Practical No. 3

Adriana Bukała
ab394064@students.mimuw.pl

April 10, 2022

1 Part 1

g)

Firstly, masks differentiate between padding tokens and everything else. Then, e_t values are being set to $-\inf$. To compute attention, softmax function is applied to e_t, which will yield 0 for all padding tokens, since $\exp(-\inf) == 0$. That way decoder will not focus on padded tokens, which is an unwanted behaviour.

i)

Corpus BLEU: 10.736178498606622

j)

Dot product attention has no learnable parameters, which makes it the least computationally exhaustive, compared to the other two. However that comes at a cost of its expressiveness - it simply measures pairwise similarities between tokens.

Multiplicative attention is more sophisticated than dot product attention - learnable parameters are enabling it to better capture compatibility between tokens. Moreover, it allows more discrepancies between encoder and decoder's embedding space compared to the previous mechanism. That being said, additional parameters are naturally making multiplicative attention more computationally expensive than the dot product one.

Additive attention allows encoder and decoder's embedding space to diverge even more than multiplicative attention. Moreover, it enables modeling coverage of tokens, which discourages model to be repetitive. However it requires more computation than previous two, especially since non-multiplication operations are less optimized (source).

2 Part 2

a)

Side-note: I focused on line-to-line translations, but unfinished sentences were a major problem with the results.

a.1

`source:` You could view this as the number 65 one time or you

`NMT translation:` "Można popatrzeć na to jako liczbe 65 jednego czasu, albo"

`problem:` misunderstanding the word "times"

`probable cause:` training data set likely contains sentences with word "times" used in different context

`solution:` enriching training data set in examples referring to "times" as in multiplication, i.e. with similar context as the source sentence above

a.2

`source:` In this case he didn't get a majority and the standard procedure in Chile is that

`NMT translation:` "W tym przypadku nie ma żadnego tego, a odchylenie procedury w *unk* jest to, że"

`problem:` omission, misinterpretation, invalid syntax

`probable cause:` 1) "Chile" was out of vocabulary, 2&3) model has not been trained long enough

`solution:` fixing invalid syntax could require longer training, same for misinterpretation. Omission of "Chile" could be fixed by enriching training data with all possible country names, maybe along with the names of famous people having pages on Wikipedia (politicians, scientists etc.)

a.3

`source:` unrecognizable

`NMT translation:` unk unk DNA

`problem:` inability to translate source sentence

`probable cause:` lack of domain vocabulary in the training data

`solution:` enriching training data in domain specific vocabulary, for example biology or geography-oriented. Some school book could be a good source of those

a.4

`source:` "one, and since it's negative one of the numbers "

`NMT translation:` "po pierwsze, ponieważ jest to ujemne jeden z liczb "

`problem:` 1) misgendering word "liczba" (number), 2) misinterpretation of the sentence

`probable cause:` 1) NMT did not have enough data or (training) time to learn, 2) meaning of this sentence was lost, when it was splitted ("can do something interesting. So it could be || "one, and since it's negative one of the numbers ")

`solution:` 1) enriching training data with Polish words along with their declination (Wikipedia could be useful here, since it contains many examples of declination), 2) translating larger chunks of data, i.e. splitting sentences into larger part

a.5

`source:` This is using the distributive law of multiplication over

`NMT translation:` To jest używajac mnożenia mnożenia mnożenia

`problem:` repeats of word "mnożenie"

`probable cause:` training data did not contain similar wording, or model has not been trained long enough and just guessed the most probable math-y word

`solution:` obviously adding training data covering possible cause for errors is not an option, so one could think of implementing penalty for the beam search function, so that repeating one word in the exact same form is discouraged. This should be done very carefully, since many common words are obviously bound to be repeated, so one could additionally implement moving window, so that only side-by-side repeats are penalized. Another fix could be using additive attention with additional coverage vector (measuring attention in previous time steps), which would encourage NMT to avoid repetitions.

b)

In comparison to the models from PolEval, these results are quite bad. Winning model, SRPOL, achieved BLEU score of 28.23; second one, Google Translate, 16.83; and the third one, ModernMT, 16.29. All of these three are significantly better even if we were to continue cheating during evaluation. Without that, results are even worse with BLEU score of 6.44126089142462.

