MT Exercise 5 - Encoder-Decoder Models

Monday, May 21. 2018

Dataset & Preprocessing

I chose not to extend the provided dataset because of the limited time I had. Also the testset is of a similar domain to the provided data, so adding out of domain data would probably not improve the results that much.

First i **tokenized** all corpora, e.g for the german training corpus:

```
cat corpus.train.de | \
perl ../preprocessing/mosesdecoder/scripts/tokenizer/normalize-punctuation.perl de | \
perl ../preprocessing/mosesdecoder/scripts/tokenizer/tokenizer.perl -l de -q > \
corpus.train.tok.de &
```

I then trained a **truecase** model for each language, and applied it to all corpora.

```
perl ../preprocessing/mosesdecoder/scripts/recaser/train-truecaser.perl \
--corpus ../data/corpus.train.tok.de --model truecase-model.de &
```

```
cat corpus.train.tok.de | \
perl ../preprocessing/mosesdecoder/scripts/recaser/truecase.perl \
--model ../models/truecase-model.de > corpus.train.tc.de &
```

Next, i trained a combined **BPE** model for both languages with a size of 70000 and applied it.

```
python ../preprocessing/subword-nmt/learn_joint_bpe_and_vocab.py \
--input ../data/corpus.train.tc.de ../data/corpus.train.tc.en \
-s 70000 -o bpe.codes.joint --write-vocabulary vocab.de vocab.en &
```

```
../preprocessing/subword-nmt/apply_bpe.py -c ../models/bpe.codes.joint \
--vocabulary ../models/vocab.de --vocabulary-threshold 50 \
< corpus.train.tc.de > corpus.train.bpe.de &
```

I applied the following **postprocessing** script to revert the preprocessing effects.

```
# remove BPE splits
sed -r 's/(@@ )|(@@ ?$)//g' $1 > $1.noBPE

# moses postprocessing
cat $1.noBPE | perl mosesdecoder/scripts/recaser/detruecase.perl | perl
mosesdecoder/scripts/tokenizer/detokenizer.perl -l $2 -q | perl
mosesdecoder/scripts/tokenizer/normalize-punctuation.perl $2 > $1.out.$2
```

Adaptations & Hyperparameters

version	**branch**	**adaptions**
1	master	Include a attention model using the attention wrapper API
2	bidirectional	+ Implement a bidirectional encoder

The adaptions are based on https://github.com/tensorflow/nmt#basic

- For both versions I did not change the other default hyperparameters.
- Because of the limited time, version 1 of the model was only trained for 5 epochs, version 2 was trained for 7 epochs.

Translation Sample and Results

With model version 2 i achieved a BLEU score of 44 on the dev set.

Samples	
source text	**daikon translation**
Wird der Rat die Position unseres Parlaments noch lange ignorieren?	Will the Council ignore Parliament's position for a long time?
Sprache und Kultur haben eben schon mal einen militärischen Konflikt in diesem Land ausgelöst, und es ist damals dank eines breiten gesellschaftlichen Kompromisses gelungen, ihn nach einigen Tagen zu beruhigen und über langfristige Aktionen der OSZE Stabilität im Land herbeizuführen.	Language and culture have already caused a military conflict in this country, and it was thanks to a broad social compromise that it was able to calm down the country's long-term stability in the country after a few days.
In einen Krieg hineingezogen zu werden, ist sehr leicht, aber aus der Logik der militärischen Intervention auszusteigen, wenn die Stimmung eskaliert, ist fast unmöglich.	In a war, it is very easy to get into a war, but to be able to move from the logic of military intervention, if the mood escalates, is almost impossible.
Ja, Herr Jarzembowski, Sie haben Recht: Der Änderungsantrag 12 ist hinfällig.	Yes, Mr Jarzembowski, you are right: Amendment No 12 lapses.

The fully translated and postprocessed testset (using version 2 of the model) can be found at `daikon/results/corpus.test.en`