LSTM-based Conon O'Brien Twitter Bot

Fengdi Li
Georgetown University
520 Prospect St NW
Washington, DC 20057
fl418@georgetown.edu

Abstract

This paper introduces a prototype LSTM neural network model built on tweets data posted by a famous talk show host—Conon O'Brien, it can be used to generate tweets imitating the O'Brien's real tweets by given appropriate starting word(s). The paper also provides the detailed data preparation processes, methodologies, and the performance evaluation of the model.

Objective

Thanks to the continuous revolution of natural language processing and artificial intelligence techniques, smarter and smarter bots have been created for research and business purposes, including Apple Siri, Google Assistant, Amazon Alexa, etc. Donald Trump, the 45th and current President of the U.S., is famous for using Twitter (*@realDonaldTrump*) to attract attention around the world. Since President Trump has a unique language style, unlike typical politicians, mimicking Trump has become a favorite activity recently. Bradley Hayes, a researcher at MIT, successfully trained a Twitter bot that 'sounds scarily like him' using the neural network (Murphy, 2016).

So, can AI automatically create the human language that bring joy to our lives? This study investigated whether deep learning techniques can imitate the style of tweets posted by Conan O'Brien (*@ConanOBrien*), an American television host/comedian who is well known for hosting several late-night talk shows and actively posts interesting short text on Twitter.

Data Source

Twitter provides a user-friendly API, which is free for collecting a maximum of latest 3200 tweets posted, including retweets and replies, by a given twitter handle using a standard Twitter developer API key. O'Brien's Twitter account was registered in February 2010, which contains a total of 3626 tweets by November 15th, 2018. To avoid the writing style changing issue, in the data collection process, only the latest 2500 direct posts were kept. At the same time, retweets and replies were also removed from the training set. This collection of scrapped tweets content forms the foundation for this study.

Methodology

In human languages, words in a sentence are mutually connected, which means their meanings are derived from the overall context. Thus, it is crucial to producing real conversation based on previous words and sentences. A recurrent neural network (RNN), unlike a typical feedforward neural network, 'is a variant of a recursive artificial neural network in which connections between neurons make a directed cycle (Monsters, 2017).' It this way, the predicted target not only depends on the current inputs but also on the previous net weights and outputs.

To accomplish this, a special RNN architecture named Long Short-Term Memory (LSTM) is designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs (Monsters, 2017). It is one of the most popular deep learning models employed in natural language processing (Young et al., 2018).

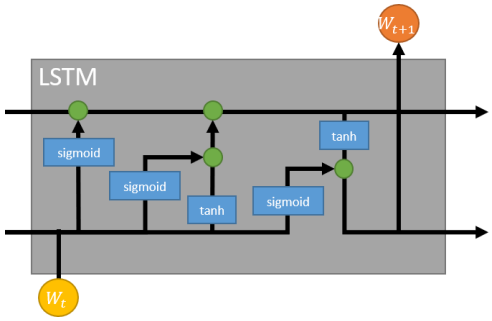

Figure 1. LSTM Unit

In this study, the LSTM-based model was trained on every tokenized N-gram in O'Brien's tweets, which were produced after lower-case conversion, punctuation encoding correction and removal except for '@' and '#', adding end-of-sentence marker, correcting unusual typographical error, and tokenization.

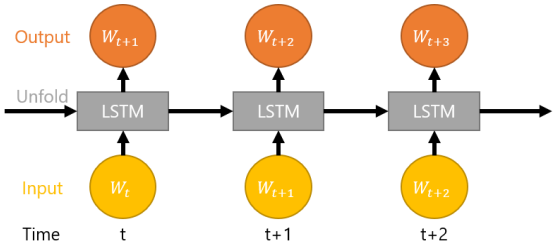

Figure 2. LSTM Model

The inputs of the model take sequences of first n words in each processed tweet at a time, where n = 1, 2, 3, …, N - 1, respectively, and N is the length of corresponding tweet, then added pre-padding to ensure each input has the same length as the maximum word length 86 in the dataset. While the outputs, known as label, are the next word in the tweet of the input N-gram (Bansal, 2018).

The model first embeds words in a 16-dimensional space, then calculates an optimized neural net, which has 150 LSTM units combined with a 10% recurrent state and a 10% input/output dropout ratio. And it will predict the best possible next word one by one.

