Maschinelle Übersetzung Übung 4 Sophia Conrad

Bericht

Ich habe in der Übung zunächst mit einer englischen Übersetzung von Homers Ilias von gutenberg.org gearbeitet und das Datenset später mit der Odyssee erweitert. Ich war daran interessiert, wie gut das Sampling von Sätzen funktioniert, wenn die Trainingsdaten aus meistens langen und oft syntaktisch verschachtelten Sätzen bestehen. Das Preprocessing (preprocessor.py) bestand nur darin, das gewünschte Format – einen Satz pro Zeile – zu erhalten und ich habe manuell die Gutenberg Credits am Anfang und Ende der beiden Bücher entfernt. Ich hatte bereits zwei Trainings auf den nicht-formatierten Daten (ein langer String statt ein Satz pro Zeile) und nach dem Preprocessing wurde das Modell wider Erwartung beachtlich schlechter. Von den hart-kodierten Hyperparametern habe ich nur die hidden layer size verdoppelt (von 1500 auf 3000) um auszuprobieren, ob dies bereits Overfitting verursacht und ob die Rechenzeit noch legitim ist. Ich habe vermutet, dass grössere hidden layers das System bloss verbessern würden; das Training hat in etwa 1.5 mal länger gedauert. Dann habe ich noch in der Architektur (compgraph.py) die BasicLSTMCell durch eine LSTMCell ersetzt, weil das sogar in der tensorflow Dokumentation empfohlen wird («DEPRECATED: Please use `tf.nn.rnn_cell.LSTMCell` instead.») Das Modell wurde dadurch ein wenig besser.

Dev Perplexität auf dem nicht-adaptierten, nicht-erweiterten (nur Ilias) Modell: 156.67

Dev Perplexität auf dem nicht-adaptierten, erweiterten (Ilias+Odyssee) Modell: 149.50

Dev Perplexität auf dem fertig adaptierten Modell: 155.24