MT Übung 4

Thema: RNNs

Wir haben uns für dieses Datenset (Shakespeare) entschieden, weil Shakespeares Werke einerseits „altes“ Englisch, sowie andererseits „gleichgebliebenes“ Englisch beinhalten. Ausserdem beinhalten Shakespeares Werke so viele Synonyme, dass manche sogar behaupten, dass diese Werke nicht nur von einer Person stammen konnten. Vokabulargrösse: 10000. Wir haben ein Datenset ausgesucht, welches mehr als die Mindestvokabulargrösse (über 1MB Text) beinhaltet, jedoch nicht allzu viel, damit der Server nicht allzu lange für seine Berechnungen benötigt.

Anschliessend haben wir das Datenset durchmischt und im Anschluss in 1/10 als Dev-Set und der Rest als Trainings-Set unterteilt. Einige Datensets hätten uns auch sehr interessiert, wie beispielsweise Kundenmeinungen über Unfälle und medizinische Unterstützung, welche jedoch nicht zugänglich waren. Einige Links gingen nicht und einige Datensets brauchten viele Pre-Processing-Schritte. Da Romanesco keine Pre-Processing Schritte durchführt haben wir als Pre-Processing-Schritt auf alle Satzzeichen verzichtet, daher ist die Variante ohne Pre-Processing etwas eleganter, da sämtliche Satzzeichen noch da stehen.

Beim Code-Verändern war es schwierig herauszufinden, wo etwas fehlt, was besser, bzw. im Kontext passender sein würde. Wir haben uns Mühe gegeben, viele verschiedene Hyperparameter z.B –e, -b zu benutzen, um die tiefere Perplexität-Werte zu erreichen. Das Training war zeitaufwendig. Wir haben ausserdem noch eine zweite Hidden-Layer hinzugefügt, welches den Perplexity-Wert gesenkt hat, jedoch zeitlich länger zur Berechnung benötigte. Aber man hatte immer einen guten Überblick mit den GPU’s, wo noch Platz auf dem Server ist.

Als Perplexität, also wie überrascht unser System auf dem Dev-Set war, haben wir folgende Perplexity-Werte erhalten:

epoch(4):406 epoch(3)1:409 epoch(3)2:261 epoch(3)3:221

- without pre-processing:perplexity:136
- with pre-processing(batch size 40):perplexity:221

```
te capability: 3.7)
2018-05-02 13:52:05,556 - INFO - Restoring parameters from model/model
hawhat autolycus than words to go by the excellent father
devil secure together though yet far himself nor long
but in the bull function of
to hima addressd as passing and to the other power
dat for the wood that is the beds in the hound
by my people may oer to use it so hoarse away the body
amongst even one sirwhy lawful voice but thought by
in your his spirit of bed grace fair strain for hundred requires
in proud passions softly swift sway
```

Mit Pre-Processing war die Perplexity höher, da wir die Satzzeichen weggelassen haben. Normalerweise wäre die Perplexity jedoch tiefer, wenn anschliessend noch ein Post-Processing stattfinden würde. Bei Shakespeare liegt die Perplexity ziemlich hoch, da es ein anspruchsvoller Text (und ausserdem nicht aus dem heutigen Jahrhundert stammt) ist.

```
2018-05-02 15:30:31,889 - INFO - Restoring parameters from model/model
whereof he so he must such fool
From all to shepherd's this dances make ring
That misery, with a natural wounds,
<eos> KENT
To is the air I hope, thou think thou fit
From my point and the face; unto him
<eos> NERISSA
I vex thee home to do my name,
I have a soul of it; for will how I we
man you to found to roof something clothes
In the king speaks,
By us with reason seeking the face
That sooner his letter?
KENT <eos>
```

Beim Sampling haben wir neuen Text generieren lassen. Die neuen Sätze die herausgekommen sind, sehen lustig aus. Es gibt Wörter die von der Reihenfolge nacheinander passen und andere wiederum die gar keinen Sinn machen. Uns sind die vielen „you“ also:

„thou“ oder „thee“ („you“) und „thy“ oder „thine“ („your“) und „ye“ (Plural „you“)

aufgefallen. Diese wurden beim Text generieren lassen nicht richtig verwendet: „…in your his spirit…“ oder „…thou think thou fit…“ sowie „…man you to found…“. Allgemein wurden die Personalpronomen nicht richtig angewendet: „…how I we man you to found…“ ist ein katastrophaler Satz. Beim Pre-Processing (im Gegensatz zu ohne Pre-Processing) wurden einige Wörter zusammengeschweisst : „hawhat“ statt „ha what“, “„hima“ statt „him a“, „sirwhy“ statt „sir why“. Dies liegt wahrscheinlich daran, dass die Satzzeichen weggelassen wurden und somit jeweils das letzte Wort und das erste Wort eines Satzes zusammengesetzt wurden. Da unser Text auf Englisch und nicht auf Deutsch geschrieben ist, ist es weniger schlimm (somit werden das „I“, Eigennamen und Nationalitäten ausnahmsweise klein geschrieben), dass alles klein geschrieben (lowercase) ist.

```
2018-05-02 16:15:53,653 - INFO - Restoring parameters from model/m
manage
than thee to she doubled stand into his good talbot
not thine shadows son as simonides leaves away
and where the free your all finger thunder cries
brought and scattred and to dwell
when i by the fair secrets to lords
earl of us to your picture ere thus youlook compound be
i know have repair of me itself knows
in emperor with such them fondly should say to
to this respect of my self this good majesty thou art no sad man
<eos> sooner raise and advise us hope as like her cousins
duke i cannot kneel of her
that shee be attendants_ of all this
and me my hope i got
and hand upon from companions time
and even to my deeds calls all my judge
<eos> whereat his grace shall dares us into my strife
<eos> as he shall look towards eldest hell
nor in this duke doth never lady in to made
and trophies that valiant and heed as according
leonatos seald in this stern beyond project
<eos> yet cares in her hastings sir
```

