دانشگاه علم و صنعت ایران

Department of Computer Engineering

Natural Language Processing

Final Project

Final Phase[1]

Alireza Moradi
Student No.: 96521479

Prof. Sauleh Eetemadi
Summer 2021

[1] https://github.com/Alireza1044/cs224n-final-project

Contents

List of Figures

List of Tables

1 Word2vec

The results in figure 1 and 2 are for 25 of common words between the two classes. The loss at the end of training was around 5.5 . The word *scranton* has a similar vector in both classes which makes sense because their work place is located at Scranton, PA and this word is usually used in the same context with pretty similar neighboring words. The word *hello* has two completely different vectors, and that is justifiable because one of the characters has a sarcastic behavior and several times during the series he uses this word in a sarcastic manner:)

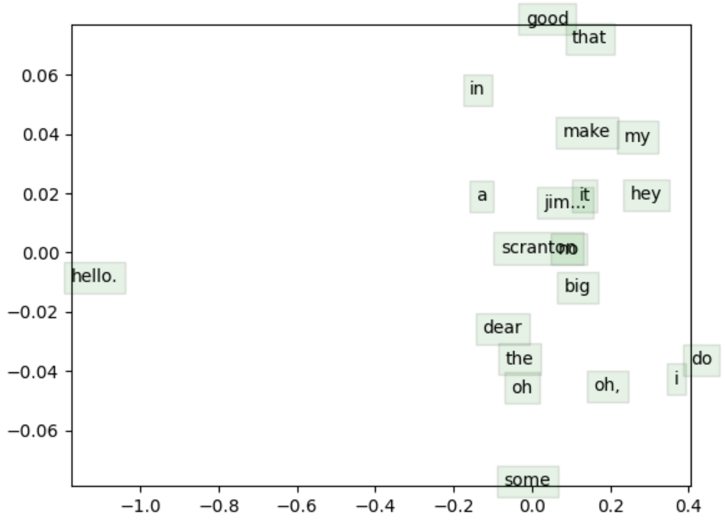

Figure 1: First class word vectors

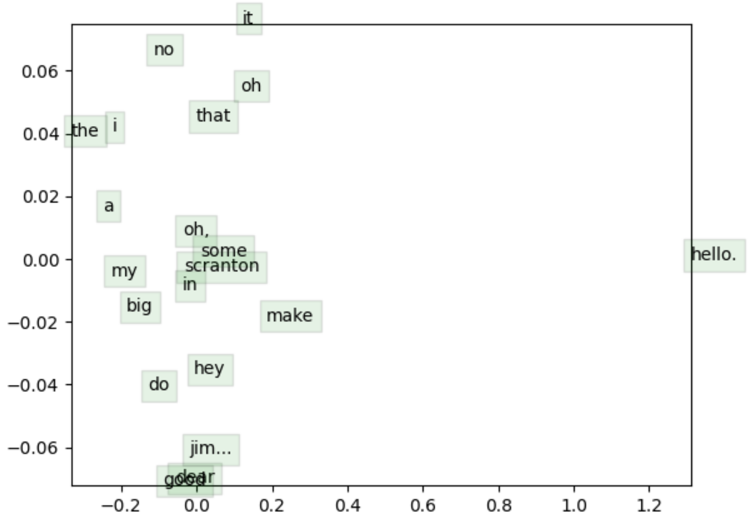

Figure 2: Second class word vectors

2 Tokenization

Vocab Size	Iteration	<UNK> %	Average
50	1	0.046	0.045
	2	0.047	
	3	0.028	
	4	0.048	
	5	0.055	
150	1	0.076	0.070
	2	0.074	
	3	0.078	
	4	0.078	
	5	0.047	
500	1	0.073	0.097
	2	0.0110	
	3	0.074	
	4	0.12	
	5	0.11	
1000	1	0.123	0.128
	2	0.124	
	3	0.143	
	4	0.143	
	5	0.11	
1500	1	0.110	0.130
	2	0.150	
	3	0.12	
	4	0.14	
	5	0.13	

Table 1: <UNK> token percentage grouped by vocab size

As it can be seen from the data in table 1, using vocab size 50 produces the best results.

