F.R.I.E.N.D.S Dialogue Generation And Classification with GPT-2 and BERT

Luping(Rachel) Zhao
https://github.com/rachelzhaolp/F.R.I.E.N.D.S_dialogue_generator_and_classifier

Abstract

Transformer architectures perform well in text classification and surpassed human accuracy in this project. But the machine is inadequate to handle more creative tasks like TV script generation.

1 Introduction

F.R.I.E.N.D.S is an American television sitcom, created by David Crane and Marta Kauffman, which aired on NBC from 1994 to 2004. The show revolves around six friends, *Chandler, Joey, Monica, Phoebe, Rachel, and Ross*, in their 20s and 30s. With the new season, *Friends Reunion Special* coming in 2021, one interesting question to answer is, does the machine capable of writing reasonable (and funny) dialogues for the characters? This project finetuned GPT-2 with the script of all seasons and evaluated the results with a multi-class classification model with BERT.

2 Related Work

Natural Language Generation (NLG) has long been considered the most challenging task, and the application of deep neural network (NN) architectures out-performs the traditional machine learning algorithms like Markov Chain. However, the evaluation of NLG is also an unsolved problem.

2.1 Natural Language Generation Models

The Markov-based approaches were one of the first algorithms used for language generation. *Liang et al. (2009)* pairs data to text based on a sequential Markov process, and conducts a probabilistic generative model.

However, due to the development of hardware and amount of data, the neural network model domains the field of NLG in recent years. The most commonly used architectures include Recurrent neural network (RNN), Long short-term memory (LSTM), and Transformers.

The inherently complex sequential path from the previous unit to the current unit of RNN and LSTM limits the capacity of their memory and disabled them to parallelize.

Transformers are widely used in Sequence-to-Sequence tasks because of their Encoder-Decoder framework *(Sutskever, Vinyals, & Le,2014)*. The application of Attention and self-attention *(Vaswani, Ashish, et al, 2017)* enables the Transformers to achieve a better BLEU score than previous state-of-the-art models in Machine Translation (MT). Generative Pre-trained Transformer 2 (GPT-2) *(Radford, A., et al, 2019)* and Bidirectional Encoder Representations from Transformers (BERT) *(Devlin, J., et al, 2018)* used in this project are both Transformer-based models.

2.2 NLG Evaluation of NN models

BLEU, a precision focused metric, and Rouge, a recall focused metric, are the most popular evaluation metrics that are used to compare models in the NLG domain. Both of them, along with other metrics such as Latent Semantic Analysis (LSA) and Translation Edit Rate (TER) evaluate the model by comparing the generated text and reference text.

Another typical method is to measuring performance on some secondary tasks, such as classification and information retrieval.

3 Dataset

The Dataset is the scripts of all ten seasons of F.R.I.E.N.D.S, including the name of episodes, settings and, dialogues. After extracting lines with no less than 3 words of the six leading characters, 42,329 valid samples remain.

3.1 Exploratory Data Analysis

The data is balanced with Phoebe has the least lines (14.84%) and Ross has the most (17.71%)

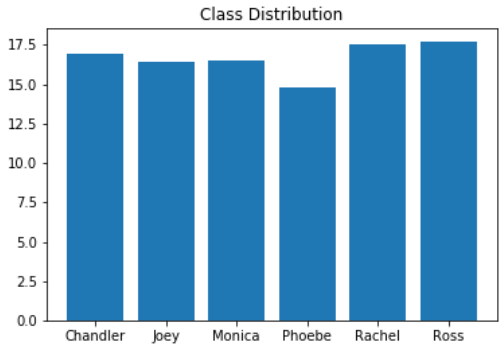


Figure 1: Class distribution

The length of each line is relatively short, with a median of 13.8 and a mean of 10.

Quantile	Length of the Line
10%	4
25%	6
50%	10
75%	17
90%	28
100%	248

Table 1: Length of the Lines

4 Methods

4.1 TF-IDF Analysis

TF-IDF is a statistic that measures how important a token is to a document in a collection of a corpus. To extract the most important word of the characters, I ran two TF-IDF experiments on the dataset.