Mit 2 Hidden Layer:

```
know not enter such afraid of dishonour
<eos> and you was for a waves he shall become
what beyond mei comes a soldier that
where let them detested friends a ponderous and lost gifts
pray him then such observing high telling thee
i are the great moans hot flavius presence
eats three
advise to pay and make the passion who feel
must i left him in a beggar
the sway of these door hath guest that my husbands house
a cheer blow we poor blessd the hare
lead the one rich one world itself
off tis all the sort of thy weight
grew she fair injuries of worse lake
she change that i thank each gloucester
not when it takes the unjust coal and fair
enter penny in himwhat is a ducat landed
i will gain with a prince of alms
scene bosom and future
yet been my honour lead it made of order itwhat afterwards appeard
thats their old coz your course a fathers days
as gives at a deed and interrupt her
most great hot be falling my night and fair good happy is sharp a wrong <eos>
```

Nach den Veränderungen im Code haben wir das Datenset ohne Preprocessing trainiert und wieder einen neuen Text generieren lassen. Der Text ist viel fliessender und klingt „menschlicher“. Grammatikalisch ist alles „richtiger“, ausser beispielsweise, dass „I“ („ich“) bleibt klein geschrieben und einiges andere. Die Abstände stimmen jedoch, d.h. dieses Mal wurden keine Wörter (jeweils das letzte Wort mit dem ersten Wort eines Satzes) zusammengeklebt. Als Perplexität haben wir 194(batchsize:40 epoch:4) erhalten. Zuletzt haben wir das Datenset mit Batchsize:30 Epoch:8 trainiert. Als Perplexität haben wir überraschend 76 erhalten. Davon gehen wir aus, dass die Perplexität viel tiefer geworden ist und der neu generierter Text etwas genauer und einlesbar ist.

```
<eos> PAGE
Well, served away you? Almost of you? Leave
<eos> VALENTINE
And for there's the friend
<eos> EVANS
Would he break within
[Exit Timon, and the KING hath to court with it
people here
<eos> FLUELLEN
Which if no rest, for you look on forty days
<eos> CADE
I'll draw forth what Helen is a infirmity,
WILLIAM
And in the hanging of a years, I yet
If we invest and th' number concluded
Do slay their door
Exit VALENTINE
She will allow the Moor upon my night,
And with your ladyship I come into th' flood
<eos> A sweet lady’s son, shows me in health,
The devil assist me with a lies,
Doth now do brought the patient years,
Think away the servant off,
My gentleman is love,
```

Zum Spaß haben wir uns für das zweite Datenset (Starwars) entschieden, um zu beobachten, ob ein neuer interessanter Text im Starwars-Stil nach dem Training generiert werden kann. Das Training war

nicht so zeitaufwendig. Beim diesem Datenset funktionierte es ziemlich schnell, denn es liegt an der Datenmenge. Beim Training haben wir die Batchsize als 1 und Epoch als 10 fixiert. Als Perplexität, also wie überrascht unser System auf dem Dev-Set war, haben wir folgende Perplexity-Werte erhalten:

- without pre-processing:perplexity:27

```
<eos>
local assortment much going to kill
interest, and more with Red arms. Han are way careful up.
<eos> Blast his downward stay
stormtrooper.
<eos> when an uniform. BAY about?
<eos>
<eos> the TIE little droid creatures sweeps in a screen, of
<eos> Luke looks at his enemy side of a loud robot has
INT. SURFACE OF THE DEATH STAR
All calmly turns for his Imperial computer little
rocky moves that moves of laser air, belt.
TARKIN
good on the Imperial checking on stroke
Vader climbs out in the time.
troopers turns a green chips.
the Death Star not is stroke.
<eos>
life the surface. A princess is Dutch.
```

Beim Sampling haben wir neuen Text generieren lassen. Die folgenden Wörter sind bei uns aufgefallen wie erwartet:

„uniform“, „princess“, „Imperial“, „Han“, „Vader“, „Death“, „robot“, „Luke“, „computer“

Kreative Sätze sind bei uns auch besonders aufgefallen. Es überrachste uns, dass nicht nur die Eigennamen richtig erkannt und korrekt generiert wurden, sondern auch manche Sätze in guter Struktur formuliert wurden wie die folgenden Sätze:

„A princess is Dutch“, „Luke looks at his enemy side of a loud robot“, „Vader climbs out in the time“.

Leider uns ist nicht gelungen, unser Sprachmodell- Schema auf Tensorboard visualisieren zu lassen. Deshalb haben wir das folgende Schema gezeichnet.