Computational Linguistics Final Project WS19/20: Skipgram Model results

Johannes Bernhard, 2544368

March 22, 2020

Contents

1 Introduction (2/2)

This project submission is split into two parts, a PDF result report and a Jupyter notebook. This PDF only contains the final results. The Jupyter notebook contains the code documentation and code needed to run the model. The notebook can be seen as part 1 and the PDF as part 2.
The submission comes with a PDF result report, Jupyter notebook, "Data" folder, "checkpoints" folder and "backup" folder. Please refer to the notebook for an exact description of how to obtain the results in this report and how to run the code.
The start of the submission is in the Jupyter notebook.

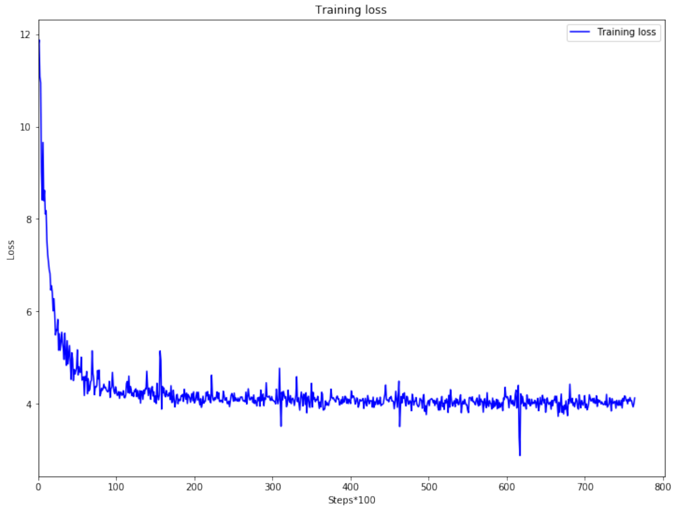
2 Results

The final model trained for 5 epochs, with 15300 training steps per epoch. Training time was 5.7 hours with around 0.273 seconds per batch. The final loss was 4.1269.

2.1 Loss

Figure 1 shows the training loss graph of the final model.
As we can see, the first training steps have the highest loss. This is obvious if one considers the fact that we initialized the matrices randomly and the model is starting from scratch. During this time, the loss lies in the range of 10-12.
Immediately afterwards, the model learns very quickly with the loss value dropping significantly after the first few iterations. The curve starts to even out around the loss values of 4-5. In all subsequent steps, the loss will continue to hover around this range. Sometimes the value drops or rises sharply only to even out again. In general, there is a lot of up and down in this range and no real progress is being made anymore. It seems that the model has already learned all it can after about 30000 to 40000 steps (x values need to be taken times 100). The curve still does have a very minor decline, but one has to ask the question whether any further training is reasonable, considering the large imbalance between computation time needed and minuscule loss drops. I think the answer is no.
We have to take into account that this model is using very generally chosen hyper-parameters for batch size, embedding size, negative samples and optimizer learning rate etc.. Perhaps fine-tuning these would give a better loss curve. In addition, the training corpus is not that large when

