BLEU_2 = BP \times \exp(0.5 \log 0.8 + 0.5 \log 0.5) = 0.6324555320336759$$

The score of candidate sentence 2 c_2 is greater than candidate sentence 1 c_1.

In my opinion, I think the sentence 2 is indeed better than sentence 1. Because it describe both of the meaning of references.

ii. Recompute BLEU scores for c_1 and c_2, this time with respect to r_1 only. Which of the two NMT translations now receives the higher BLEU score? Do you agree that it is the better translation?

- 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: $p_2 = \frac{1}{4} = 0.25$

Because $c = 5, \quad r^* = 6$ thus

$$BP = \exp(1 - \frac{6}{5}) = 0.8187307530779819$$

$$BLEU_1 = BP \times \exp(0.5 \log 0.6 + 0.5 \log 0.5) = 0.448437301984003$$

$$BLEU_2 = BP \times \exp(0.5 \log 0.4 + 0.5 \log 0.25) = 0.25890539701513365$$

The score of candidate sentence 1 c_1 is greater than candidate sentence 2 c_2 now.

I'm not agree the sentence 1 is now better than sentence 2. In my opinion, I think it is because the lack of human labeling. A sentence should be able to express in many kind of ways especially in translation.

iii. Please explain (in a few sentences) why "NMT systems are often evaluated with respect to only a single reference translation (due to data availability)" may be problematic.

As the last exercise shows, when we have only one single reference translation, then it will probably restrict the expression. Even if we have a better translation but it will end up receives lower score.

iv. List two advantages and two disadvantages of BLEU, compared to human evaluation, as an evaluation metric for Machine Translation.

- Advantages
  - Make the evaluation quick, inexpensive, and absolutely objective.
  - Scoring become language-independent (just input references and candidates, and we don't have to care about what language we use)
- Disadvantages
  - Scoring is not flexible. There should be plenty of solution but it only evaluate based on the given references.
  - Can't evaluate too advanced translation. Because BLUE is comparison-based evaulation, it can't capture synonymous or similar phrase. Additionally, some more abstract metrics like adequacy, fidelity and fluency is even harder to scoring. (Even if evaluate by human may have different opinions.)