Practical No. 3

Emilia Wiśnios
ew407219@students.mimuw.edu.pl

April 4, 2022

Part 1

g)

Mask operation sets e_t values for padded tokens to negative infinity. It's because we want the probability of 'pad' token in the attention vector to be zero ($\exp(-\infty) = 0$).
If we don't apply the mask, the decode will use the information of padded tokens of hidden states.

i)

Obtained BLEU score: 11.649239633804452.

j)

- Dot product attention
  - Advantage: Space-efficient. Don't need additional weights with respect to the other two attention mechanisms.
  - Disadvantage: Query and values must have same dimension.
- Multiplicative attention
  - Advantage: Similar in complexity with additive attention, although multiplicative attention is faster and more-space efficient in practice as it can be implemented more efficiently using matrix multiplication.
  - Disadvantage: Additional storage of the weight matrix W with respect to the dot product attention.
- Additive attention
  - Advantage: Performs better for larger dimensions with respect to the multiplicative attention.
  - Disadvantage: Heavier complexity with respect to the other two attention mechanisms.

Part 2

a)

- Original sentence: *And there's two ways to interpret this.*

  Correct translation: *I są dwa sposoby interpretowania tego.*

  Model's translation: *I **jest** dwa sposoby przedstawienia tego.*

  In this case we have literal translation of *there's* for *jest*. In Polish language we conjugate to *są* in plural form. We can add more examples to the training set with examples like that.

- Original sentence: *We could make it work in all cases.*

  Correct translation: *działałyby prawidłowo w każdym przypadku*

  Model's translation: *Możemy to zrobić **w każdej** przypadkach.*

  Here we have declination problem. To make this translation correct it should be *we wszystkich* instead of *w każdej*. Again, the model made literal translation. Again, we could fix it by adding more complex examples to the dataset.

- Original sentence: *increase in temperature.*

  Correct translation: *wzrost temperatury.*

  Model's translation: *wzrost **winogron**.*

  The model mistakenly translated *temperatury* to *winogron*. Grapes occur in training set in economical context. Model probably associated increase with grapes. We could add more examples like this to the dataset.

- Original sentence: *Newton's first Law describes*

  Correct translation: *to o czym mówi I zasada dynamiki Newtona*

  Model's translation: *<unk> pierwsze prawo **NATURY***

  Model mistakenly translated *Newton* to *NATURY*. It is probably due to the examples from the training set. The word *prawo* often occurs together with *natury*. In order to fix it we should add more examples with *Newton*.

- Original sentence: *Which we call breast cancer.*

  Correct translation: *którego nazywamy rakiem piersi.*

  Model's translation: *Co nazywamy **piersi piersi <unk>**.*

  Model mistakenly translated *rakiem* to *piersi*. It may be caused by the fact that *rak* often occured together with *piersi* in the training set. We may try to fix it by adding more examples with different kinds of cancer to the dataset.

b)

In comparison to the models from PolEval our results are quite bad. The winning model obtained 28.23 BLUE score. In the second place there is Google Translate with the score 16.83.
Obtained BLUE score on full test set: 0.1928445876999296.
In my opinion, the difference is so large because Polish language has complex syntax, grammar and extensive flection system. To avoid this issue we can train our model on bigger dataset with more complex examples. We can also add some regularization to our model to avoid overfitting.

c)

1. For candidate 1:

   Unigrams:

Candidate Count	Count	Ref1 Count	Ref2 Count	Max Ref Count	Clip Count
czyli	1	1	0	1	1
to	1	1	1	1	1
podzbiór	1	1	1	1	1
właściwy	1	1	1	1	1
4					4

   Bigrams:

Candidate Count	Count	Ref1 Count	Ref2 Count	Max Ref Count	Clip Count
czyli to	1	1	0	1	1
to podzbiór	1	0	0	0	0
podzbiór właściwy	1	1	1	1	1
3					2

   So the modified precision score for the unigram is 1 and for bigram is $\frac{2}{3}$.

   Let's compute the brevity penalty BP. In our case the lenght of candidate sentence is 4 and the length of closest reference translation is 5. So BP is equal to $\exp\left(1 - \frac{5}{4}\right) = 0.78$.