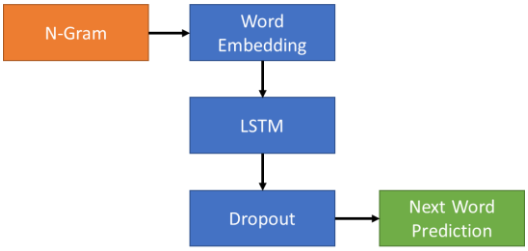

Figure 3. Model Structure

Next, a function was created using this model to generate a sentence with given start word(s) mimicking O'Brien's tweeting style, and the length of each result will be automatically defined by the model with the help of end-of-sentence marker.

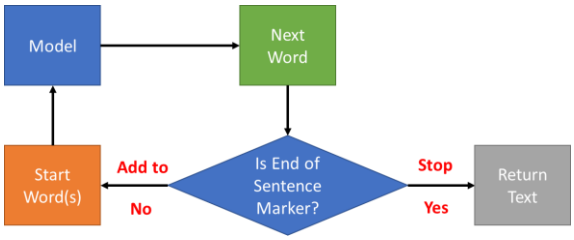

Figure 4. Word Generation Function

Result

Twenty pairs of common start words within the training vocabulary, including "Today", "I like", "How", etc., were given as the input to feed in the text generation function using the model, and the length of the generated sentence will be automatedly set based on the end of sentence token trained in the model. Some examples of the text result are "*This is the third time i think we re not sing by ikea*", "*Good luck tonight @lesdoggg #betawards*", and "*I ll be discussing it with color guard with extra tips as if you bring up discuss*" (see Appendix A for details of the tested start words and generated text pairs).

Evaluation

The training model has a comparably minimum cross entropy of 1.94 at the 60th epoch, reduced from 9.82 at the first epoch. After producing the collection of result sentences, two approaches were conducted to evaluate the performance of the model.

First, 'Readability' score. The score, ranged from 0 to 1, is based on the manually reading by the author to validate if the sentence is understandable by a human, or if the sentence can be split into a few shorter meaningful sub-sentences, instead of an entirely non-sense sentence (see Appendix A for details of the tested generated text and Readability score pairs).

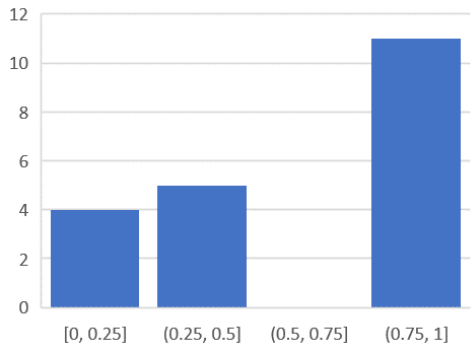

Figure 5. Histogram of Readability Scores

Second, 'SimOver85' ratio. Latent Dirichlet Allocation (LDA) is a probabilistic model that assigns each document a mixture of latent topics (Blei et al., 2003). Based on the LDA topic model, a similarity score can be calculated by comparing word and topic composition of two documents (Celikyilmaz et al., 2010). Implementing the topic modeling on original tweets data, and the proportion of tweets associated with a similarity score higher than 0.85 per generated tweet produced by the model is a reasonable metric measuring the performance of the model (see Appendix A for details of the tested generated text and SimOver85 score pairs).

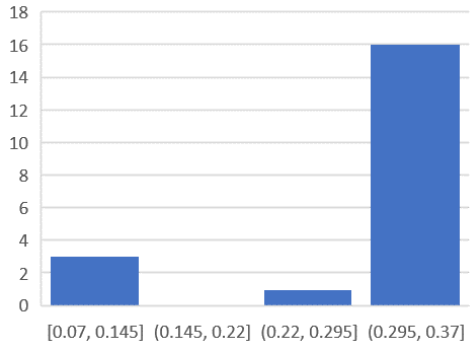

Figure 6. Histogram of SimOver85 Scores

The histogram graphs of the 'Readability' scores and 'SimOver85' scores represent the performance twenty generated text using the model created in this study. Depicted by the graphs, the majority of both evaluation values lie in the highest bar, which indicates the quality of the texts are passable.

