Machine Translation Übung 5

Lukas Yu 14-720-866

13. Mai 2019

1 Github repository

`https://github.com/Gizmondd/daikon`

2 Preprocessing

Für das Preprocessing habe ich die gleichen Schritte analog zur zweiten Übung angewandt. Das heisst:

- Normalisierung
- Tokenisierung
- Cleaning
- Truecasing
- BPE

Der einzige Unterschied bestand darin, dass ich die Symbolanzahl beim Lernen des BPE-Modells auf 100'000 erhöht habe. Ich habe mir damit erhofft, dass die vergrösserte Symbolanzahl das Vokabular erweitert wird, das vor allem bei deutschen Komposita eine bessere Performance aufweist. Das BPE-Modell habe ich wie in der zweiten Übung schon mit den Trainingsdatensets beider Sprachen gelernt.

3 Training

Für das Training wurde die Learning-Rate von 0.0001 auf 0.001 erhöht da der Defaultwert mir etwas zu tief erschien und das Training somit auch schneller geht. Zusätzlich wird die Input-Sequenz in umgekehrter Reihenfolge eingelesen, da dies nach Sutskever et al. (2014) die Performance der LSTM erhöht.

Alles anderen Einstellungen blieben identisch zur Default-Konfiguration. Gelernt wurde mit 10 Epochen.

4 Resulate

Auf dem Devset erreichte ich einen BLEU-Score von 14.4, was eine deutliche Verbesserung gegenüber dem Baseline darstellt.