Kennedy Ellison

Speech Synthesis and Recognition

Professor Jane Chandlee

December 21, 2018

Dr. Phil Language Model

I. Summary of Project

The goal of the project was to train a bigram language model on tweets of Dr. Phil to generate sentences. Dr. Phil is known by the majority of Americans for his TV show based on psychology so a survey was used to gage the effectiveness of the model by comparing actual tweets of Dr. Phil to the ones generated by the language model. It was supposed that the domain that Dr. Phil talks about would lead to some interesting sentences and would make for an enjoyable project. One of the main motivations for the project was to learn how to use the twitter APIs. In most courses, data is given to the students, so this was an opportunity to learn how to collect social media data and then use it in an interesting way.

The majority of the time spent on the project was on the data collection because it was unfamiliar. After getting Twitter developer access, there was a learning curve to get the tweets. Alex Kras had a Twitter api tutorial that was helpful in learning how to get a user's tweets.[1] The next challenge was finding a way to get the python twitter modules to interact my computer. There were issues with how the modules interacted with my computer that required a lot of debugging and manually changing certain parts of the Twitter module code. Another challenge in this project was the tweet limits. The api

[1] https://www.alexkras.com/how-to-get-user-feed-with-twitter-api-and-python/

function that was used (GetUserTimeline) only grabs 200 tweets at a time. I worked around this by using a loop to keep grabbing tweets until I hit the 3200 tweet limit.

The last major step was to try to improve the language model. One of the first things I did was try to create a regular expression to remove urls from the sentences. Every tweet of Dr. Phil has a url to the story on his site which means the earliest tweets that my program generated contained a lot of urls. Once the urls were removed, I created a function to find all of the words that Dr. Phil uses to begin sentences. The goal of this was to make sure that the generated sentences begin with a reasonable word. Otherwise the program would generate sentences that began with “him” or something else ungrammatical. These additional steps improved the intelligibility of the generated sentences. After this, I created a survey to evaluate my sentences.

II. Data: The data is attached in a file called “tweets.txt”.

III. Modules Used

1. twitter module: python module to interact with the twitter APIs
2. re: used for creating regular expressions for processing the data
3. random: used in generating sentences by picking a random bigram
4. nltk: used the sentence tokenizer to determine sentence boundaries to find what words begin sentences

IV. Evaluation of Language Model

The evaluation consisted of ten sentences, four sentences tweeted by Dr. Phil and six generated by my language model. These sentences are displayed in Table 1. Ten people evaluated the language model and they were all familiar with who Dr. Phil is and what he does. This way they knew what kind of sentences to expect from the

survey. Tweets by Dr. Phil were included to provide a standard/baseline for what the users thought of his actual tweets. The evaluators were asked to rate each sentence from one to five, with one being a computer made sentence and five being a sentence that Dr. Phil tweeted. Three was considered “neutral” meaning they had no idea if it was made by the language model or Dr. Phil himself.

Table 1. List of Sentences included in the survey.

Language Model Generated Sentences	Dr. Phil’s Actual Sentences
1. brittanie is driving	3. cassidy says over the years her mother melinda has been more of a friend to her than a parent
2. vehemently deny posting scary music wear	4. paige says that she and her girlfriend got kicked out of their apartment
5. i always open up things without his five drphil mark	9. how to stay married 42 years 🤫😂
6. tina claims bobbys 8yearold son	10. angie says her husband of 25 years butt dialed her thus revealing that he had a secret inheritance and he spent over $100,000 without her knowledge
7. dr freda lewishall discusses her journey with doctors warned lisa and mary	
8. risk factors of social situations	

The average rating for generated sentences compared to Dr. Phil’s actual sentences differed significantly. The expectation was that Dr. Phil’s sentences would be rated close to five and that the generated sentences would be closer to five if the model was good or closer to one if the model did not reflect Dr. Phil very well. Dr. Phil’s sentences were rated

4.23/5 which is pretty high, though not five! Sentence nine tripped people up as seen in Figure 1 which probably dropped the overall rating for Dr. Phil's sentences. The emojis might have seemed like something Dr. Phil does not use much, which could explain why three people did not think Dr. Phil tweeted sentence 9.

