Universität Zürich, Institut für Computerlinguistik
Maschinelle Übersetzung Vera Bernhard
Mathias Müller 30.04.2019

MT Übung 4: RNNs

Datenset

Für das Datenset habe ich den Archimob Korpus verwendet. Dabei handelt es sich um das grösste Oral-History Projekt der Schweiz, welche aus einer Zusammenarbeit von Historiker_Innen mit Filmacher_Innen enstand: Es wurden die Zeitzeug_Innen vom zweiten Weltkrieg zum Thema «Swissbankclaims» («das Verfahren um die Vermögenswerte der jüdische Opfer bei Schweizer Banken»[1]) befragt, um eine alternative Sichtweise zu erhalten. Die Gespräche wurden aufgenommen und werden nun nach und nach unter anderem in Schweizerdeutsch transkribiert und annotiert. Momentan ist eine Version des daraus entstandenen Korpus in der Grösse von 528'381 Token im XML-Format erhältlich[2]. Da die Zeitzeug_Innen aus der ganzen Schweiz kommen, wir das Datensetz vor allem für dialektologische Forschung des Schweizerdeutschen gebraucht. Ich fand es spannend das Sprachmodell auf Sprachdaten zu trainieren, welche in unterschiedlichen Dialekten verfasst wurden und dann zu sehen wie gut bei daraus generierten Text, die Vielfältigkeit der einzelnen Dialekte noch erkennbar ist. Oder vielleicht kann man so sogar das Problem des nicht vorhandenen Standarddialekts des Schweizerdeutschen lösen?;)

Pre-Processing

Da die einzelnen Aufnahmen des Korpus im XML-Format sind, musste ich zuerst den Rohtext, also die in Schweizerdeutsche transkribierten Token, aus den Files extrahieren (s. Skript `pre-process.py`). Die Transkriptionen sind schon annotiert, das heisst um Tokenisierung und Satztrennung musste ich mich nicht kümmern. Das daraus entstandene Textfile enthielt noch einige leere Zeile, was beim Trainieren zu eine DivisionByZero Error führte und darum gibt es im Repository noch ein kleines Skript, um nachträglich New Lines zu entfernen (s. Skript `del_empty_line.py`).

Performance vor & nach Modifikation Romanesco

	Romanesco ohne Modifikation:	Romanesco mit Modifikation:
Perplexity on dev set:	171.66	251.47

Hyperparameter:

```
VOCAB_SIZE = 2000
```

Ich habe die Vokabulargrösse vergrössert, da meine Daten aus 41'607 Types besteht. Somit werden die Trainingsdaten besser erfasst.

```
BATCH_SIZE = 32
```

Die Batch Size hab ich verkleinert, da die Sätze in gesprochener Sprache eher kurz sind und darum die Token von nicht so vielen vorherigen Token abhängen; es muss weniger vorherige Timesteps einen Hidden State beeinflussen.

```
NUM_EPOCHS = 13
```

Die Epochen haben ich leicht erhöht, weil es beim Ausprobieren mit den restlichen Settings bessere Werte gab.

Denn Rest hab ich gleich gelassen, weil beim Ausprobieren keinerlei Anpassungen bei denen einen positiven Einfluss aufs Ergebnis hatte.

[1] "Verfahren Um Jüdische Vermögen Bei Schweizer Banken." Wikipedia, de.wikipedia.org/wiki/Verfahren_um_j%C3%BCdische_Verm%C3%B6gen_bei_Schweizer_Banken.
[2] https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html

```
NUM_STEPS = 100

LEARNING_RATE = 0.001

HIDDEN_SIZE = 1024

EMBEDDING_SIZE = 256
```

Architektur:

Die LSTM-Zelle hab ich durch eine GRU Zelle ersetzt, da die Performance oft im ähnlichen Bereich ist wie LSTM, aber weniger rechenintensiv. Das Training war jedoch weder kürzer noch war die Perplexität kleiner.

Allgemein haben jegliche Veränderungen – ich habe vor allem mit dem Hyperparametern gespielt – keinerlei Verbesserungen in Perplexität erzielt. Ich gebe jetzt einfach die Ergebnisse ab, wo ich trotz Veränderungen noch die niedrigste Perplexität erhalten haben.