Figure 1: Loss graph of the Skipgram model

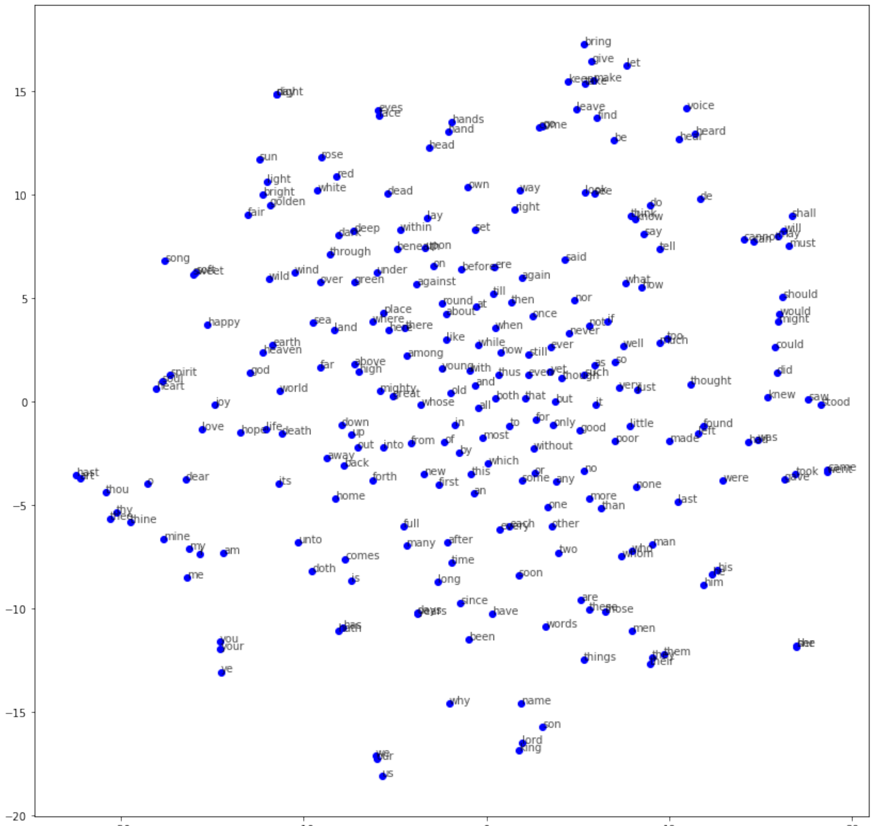


compared to the corpora other word2vec models have been trained on (783M words in the original word2vec paper). This further limits the model's learning potential.

2.2 Two-dimensional Embeddings

Figure 2 is a plot of the 250 most frequent words in the embedding space and how they relate to each other based on their distances. This is a way of quickly and easily verifying that the model has actually learned something. If we take a closer look at the given plot, we can see that this is the case.

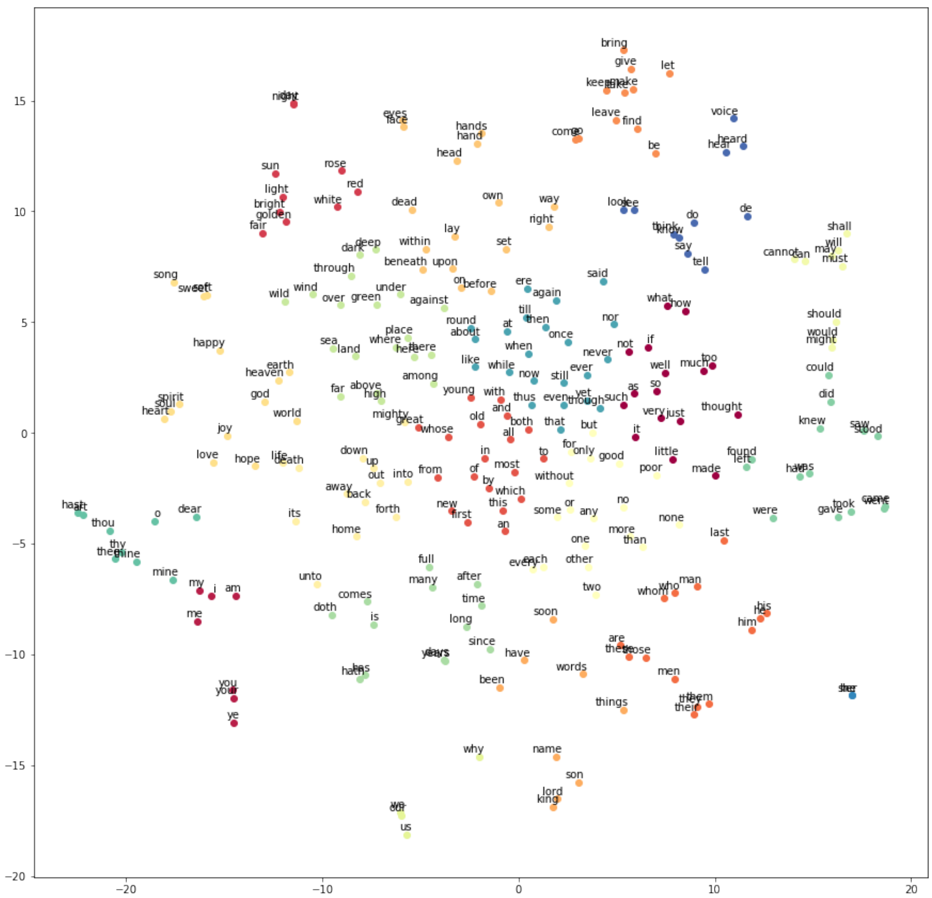
Figure 2: Basic embedding visualization in two dimensions.