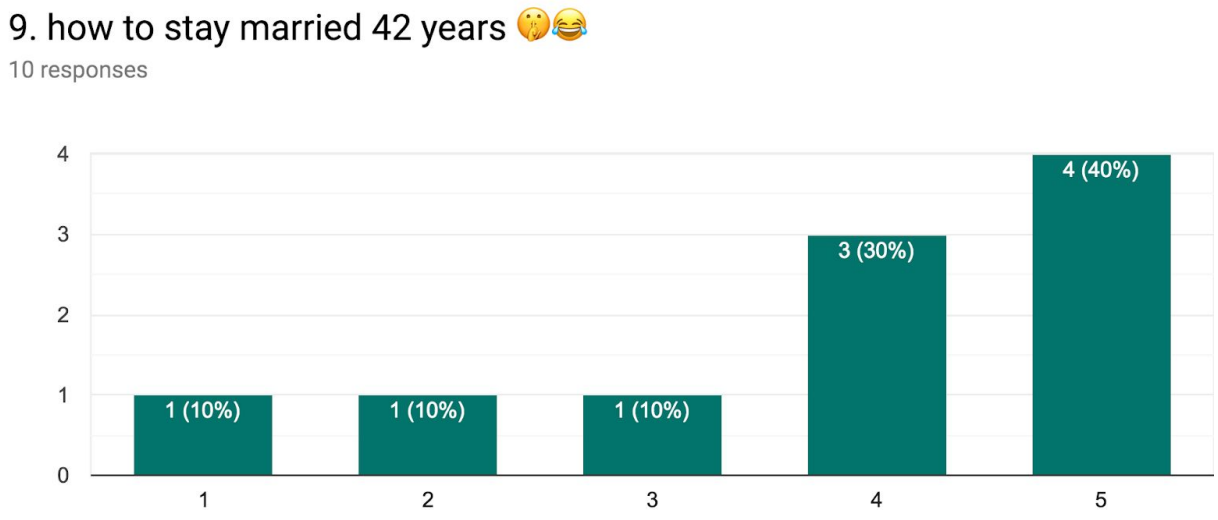

Figure 1. The lowest rated sentence of Dr. Phil.

The language model was rated 2.45/5 which is not very high. It indicates that there is a lot of room for improvement within the language model. The highest rated sentence generated by the language model was sentence eight as seen in Figure 2. It's average rating was 3/5. One reason for it being ranked higher than the others is that the content talks about social situations. Evaluators may have thought that it was a topic Dr. Phil would talk about frequently. It is also a short phrase that is grammatically correct. There was no ungrammaticality to signal to the evaluators that it was computer generated.

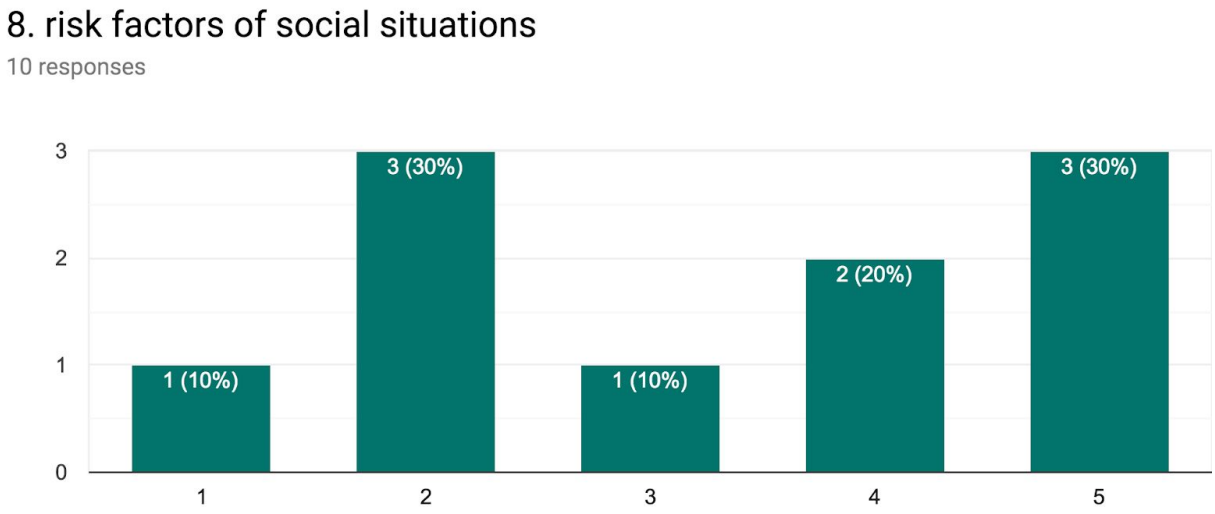

Figure 2. The ratings for sentence eight.

V. Conclusion

Though the language model was far from perfect, it generated some intelligible sentences and even made some sentences that evaluators thought were real sentences tweeted by Dr. Phil. However, there is a lot of room for improvement. The words at the end of sentences could be included in the decision making for generating a sentence like words at the beginning of sentences were. If more than 3200 tweets could be gathered, it could be possible to use trigrams or 4-grams to generate more realistic tweets. With only 3200 tweets, it is difficult to use trigrams or 4-grams because of the data sparsity. Smoothing could be incorporated to deal with the data sparsity.

Moving away from n-grams, other methods could be used in the artificial intelligence realm. Neural networks that generate sentences character by character have proven to be incredibly successful, but do require large amounts of data. This seems to

be a tough problem to solve for language models of specific people. There is only a certain amount of publicly available data to train these programs so dealing with data sparsity is very important. As my language model said, “standup from zero future.” I’d like to think it was trying to say something like “there’s nowhere to go but up”, so as to reflect this need to further explore language model improvement.