In the first experiment, I concatenate the lines for each character and generated six documents, trying to illustrate what matters the most to them cross that ten years of their life, comparing with others.

In the second experiment, I grouped the lines by characters and seasons, expecting to show the growth of the friends.

4.2 Data Augmentation

The amount of the data is small, even for fine-tuning a pre-trained model. In this project, I used the *nlpaug* package in python to augment the original dataset. This library supports contextually insert/substitute words by feeding surrounding words to BERT to find out the most suitable ones for augmentation. To add randomness to the new data, I also applied random augmentation in the pipeline.

4.3 GPT-2 Generator

The projects ran experiments with both published GPT-2 models (124M and 355M). The input is the full script, including both settings and dialogues that enable the model to generate short scenes.

To evaluate the performance of the GPT-2 Generators, I randomly picked 500 lines as the seed for the generator, then preprocessed the generated text the same way as that for the BERT classification model (section 4.4), and compared the test accuracy of original lines and the generated lines.

4.4 BERT Classification

The text was cleaned by separating the characters and the lines, removing non-alphabetic words, ignoring short lines with less than 3 words, and tokenizing with BERT tokenizer. Then I fine-tuned the pre-trained Bert-base-uncased model for the multi-classification task.

5 Results

5.1 TF-IDF Analysis

The TF-IDF Analysis is surprisingly insightful. Most of the top 5 words are names, locations, and special events. When looked into the dialogues in a time series, the words with high TF-IDF scores show the changes in the characters' lives.

Top 5 Words of Each Characters

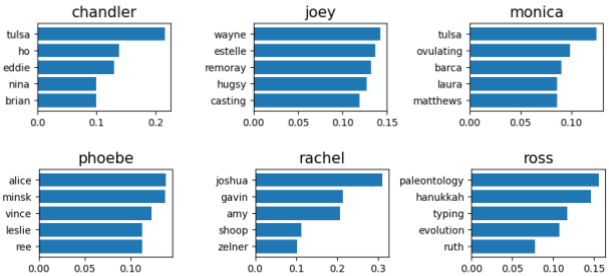


Figure 2: TF-IDF scores (experiment 1)

Names: 17 out of all 30 words are names of their dates, rivals, coworkers and, other important supporting roles

Locations: *Minsk* is a city in Russia. It is important to Phoebe because Vince, the love of her love, works in a lab there.

Only Chandler and Monica have an overlapping word *Tulsa* because they are a couple, and Tulsa is where Chandler works for a while. The distance betweenTulsa and New York, where Monica lives, is the reason why Chandler finally makes the lifetime decision of quitting the job he complains, but good at, for years, and start fresh in the advertising industry.

Special Events: Casting is among the top 5 words of Joey because he is an actor while ovulating is important for Monica as she wants to have a Baby since Season one.

Paleontology and *evolution* are important parts of Ross's life as he has been obsessed with dinosaurs and science since he was a child. He has a Ph.D. in this field and works as a professor at NYU.

Top 5 Words of Chandler [1] in Each Season

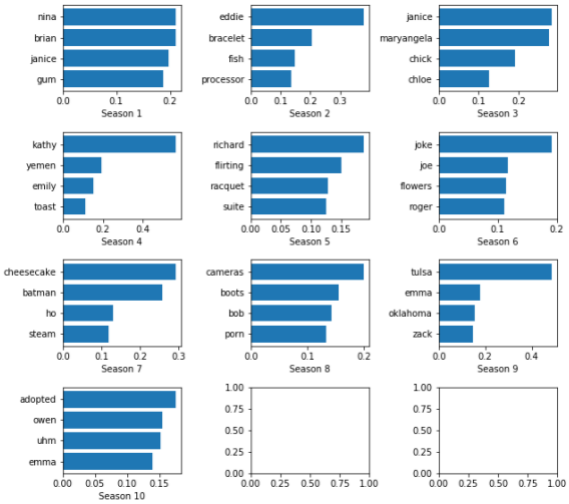


Figure 3: TF-IDF scores (experiment 2)