In the bottom left, we can see a range of pronouns clustered closely together. "You", "your", "ye" and "mine", "my", "i", "me", etc. are all related in that they are all personal or possessive pronouns. Interesting to note in this cluster and for many words in general, is that there are also many older word forms present in the training data. "Ye", "mine" and "thine" are examples of this. We can see that there was no problem to correctly group them with their "newer" counterparts.
On the left, slightly above the pronoun section, we see "soul", "spirit" and "heart" which are all connected in their more abstract philosophical/esoteric meanings.
In the middle at the top, we see a body section, with "hands", "head", "eyes" and "face" occurring close to one another. Directly left of these, we see clusters for "light" with frequently accompanying nouns and adjectival descriptors as well as "rose" with its trademark colors "white" and "red".
Finally, another interesting bunch of words is on the far right, where we find "should", "would", "could", "shall" etc.. All of these are modal verbs.
As we move closer to the center, it becomes more difficult to distinguish certain word groups. Let us therefore move on to a clearer visualization of these word clusters in the form of the K-Means++ plot.
Figure 3 shows us the same plot as in figure 2 only with coloured clusters determined via K-Means++.
Immediately, we can see that the perceived clusters from the previous figure are more or less accurate. The personal pronouns in the bottom left have been split between the lines of variations of "they" and "me"+"you". The "light" and "rose" bunch are counted as their own cluster and the body cluster has been expanded with complementing words that fit very well to the initial words. Examples that come to mind are "dead eyes", "lay hands upon" or "set eyes (on)".

Moving closer to the center, we can identify a time-themed cluster in light blue just below the

Figure 3: Clustered embedding visualization in two dimensions.



orange body cluster. Here, we see an over abundance of time related words, such as "at", "till", "never", "while" or "now". Directly next to the time cluster, we can see a place-themed cluster in

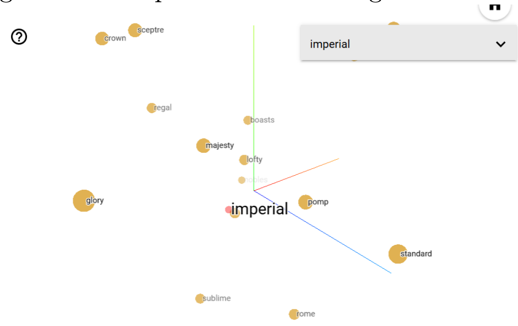
light green with "far", "high", "here" or "there".
We can conclude that the model has learned many interesting word relationships. Now is a good time to move on the embedding projector and have a look at some words that might occur more frequently in poetry texts or just didn't appear in the 250 most frequent words.

2.3 Embedding Projections

Table 1 shows some example terms and a selection of related words. These related terms were taken from the embedding space by choosing 6 of the top-15 nearest neighbors of an example term. An easy way to see these in the projector is to find the term via the search bar, limit the neighbors to 15 and then click on "isolate 16 points" (example term and 15 neighbors). An example picture is provided in figure 4.

Example terms were inspired by a web-page called "Top poetry words" [1] and words that I found interesting while navigating the embedding space. There are a total of 14 example terms, just to drive home the point that the embeddings are meaningful for multiple different words.
In general, there are many interesting word relations to discover in the embedding space and it was quite easy to get lost in the cloud by just clicking through from data point to data point via the nearest neighbors list. The example term choice is intended to highlight the fact that the model has learned various word relationships. As such, related words that might be more interesting than others are highlighted in bold and will be discussed in a bit more detail.

Figure 4: Example of an embedding visualization.



