Alla Stöckli, Michael Stähli

MT Übung 5

Thema: Encoder-Decoder-Modelle

## Preprocessing

Für diese Übung haben wir die bereitgestellten Trainingsdaten mit 20’000 neuen Sätzen aus einem weiteren parallelen Korpus auf OPUS (EU Bookshop de-en) vergrössert. Anschliessend haben wir alle Dateien mit Moses tokenisiert. Dann haben wir das Truecasing-Modell von Moses trainiert und auf neuen Trainingsdateien laufen lassen. Als nächsten Schritt haben wir ein gemeinsames BPE-Modell auf den Trainingsdateien beider Sprachen trainiert und zusätzlich das Modell vergrössert. Dafür haben wir folgenden Befehl verwendet:

```
python subword-nmt/learn_joint_bpe_and_vocab.py --input corpus.train.tc.de corpus.train.tc.en -s 75000 -o joined.bpe.codes --write-vocabulary vocab.de vocab.en
```

Anschliessend haben wir BPE-Encoding auf unseren Dateien durchgeführt. Schliesslich haben wir das MÜ-System mit 6 Epochen trainiert, wie in der Vorgabe. Weitere Hyperparamter haben wir nicht verändert.

## Code-Veränderungen

Um den Code zur verändern hatten wir uns vorgenommen, erstens die Source-Sequenz umzukehren und zweitens unbekannte Wörter im Decoding zu unterdrücken. Um den Code zu verändern und die Implementierungen zu testen, hatten wir uns überlegt, das lokal auf unseren Computern zu machen, da es auf den Servern etwas umständlich ist. Wir haben damit aber zu spät begonnen und die Installation von Tensorflow auf Pycharm hat sehr viel Zeit in Anspruch genommen, sodass wir am Ende keine Veränderungen mehr vorgenommen haben. Vielleicht wäre es effizienter gewesen, den Code direkt auf dem Server zu verändern und testen.