Jasmin Engelmann 11-747-334

Maschinelle Übersetzung, Übung 4

Datenset

Ich habe die gesammelten Werke Shakespeare's von Project Gutenberg verwendet. Meine Gründe dafür waren, dass dieses Testset gross genug war (5.53 MB), der Text in einem plaintext-Format vorlag und der Schreibstil sich von anderen Stilen unterscheidet.

Preprocessing

Ich habe mich dafür entschieden, auf Character-Level zu arbeiten. Denn so besteht auch die Möglichkeit, dass neue Namen erfunden werden könnten. Ausserdem ist es in meinen Augen eine grundlegendere Art, ein Sprachmodell zu trainieren. Auch Schreibfehler fallen dann weniger ins Gewicht, als wenn auf Wort-Lever trainiert würde. Vor dem ersten Training habe ich nur alle Zeichen durch Whitespaces getrennt und existierende Whitespaces durch das Symbol <space> ersetzt. Für das zweite Training habe ich den NLTK tokenizer verwendet, um auf Satz-Ebene zu tokenisieren und damit sicher zu gehen, dass eine 'line' nur einen Satz beinhaltet. Ausserdem habe ich leere Zeilen und Seitenzahlen heraus gelöscht. Da der Text vor dem Training geshuffelt wird, ist es unnötig, das Layout beizubehalten. Ich musste ziemlich lange herum probieren, bis ich es geschafft habe, leere Zeilen zu löschen. Wieder einmal habe ich bemerkt, wie wichtig die Reihenfolge der einzelnen Preprocessing Schritte ist; die leeren Zeilen müssen zuerst entfernt werden, dann wird tokenisiert. Auf die nun so entstandene Liste von Sätzen kann dann elementweise zugegriffen werden, um dann unnötige Whitespaces zu löschen, nach jedem Zeichen wieder Whitespaces einzufügen und dann dort, wo zwei hintereinander stehen, <scpace> einzufügen.

Postprocessing

Beim Postprocessing werden zuerst die <eos> Tags entfernt, dann die Whitespaces und schlussendlich <space> wieder durch Whitespaces ersetzt.

Hyperparameter

Ich habe für meine Trainings keine Hyperparameter benutzt, was sich im nach hinein als Fehler herausgestellt hat. Denn da ich beim zweiten Training keine neuen Speicherorte für die Logs und das Modell angegeben habe, konnte Tensorboard die Dateien nicht interpretieren und damit keine Loss-Übersicht darstellen:

plugin_event_accumulator.py:294] Found more than one graph event per run, or there was a metagraph cont aining a graph_def, as well as one or more graph events. Overwriting the graph with the newest event. W0430 04:15:57.371688 140117728683776 plugin_event_accumulator.py:302] Found more than one metagraph event per run. Overwriting the metagraph with the newest event.

Perplexity

Erstes Training: 5.30

Zweites Training: 4.08

Evaluation

Wie ich erwartet habe, enthält der gesampelte Text Namen, die nicht in Shakespeare's Werken vorkommen (z.B. Pandarus, Arcite). Der Text liest sich wie Shakespeare, jedoch macht er noch wenig Sinn. Es werden Wort-Neuerfindungen wie 'mansio' oder 'philo' erzeugt – ganz nach Shakespeare!

Die Perplexity war im zweiten Training etwas besser als im ersten Durchlauf.