3 Parsing

The model was trained on the original data for assignment 3. Below are 10 sentences from my own dataset which were parsed by the model. The UAS score on these samples was 86.81%. The model performs well (UAS 100%) on short sentences or sentences that are only a single part (e.g. have only one verb and no , in the middle of the sentence). On long sentences or complex sentences, the model struggles. I believe this problem can be fixed with more training data.

1. **your quarterlies look very good.**
   [(2, 1), (3, 2), (5, 4), (3, 5), (3, 6), (0, 3)]

2. **how are things at the library?**
   [(3, 2), (3, 1), (6, 5), (6, 4), (3, 6), (3, 7), (0, 3)]

3. **you are a gentleman and a scholar.**
   [(4, 3), (4, 2), (4, 1), (4, 5), (7, 6), (4, 7), (4, 8), (0, 4)]

4. **so this is my kingdom, as far as the eye can see.**
   [(5, 4), (5, 3), (5, 2), (5, 1), (5, 6), (8, 7), (11, 10), (13, 12), (13, 11), (13, 9), (8, 13), (5, 8), (5, 14), (0, 5)]

5. **you are going to be my accomplice.**
   [(3, 2), (3, 1), (7, 6), (7, 5), (7, 4), (3, 7), (3, 8), (0, 3)]

6. **i think the conference room should be fine.**
   [(2, 1), (5, 4), (5, 3), (8, 7), (8, 6), (8, 5), (2, 8), (2, 9), (0, 2)]

7. **i slashed benefits to the bone.**
   [(2, 1), (6, 5), (6, 4), (3, 6), (2, 3), (2, 7), (0, 2)]

8. **the employees went crazy, i got no help from corporate.**
   [(2, 1), (3, 2), (3, 4), (7, 6), (7, 5), (7, 3), (9, 8), (11, 10), (9, 11), (7, 9), (7, 12), (0, 7)]

9. **if you do not raise your hand, it will not be covered.**
   [(5, 4), (5, 3), (5, 2), (5, 1), (7, 6), (5, 7), (13, 12), (13, 11), (13, 10), (13, 9), (13, 8), (13, 5), (13, 14), (0, 13)]

10. **you have forfeited that privilege.**
    [(3, 2), (3, 1), (5, 4), (3, 5), (3, 6), (0, 3)]

4 Language Model

Michael	the office examination, perfect try. is, come breakfast. oh, what is the
	corporate is stay very acupuncture, minority. it is not not to start
	they want us to youtube? john copy, delivery for win. here you manage. did
Dwight	the office is string with between hunter. i am an little traditional cake
	corporate is way! instructions that is michael! enough i should sailed. coolio.
	they want us to once, of my bullet oh, oscar chop you smell him

Table 2: Generated text samples

The model is a simple LSTM Language Model which was trained for 5 epochs for each of the classes. As it can be seen, sentences generated by the model don't make sense and most of them are also grammatically incorrect (pre-training on a large corpus would be helpful to overcome these problems). On the other hand, generated sentences show some characteristics of the characters which indicates that [much] more training data can improve this.

5 Fine-Tuning

Michael	i am very happy!pam has a good job.i know.
	i wanted to sit in that bathroom because she was kind of a jerk.
	warehouse has a great, fantastic, very nice office.
Dwight	i have to be on vacation with my friend in order to get some cookies.
	i am assistant regional manager.no.no, you are the scranton branch manager.
	my farm is an island of passion, fun and hope.

Table 3: Generated text samples

Generated sentences after fine-tuning are in most cases grammatically correct. They also represent a better characteristics of the speaker. This is because we have a much more complex network (GPT-2) which was also pre-trained on a huge corpus. Despite producing better outputs, we still need more data to have satisfactory outputs in the end.