I think that the main reason of the drop is that Polish is a complicated language; starting with flection system (for example "rysował", "narysuje", "narysowawłszy", "rysujac", "dorysowały"), through strong genderization (for example "zaśpiewał", "zaśpiewała", "zaśpiewało", "zaśpiewali", "zaśpiewały"), ending with knotty grammar and syntax. All these nuances can not by comprehended by a NMT, if we do not feed it with well prepared training data set, containing a lot of complex sentences. Our model would probably benefit from regularization and longer training, but I think that lack of data is the largest bottleneck in this case (i.e. while not considering to switching to a different class of models - transformers).

c)

Question 1

Source Sentence s: So this means a strict subset.

Reference Translation r_1: Czyli to oznacza podzbiór właściwy.

Reference Translation r_2: W takim razie to oznacza podzbiór właściwy.

NMT Translation c_1: Czyli to podzbiór właściwy.

NMT Translation c_2: W takim razie to oznacza jest zbiór właściwy.

Side-note: all results were rounded to two significant figures.

`BLEU for c1`
1gram ϵc_1 := {Czyli, to, podzbiór, właściwy}

$$p_1 = \frac{1+1+1+1}{4} = 1$$

2gram ϵc_1 := {Czyli to, to podzbiór, podzbiór właściwy}

$$p_2 = \frac{1+0+1}{3} = 0.67$$

$$c = 4$$

$$r^* = 5$$

$$BP = exp(1 - \frac{r^*}{c}) = 0.78$$

$$BLEU = 0.78 * exp(0.5 * \log 1 + 0.5 * \log 0.67) = 0.78 * exp(0 - 0.2003) = 0.78 * 0.82 = 0.64$$

`BLEU for c2`
1gram ϵc_2 := {W, takim, razie, to, oznacza, jest, zbiór, właściwy}

$$p_1 = \frac{1+1+1+1+1+0+0+1}{8} = 0.75$$

2gram ϵc_2 := {W takim, takim razie, razie to, to oznacza, oznacza jest, jest zbiór, zbiór właściwy}

$$p_2 = \frac{1+1+1+1+0+0+0}{7} = 0.57$$

$$c = 8$$

$$r^* = 7$$

$$BP = 1$$

$$BLEU = 1 * exp(0.5 * \log 0.75 + 0.5 * \log 0.57) = exp(-0.14 - 0.28) = 0.66$$

According to $BLEU$ score, c_2 is a better translation, which is counter-intuitive, since c_2 has invalid syntax and mistakes a subset (podzbiór) with set (zbiór).

Question 2

Source Sentence s: So this means a strict subset.
Reference Translation r_1: Czyli to oznacza podzbiór właściwy.
NMT Translation c_1: Czyli to podzbiór właściwy.
NMT Translation c_2: W takim razie to oznacza jest zbiór właściwy.

`BLEU for c1` Nothing changes for c_1.

$$p_1 = \frac{1+1+1+1}{4} = 1$$

$$p_2 = \frac{1+0+1}{3} = 0.67$$

$$c = 4$$

$$r^* = 5$$

$$BP = 0.78$$

$$BLEU = 0.64$$

`BLEU for c2`

$$p_1 = \frac{0+0+0+1+1+0+0+1}{8} = 0.38$$

$$p_2 = \frac{0+0+0+1+0+0+0}{7} = 0.14$$

$$c = 8$$

$$r^* = 5$$

$$BP = 1$$

$$BLEU = 1 * exp(0.5 * \log 0.38 + 0.5 * \log 0.14) = exp(-0.48 - 0.98) = exp(-1.46) = 0.23$$

Now c_1 is a better translation, according to $BLEU$ score, and I agree with it.

Question 3

Natural language is not a deterministic entity, meaning that there is always more than one way of conveying specific information. That property of language would not be taken into the consideration of machine translation, if model has access to only one reference. That handicaps NMT of the ability to generalize well.

Question 4

Advantages:

1. automation of the evaluation process, making it faster and usually cheaper than expert evaluation,

2. BLEU is deterministic, i.e. results would be the same today and in a month, while human perspective and judgement is more flexible. That makes BLEU a solid comparison mechanism.

Disadvantages:

1. BLEU cannot differentiate between typos, minor and major errors, i.e. "wziać" (to take) and "wziaść" (which is an incorrect, but common spelling of the word "wziać") will be treated the same as "wziać" and its antonym - "dać" (to give). Human expert would easily notice the difference between those errors,

2. BLEU does not take word order into consideration, so nonsense translations will get high scores, if they contain the right words. That would not be tolerated by a human evaluator.