Bericht für Übung 04 Maschinelle Übersetzung

Datenset

Mein Datenset ist ein Datenset der Uni Leipzig in Englisch. Es ist eine Sammlung von Texten aus News, die in verschiedenen Format zugänglich sind: Wörtern, Sätze usw. Für diese Aufgabe habe ich natürlich die Sätze verwendet. Alle waren auf einer Linie, mit einem ID und einem Tabulator. Diese habe ich (direkt im Editor) weggenommen (also kein Skript dafür). Der Rest war eigentlich ganz okay. Die Datei war alphabetisch geordnet.
Da sie aber ziemlich gross war (>1'000'000) und ich aus Zeitdruck eher kurze Trainingszeiten haben wollte habe ich deren Grössenstark reduziert. Alle Sätzen, die mit einer Nummer anfingen, habe ich weggelöscht (immer im Editor), und dann habe ich mein Datenset auf 80'000 Linien reduziert. Diese wurden im dataset.train (78'000 Linien) und dataset.dev (2'000 Linien) verteilt. Mit diesen habe ich das Modell trainiert.
Ich habe dieses Datenset ausgewählt, da ich damit minimal Preprocessing haben würde, da die Arbeit schon vorher gemacht wurde (immer Zeitdruck ;-)).

Code-Veränderungen

Warnung: die Commits habe ich etwas künstlich einen nach dem anderen gemacht, damit ihr das Prozess sieht.
Ich habe gedacht, dass die Datensetgrösse sicher mein grösstes Problem war. Deswegen habe ich mich auch entschieden, diese zu erhöhen (von 80'000 auf 100'000 Linien). Deswegen habe ich mich dann dafür entschieden, das Datenset grösser zu machen. Ausserdem hatte ich noch ein bisschen weiter mit den Hyperparametern experimentiert:

Erster Versuch: num_steps:100, epochs:13, hidden_size: 1024, batch_size:128, vocab_max_size: 10000, embedding_size: 256. Ich dachte, mit einer grösseren batch_size würden mehr Informaiton dem System «gefüttert», und mehr Epochs aus verschiedenen Gründen: die Datei war grösser und ich erwartete, dass das System wegen der vergrösserten batch size komplexer wäre, deswegen konnten mehr Durchgängen helfen. Es ist mir aber ehrlich gesagt noch nicht super klar, wie alle diese Hyperparameters zusammen spielen, von dem her kann es auch Unsinn sein. Perplexität : 106.61 → Das Sample habe ich mit diesem Modell versucht zu erzeugen (beste Resultate)

Zweiter Versuch: die batch-Grösse wieder wie in das Originale Skript zu setzen. Dies war mehr als Neugier, ich wollte nämlich wissen, ob das System viel schneller bzw. schlechter werden würde. num_steps:100, epochs:13, hidden_size: 512, batch_size:64, vocab_max_size: 10000, embedding_size: 256. Perplexität: 110.50

Perplexität beim Originalen war: 113.5
Perplexität beim Originalen Skript aber mit grösseren Datenset: 112.77 → das heisst, meine Änderungen am Skript haben das System mehr verbessert, als wenn ich nur die Datensetgrösse geändert hätte!