Discussion

The prototype of LSTM-based model, built in this study, can be used to generate tweets imitating the real tweets posted by O'Brien. However, the performance is insufficient due to the limited amount of accessible training data, and the compromise on accuracy to reduce the computational expenses in building this prototype model, a small number of oldest tweets were excluded and the majority of the punctuations were removed, which makes the generated text harder to read and unable to interpreter sometimes. Thus, by selecting a Twitter account with more tweets data available to collect, and applying a better tokenization approach, the result will be more readable.

In addition, the readability scores in the evaluation section are very subjective. This problem can be solved by recruiting at least 20 randomly selected volunteers and defining the ultimate score as the average value of readability scores assigned by these volunteers.

Another issue with the tweet generation is that the original text can commonly be fragmented, mistyped, mixed with emojis and Internet slangs, or in poor grammar. Thus, a more proper way to evaluate the quality of the generated text should be adopted in this study.

Acknowledgements

I would like to thank Conon O'Brien for his interesting talk shows and tweets, which form the foundation of this study. I would also thank Dr. Daniel Loehr and Dr. Linda Van Guilder for the guidance of natural language processing theories and techniques.

References

Murphy, M. (2016, March 04). MIT built a Donald Trump AI Twitter bot that sounds scarily like him. Retrieved from https://qz.com/631497/mit-built-a-donald-trump-ai-twitter-bot-that-sounds-scarily-like-him/

Monsters, D. (2017, September 26). 7 types of Artificial Neural Networks for Natural Language Processing. Retrieved from https://medium.com/@datamonsters/artificial-neural-networks-for-natural-language-processing-part-1-64ca9ebfa3b2

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Computational Intelligence Magazine, 13(3), 55-75. doi:10.1109/mci.2018.2840738

Bansal, S. (2018, March 26). Language Modelling and Text Generation using LSTMs - Deep Learning for NLP. Retrieved from https://medium.com/@shivambansal36/language-modelling-text-generation-using-lstms-deep-learning-for-nlp-ed36b224b275

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Celikyilmaz, A., Hakkani-Tur, D., & Tur, G. (2010, June). LDA based similarity modeling for question answering. In Proceedings of the NAACL HLT 2010 Workshop on Semantic Search (pp. 1-9). Association for Computational Linguistics.

Appendix A. Table: Details of Experiment Result

Start Word(s)	Generated Text	Readability	SimOver85
Today	Today i had a better last video but i m taking a batman #conancon'	0.5	0.35
I like	I like a ladybug in the streets and the fbi chambray the sitting place me suddenly in vr who have to get your eyes	0	0.07
Just saw	Just saw the red band trailer for minions cute but a few red makes #conanjapan	1	0.29
This	This is the third time i think we re not sing by ikea	1	0.37
If I	If I die suddenly in order to avoid embarrassment to the brave men women who ve served my guests	1	0.30
My friend	My friend greg daniels and i produced #peopleofearth for tbs s larry the woman in fire	1	0.36
These	These i can t believe i didn t win the super bowl office pool i got the gold statues fracking the soil of western pa is lousy with em	0.5	0.36
It	It s a small comfort but with a caftan excerpted from a factory that life begins	0.5	0.35
The show	The show millions drawing are cool and mine now said if i can t announce that my lawsuit against panda express	0	0.31
Why I	Why I am a better actor than people who don t know what my fitbit is coming	1	0.35
Good	Good luck tonight @lesdoggg #betawards	1	0.31
How	How much wood would a woodchuck chuck if he won to totally committed to share it s	0.5	0.36
A	A new study claims coffee fights erectile dysfunction that explains starbucks new slogan who s going to be a giant skull	1	0.30
President	President trump has been consulting with the user s brain you re all cat brady #dblegends	0	0.31
Tonight	Tonight on #conan i play @callofduty #advancedwarfare watch and become erect	1	0.31
It s	It s a small comfort but with a caftan excerpted from a factory that life begins	0	0.35
Thanks	Thanks to the winner of the 2016 presidential election i hope nobody won t even the nomination of my kids then go wrong around having #conanatl	0.5	0.08
I m	I m not sure i m going to think i m chomping up some grape s grandmother	1	0.10
I ll	I ll be discussing it with color guard with extra tips as if you bring up discuss	1	0.31
Look	Look steve bannon is a good twin	1	0.31