Maschinelle Übersetzung FS 2018
Institut für Computerlinguistik
Anja Ryser Universität Zürich

MT Übung 4

Thema: RNNs

Zu Beginn wollte ich das System auf Liedtexte trainieren, allerding fand ich kein geeignetes Korpus. Danach wollte ich ein Set mit Witzen aus den vorgeschlagenen Sets nehmen, allerdings war dieses viel zu klein. Schlussendlich entschied ich mich für das Gesamtwerk von Edgar Allan Poe, das war gross genug. Später merkte ich, dass die gemischte Textform (Fliesstext und Gedichte) ein Problem sein könnte, Um das etwas einzudämmen fügte ich die Gedichte im Preprocessing auf eine Linie zusammen, jede Verszeile von den anderen durch ein Leerzeichen getrennt. Weitere Probleme, die mir im Preprocessing begegneten waren Abkürzungen. Die, die mit Grossbuchstaben geschrieben worden waren, konnte ich als Wörter verarbeiten, allerdings hatte es auch solche, die kleingeschrieben waren. Die sind jetzt als einzelbuchstaben im Datenset enthalten. Um dieses Problem zu lösen müsste man wahrscheinlich einen Dictionary mit allen Abkürzungen, die vorkommen, erstellen, und das Preprocessing müsste überprüfen, ob das, was es vor sich hat, in diesem vorhanden ist. So wie ich das Preprocessing aufgebaut habe, ist das allerdings nicht möglich, da ich Zeichen für Zeichen vorgehe und nur das letzte speichere. Dies führte auch zu einem anderen Problem: Wenn nach einem Satzpunkt noch ein Satzzeichen folgte, schrieb er auch dieses auf die nächste Zeile.

Nach anfangs vielen Fehl Versuchen machte ich folgendes:

Zu Anfang trainierte ich das System mit folgenden Parametern:

Vokabular Grösse beschränkt auf 10 000

Hidden Layers = 150

Learning Rate = 0.001

Num_steps = 100

Ich überlegte mir, dass das System besser wird, wenn man das Vokabular einschränkt, da es dann viel weniger Wörter, die nur einmal vorkommen lernen muss. (Und stattdessen ein unknown macht.) Die 1500 Hidden Layer dünkten mich ein bisschen viel und mich nahm Wunder, was passieren würde, wenn man diese auf 150 beschränkt. Die learning rate erhöhte ich, um den Fortschritt des Systems zu beschleunigen, allerdings nicht zu viel, damit es nicht über das Ziel hinausschiesst. Die Backpropagation setzte ich auch höher an, da die ersten Layer die wichtigsten sind und mit einer tiefen Zahl erreicht man die gar nie.

Das zweite Training führte ich mit denselben Werten durch, ich veränderte das Preprocessing allerdings so, dass der Punkt am Wort dran bleibt, um zu berücksichtigen, dass einige Worte eher am Schluss vorkommen als andere. Dadurch stieg aber die Perplexität auf dem Trainingsset (von 143.95 auf 154,05).

Bei dritten Mal änderte ich nochmals die Werte, die Hidden Layers erhöhte ich zum vergleich nochmals auf 1500, die Numsteps im gleichen verhältnis wie oben angepasst (1000), und ich trainierte mit ersterem Preprocessing. Dieser Durchgang war sehr langsam, und zeigte keine

Iteration an, nur die Epochen. Ich nehme an, dass das mit dem Num_step zu tun hat. Dieser durchgang erreichte eine Perplexity im Trainigsset von 120.12. Dieses Modell benutzte ich für den Rest der Aufgaben.

Scoring: 126.33