All turning points in Chandler's life are in Figure3. Chandler and Janice have an on-again-off-again relationship in Season 1 to Season 3. After their last break-up, he falls in love with Kathy in Season 4, but it doesn't work out. Richard becomes important to Chandler in Season 5 after he and Monica become a couple. He is jealous of Richard as he witnesses how Monica and Richard used to love each other. He changes his career after being transferred to the Tulsa office in Season 9 and also become the godfather of Emma, daughter of Ross and Rachels'. In Season 10, Monica and Chandler find out they can't have a baby of their own, and they have to adopt. They want to raise the incoming twins in the suburbs. Hence, they move out of the most important scene of F.R.I.E.N.D.S, their apartment in New York City in the last episode.

5.2 BERT Classification

The experiments and results are listed as follows:

Models	1	2	3	4	5	6	7
Augmentation	F	F	T	T	T	T	T
# of Samples	42K	42K	77K	77K	84K	84K	84K
Max Input length	32	64	32	64	32	64	128
# of Epochs	10	10	10	10	10	10	10
Batch Size	128	128	512	128	512	128	64
Accuracy	0.328	0.334	0.340	0.363	0.351	0.406	0.410

Table 2: BERT classification Results

Conclusions:

1. Data Augmentation is powerful for Natural Language Classification. Increasing the number of samples improves model performance.
2. Increasing the maximum length of the input vector with padding also contributes to model improvements. However, the positive correlation only exists in a range that limited by the length of the original lines
3. An epoch of 10 is not enough for this task. Test accuracy reaches the plateau of 47% after 35 epochs for Model 6

Figure 4: Test Accuracy of Model 6

4. Test accuracy of each class, except Monica, are around 48%. And that of Monica is 42%.

[1] You can find the visualization of others in the EDA folder

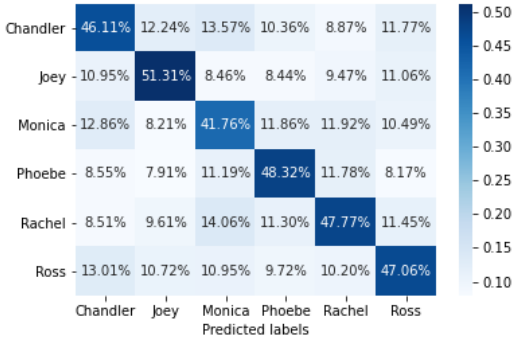
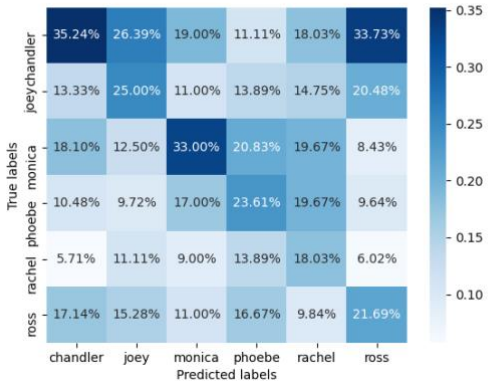


Figure 5: Confusion Matrix of Model 6(40 epochs)

5.3 GPT-2 Generator

I evaluated the text generated by GPT-2 with a BERT classifier that has a test accuracy of 42.78%. The accuracy of the generated samples is 26.17%. And the confusion matrix is very different from that of the test dataset.

Besides, when reading the generated text as a human, I found that the GPT-2 generator can generate meaningful lines. But the logic between lines is far from satisfying.



6 Discussion

Natural Language Generation is much more complicated than Text Classification. With the same corpus, Transformers perform nicely in classification. An accuracy of 47% is great. The training samples are short, and a lot of the lines are possible to come from anyone. However, with such a small dataset, the machine fails to write reasonable dialogues. Increasing the amount of data is too expensive. And we can't use scripts from other shows because of the differences in characters and settings. Deep learning has been proved powerful in many areas, however, for more creative tasks like writing a story, humans still outperform machines by a lot.

References

Liang, P., Jordan, M. I., & Klein, D. (2009, August). Learning semantic correspondences with less supervision. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (pp. 91-99). American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008). Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.