MT Exercise 5

Topics: Byte Pair Encoding, Beam Search

Repository link: *https://github.com/k-horn/mt-exercise-5*

Important Note:
We put in a lot of time and effort trying the complete this assignment but due to facing a few difficulties in understanding how to implement the tasks and also because of the exams we have during this week, we were unable to complete exercise 5 entirely. Part 1 is complete but for part 2 we were not able to program the graphs for the beam sizes and Bleu scores.

1 Experiments with Byte Pair Encoding

For the model experiments we chose the language pair nl-de. For the third model we decided to use the vocabulary size 6000 because while looking through the JoeyNMT documentation we didn't find a number that was considerably higher and thus increased the vocabulary size by three times. The beam size for all our models were set to the default of *beam_size=5*.

	Use BPE	Vocabulary size	BLEU
(a)	no	2000	6.6
(b)	yes	2000	14.7
(c)	yes	6000	16.6

Changes and additional steps taken for the experiments in Part 1:

1. First, we wrote the preprocess.sh script in which 100'000 sentences were extracted from the raw data-files, once for the source language and one for the target language train file.

2. For model 1 we adapted the transformer_model1.yaml and the train_model1.sh.

3. To build the vocab_file for models 2 and 3, we used the *joeynmt/script/build_vocab.py* file. This script expects a yaml config file as a positional argument in the command line. To get a single vocab_file we added the *'--join'* keyword argument:
   python3 joeynmt/scripts/build_vocab.py configs/model-name.yaml --join

4. To train model 2 and 3 we also adapted the corresponding config files and shell scripts.

5. To analyse the BLEU scores of the three models, we also adapted the evaluate_modelX.sh shell scripts.

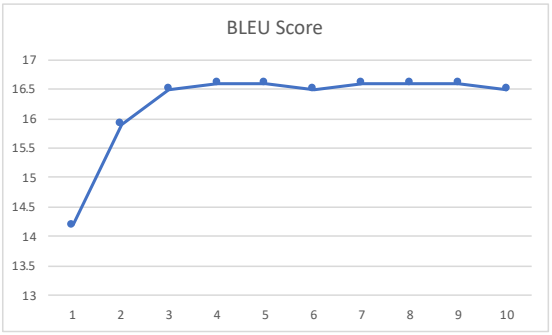
Our findings in Part 1:

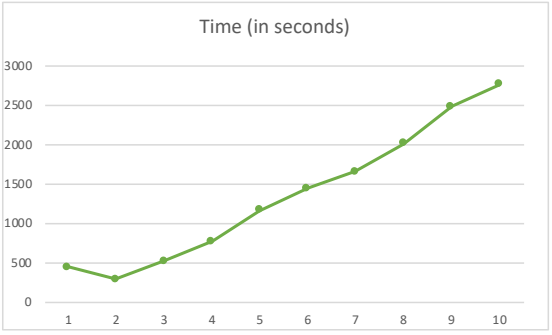
- Looking at the BLEU score it seems that the BPE models are significantly better. You can also observe that increasing the vocabulary size also has a positive effect on the BLEU score.
- Looking at the translation of model 1, there are many 'unk' tokens (7719 tokens) in the file. Furthermore, looking at the rest of the translation, it seems that the translation is grammatically correct.
- Model 2 has way less 'unk' tokens (186 tokens) and you do have sentences but aren't coherent with each other. You may get a rough idea what the text is about but the translation is not grammatically correct.
- Model 3 does not contain any 'unk' tokens but it is important to note that the quality of the translations are the same as with model 2;
- both models 2 and 3 have a lot of repetitions in their translations.

2 Impact of beam size on translation quality

We adapted the beam sizes in our model 3 which had the highest Bleu score:

Beam Size	BLEU Score	Time (in seconds)
1	14.2	434
2	15.9	283
3	16.5	515
4	16.6	764
5	16.6	1157
6	16.5	1442
7	16.6	1649
8	16.6	2005
9	16.6	2478
10	16.5	2762





Our findings in Part 2:

- Because we have a small vocabulary, there are not many hypothesis that can be built in parallel and therefore increasing the beam size doesn't drastically change the BLEU score after a certain size. The translations also become somewhat similar .
- Even looking at the translations of beam sizes 1 and 3, they show a slight improvement of the translation choices made (*'gibt es uns tiefe Einsichten'* vs *'gibt uns die tiefgreifende Einsichten'*). The translation quality of beam_size=5 and up doesn't significantly change. The time stamps show an increase with each bigger beam size. This makes sense because the algorithm has to compare more possible translations against each other, the bigger the beam size is.
- According to these observations, we would recommend choosing a beam size of 5 with our low-resource setting. This also corresponds with the default setting you will find in the JoeyNMT example-files.