   Finally, the BLEU score is equal

$$\text{BLEU} = 0.78 \cdot \exp\left(0.5 \cdot \log(1) + 0.5 \cdot \log\left(\frac{2}{3}\right)\right) = 0.64$$

   For candidate 2:

   Unigrams:

Candidate Count	Count	Ref1 Count	Ref2 Count	Max Ref Count	Clip Count
w	1	0	0	1	1
takim	1	0	1	1	1
razie	1	0	1	1	1
to	1	1	1	1	1
oznacza	1	1	1	1	1
jest	1	0	0	0	0
zbiór	1	0	0	0	0
właściwy	1	1	1	1	1
8					6

Bigrams:

Candidate Count	Count	Ref1 Count	Ref2 Count	Max Ref Count	Clip Count
w takim	1	0	1	1	1
takim razie	1	0	1	1	1
razie to	1	0	1	1	1
to oznacza	1	1	1	1	1
oznacza jest	1	0	0	0	0
jest zbiór	1	0	0	0	0
zbiór właściwy	1	0	0	0	0
7					4

So the modified precision score for the unigram is $\frac{6}{8}$ and for bigram is $\frac{4}{7}$. Let's compute the brevity penalty BP. In our case the lenght of candidate sentence is 8 and the length of closest reference translation is 7. So BP is equal 1.

Finally, the BLEU score is equal

$$\text{BLEU} = 1 \cdot \exp\left(0.5 \cdot \log\left(\frac{6}{8}\right) + 0.5 \cdot \log\left(\frac{4}{7}\right)\right) = 0.65$$

According to the BLEU score the second candidate is considered the better translation. I don't agree that it's better due to grammatical mistake in the second translation.

2. For candidate 1:

   Unigrams:

Candidate Count	Count	Ref1 Count	Max Ref Count	Clip Count
czyli	1	1	1	1
to	1	1	1	1
podzbiór	1	1	1	1
właściwy	1	1	1	1
4				4

Bi-grams:

Candidate Count	Count	Ref1 Count	Max Ref Count	Clip Count
czyli to	1	1	1	1
to podzbiór	1	0	0	0
podzbiór właściwy	1	1	1	1
3				2

So the modified precision score for the unigram is 1 and for bigram is $\frac{2}{3}$.

Let's compute the brevity penalty BP. In our case the lenght of candidate sentence is 4 and the length of closest reference translation is 5. So BP is equal to $\exp\left(1-\frac{5}{4}\right) = 0.78$.

Finally, the BLEU score is equal

$$\text{BLEU} = 0.78 \cdot \exp\left(0.5 \cdot \log(1) + 0.5 \cdot \log\left(\frac{2}{3}\right)\right) = 0.64$$

For candidate 2:

Unigrams:

Candidate Count	Count	Ref1 Count	Max Ref Count	Clip Count
w	1	0	0	0
takim	1	0	0	0
razie	1	0	0	0
to	1	1	1	1
oznacza	1	1	1	1
jest	1	0	0	0
zbiór	1	0	0	0
właściwy	1	1	1	1
8				3

Bigrams:

Candidate Count	Count	Ref1 Count	Max Ref Count	Clip Count
w takim	1	0	0	0
takim razie	1	0	0	0
razie to	1	0	0	0
to oznacza	1	1	1	1
oznacza jest	1	0	0	0
jest zbiór	1	0	0	0
zbiór właściwy	1	0	0	0
7				1

So the modified precision score for the unigram is $\frac{3}{8}$ and for bigram is $\frac{1}{7}$. Let's compute the brevity penalty BP. In our case the lenght of candidate sentence is 8 and the length of closest reference translation is 5. So BP is equal 1.

Finally, the BLEU score is equal

$$\text{BLEU} = 1 \cdot \exp\left(0.5 \cdot \log\left(\frac{3}{8}\right) + 0.5 \cdot \log\left(\frac{2}{7}\right)\right) = 0.23$$

According to the BLEU score, the first candidate is considered better translation. I agree that it's better.

3. Uncertainty is the core challenge in machine translation. Usually we can translate a sentence in many ways (all grammatically correct). Also in some languages the context is needed to properly translate a sentence (for instance to say who the object is). Evaluating NMT system with respect to only a single reference translation is causing an overfitting in some sense – we evaluate the model on single reference in a one-to-many task. [2]

4. Advantages:

   - Quick and inexpensive.
   - The score often correlates well with a human judgement. [1]

   Disadvantages:

   - Doesn't compare meaning.
   - Doesn't incorporate sentence structure.

References

[1] Deborah Coughlin. Correlating automated and human assessments of machine translation quality. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA, September 23-27 2003.

[2] Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation, 2018.