Looking at table 1, we can see that many immediate neighbors are related to the example terms in meaning, concept and context.
For "air", we can observe many words that may "lie in the air" such as "music", "melody" or "fragrance". Also, "air", "sky" and "breeze" are linked via their connections to nature.
Next, "body" is closely connected to the soul, especially in more artistic texts where one can often read about the connection between body and soul. "Curse"'s neighbors are related to revenge which is not surprising when one considers that a person is most likely to curse someone else if they are upset with them and want them to suffer, the very definition of revenge. In this context, "doom" is a very telling indicator as to what a curse entails.
Moving on, "damsel" is closely connected to "knight" which, considering some of the older texts in the corpus and the wide-spread damsel in distress setting of many medieval stories, is not very surprising. A close word to "eyes" is "countenance" which is interesting as it identifies the eyes as being a crucial part of the facial expression. The many different adjectives in close vicinity to "fragrant" point to the varied word use of the different poets. The same goes for "gorgeous", where all the related words can be seen as synonyms.
Continuing, in posh and noble settings a hand is to be kissed in order to show one's subservience or respect, thus the connection between "hand" and "kiss". Empires never see themselves as anything less than great and glorious. Thus, it is only natural that words like "glory" and "illustrious" appear, with "crown", "majesty" and "nobles" pointing to the governing structure of most historic empires. Interesting is also the mention of Rome, further reinforcing its position as one of the most influential empires of all time. The connection between "juice" and "poison" is quite amusing and the presence of other liquids ("wine", "nectar") as well as words related to the action of drinking ("drink", "cup") show that the model was able to build a solid relationship between all of these words.
Furthermore, the nearest neighbors of the last few words are also populated with many meaning and concept related words such as "king" with "monarch", "love" with "passion" and "hope", "zeal" with "ardour" and "purpose" (zealous purpose) or "nymphs" with "virgins" and "beauteous". Especially the expanded range of neighbors for "nymphs" points to their mythological

Term	Example 1	Example 2	Example 3	Example 4	Example 5	Example 6
air	music	melody	breath	sky	breeze	fragrance
body	**soul**	heart	flesh	limbs	bruised	blood
curse	vengeance	**doom**	wrath	scorn	murder	death
damsel	maiden	queen	fairest	lovely	**knight**	mistress
eyes	lids	features	**countenance**	sight	blinded	glance
fragrant	**ambrosial**	delicious	blossoms	dewy	odorous	roses
gorgeous	splendid	emerald	radiant	resplendent	fair	dazzling
hand	weapon	fingers	arm	clasp	touch	**kiss**
imperial	Rome	glory	**illustrious**	nobles	crown	majesty
juice	**poison**	wine	cup	drink	oil	nectar
king	prince	**monarch**	valiant	queen	warrior	baron
love	passion	joy	**hope**	desire	soul	tender
mysterious	trance	spell	veiled	silence	universe	**infinite**
nymphs	virgins	maidens	sylvan	goddess	beauteous	groves
zeal	pride	ardour	ambition	manly	service	purpose

Table 1: Selected words and their nearest neighbors

origin and traits. The connection between "mysterious" and "infinite" is amusing, pointing to the problem that many people seem to have with grasping infinity. The other related words for "mysterious" are also very relevant, with "trance" and "spell" being more overt connections; trances are mysterious after all and spells aren't natural; and "silence" being a more abstract connection but nonetheless plausible, especially considering a more poetic environment.

Concluding, we can see that the model has learned some interesting relations and that the implementation was successful. The most refined results seem to have been achieved for nouns, verbs and adjectives that are generally more frequent. Many words display interesting connections to their nearest neighbors. Such connections might be synonyms, hypernyms and related concepts, explicit and abstract. However, some remarks have to be made.

First, while many of the higher frequency words have very plausible related words, many lower frequency words do not have meaning related words in their nearest neighbors. This is simply due to the lack of sufficient training data and thus lack of different contexts.

Second, in a similar vein with frequency, the embedding space is extremely dense in the center. Many of these data points represent lower frequency words which do not really contribute to the learned relationships since they do not occur often enough to infer meaningful contexts. In order to improve the results, it would be best to increase the pre-processing absolute frequency threshold and implement the sub-sampling technique proposed by Mikolov et al. in a second paper on word2vec [2].

Third, the hyper-parameters were not tuned at all and chosen as simple rough estimates. Increasing the embedding size, negative samples and window size might yield more refined results even for lower frequency words.

References

[1] Robert Peake. *Top Poetry Texts.* https://www.robertpeake.com/archives/6676-top-poetry-words.html. 2014.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. *Distributed Representations of Words and Phrases and their Compositionality.* https://arxiv.org/abs/1310.4546. 2013.