Natural Language Generation with Charles Dickens

NLP Final Project: Wesley Stevens

Introduction

Language Generation is a sub-field of the field of Natural Language Processing (NLP), Artificial Intelligence (AI) and Cognitive Science (CS) that has been studied since the 1960s. Natural Language Generation (NLG) is still one of the major challenges blocking the path towards achieving Artificial General Intelligence (AGI) and it entails incorporating fundamental aspects of artificial intelligence and cognitive science to create coherent sentences or creative material [1].

The field of text generation systems started gaining traction through various statistical methods, one being Markov Chains, which we will discuss in detail later. These statistical methods were much better than the previous method of deterministically selecting words, but still not ideal. Decades later, when neural networks became useful through technological advancements, researchers were able to achieve astounding results [2].

Markov Chains were first introduced in 1906 by Andrey Markov, with a goal of showing that the law of large numbers can apply to random variables that are not independent. How they function in NLG is by generating a probability distribution from word to word. By the law of large numbers, and with enough data, one can create an accurate representation of how words flow probabilistically [5].

One of the most current of these neural networks is called GPT-2, a large transformer network with over 1.5 billion parameters. It was trained on a dataset of 8 million web pages and took almost 9 months to train with the objective to predict the next word. It has reportedly achieved results far outpacing previous statistical methods and the popularized recurrent neural networks [3].

Interesting capabilities came forth as a result including the ability to generate conditional synthetic text samples of unprecedented quality. On language tasks like question answering, reading comprehension, summarization, and translation, GPT-2 far outperformed previous state-of-the-art methods and models. The creators were initially hesitant to release this model as they were concerned over the ethical implications of giving such a unique and powerful tool to the public. It was released a few months later, however with no explanation of what caused them to change their minds [3].

After processing the data and training a Markov Chain, we compare and contrast how it performs at generating text in the voice of Charles Dickens against a small GPT-2 model with pre-initialized weights [4].

Methods

We use six books by Charles Dickens pulled from Project Gutenberg: A Tale of Two Cities, Our Mutual Friend, Oliver Twist, A Christmas Carol, Bleak House, and The Pickwick Papers. These books are used as the training data for each of the NLG methods.

Preprocessing

For the preprocessing, we removed the Gutenberg tags, page headings such as “iv”, “iii”, and “xxi”, and special characters. We also put everything in lower-case. Exclamation marks, periods, and question marks were chosen to indicate the end of a sentence [4]. After preprocessing the text corpuses, we tokenized each corpus and fed them into each method.

Markov Chains

The Markov model is represented as a giant matrix where each row and column represent a unique word. We also include a row and column for words that begin sentences, and words that end them. For n unique words, there are n+2 rows and n+2 columns. To train the Markov Chain, we increment each entry where the *ith* word leads to the *jth* word. Thus, we are able to build sentences by randomly drawing from the probability distribution of sentence-starting words and recursively drawing until we draw a sentence finishing word.

GPT-2

GPT-2 pre-trained weights have been released for small, medium, and large versions of GPT-2. There are already many wrappers around using GPT-2 on GitHub, so we will utilize one to make interacting with it simple [6]. We will also download the pre-trained weights for the small model and transfer learn onto the same books by Charles Dickens. We will similarly see how this performs in NLG given various amounts of starting text.

Results

We first note that the Markov Model trained in seconds whereas GPT-2 took several hours to transfer learn. Randomly generating sentences with no prompt, the Markov Chain method gave multiple sentences such as the following:

```
Though recalled first, attained to duty wine-shop lithe contents
night. Herself a an tread recalled mail recalled first matters years
to belief book shoemaker life fellow wine-shop mail thread book
recalled to blunderbuss recalled chapter life. Necessaries book
sight period the age shoemaker book gorgon contents mail shoemaker
chapter. Newyear's contents the lord jackal mail second i first sea.
Because Monseigneur.
```

GPT-2, on the other hand, generated the following text when given no prompt.

```
He sat at his desk, and looked down at the floor, and watched the
faces. Some of his friends came and sat down, so that the next
shadows appeared from a corner of a corner, where there were still a
few takes, and some of them turned to look round for the other. The
dear old lady had not the least idea of what to look for, when she
looked round in the same place, at the same time and from the same
place, or to look at it once in a while, and wonder what it was, or
what it had looked like. To be sure, the lady who had taken the
paper on was an old woman herself, of a very high complexion and
very long hair. But there were certain features of that hair, that
were not noticed, but there were certain features that were. They
were always there. Even though their faces were not at all alike,
```

> when the women who were looking in on her looked up at her in the morning, and said, "I was looking down the street," she went on: "I was looking at a woman who looked down the street for herself."

We then presented each method with the word “And” and generated several samples of the results. The Markov chains generated the following text.

> And gloom wine-shop life drawn life i coincidence chapter period shadows years still outhouses contents shoemaker secret age. And later the description later night hundreds contents the boot shadows the some took two. And because second two five first wine-shop years recalled book sight book to stock contents.

And GPT-2 generated the following text for this experiment.

> And looked and looked, for the sternest time, and for the most exquisite time, and the most beautiful and happy and beautiful. And beautiful before they had said something in their hearts, and felt them with her face. After an hour's long silence, Mrs. Winkle looked at her brother. And as they thought, perhaps for the first time were going to go on looking at each other's faces the same way, and perhaps for the second to go on looking at them that way for the first time. So they went upstairs, and looked and looked, for the sternest time, and for the most exquisite time, and the most beautiful and happy and beautiful.

As a final test, we presented each method with the phrase “It had been a long time”. The results for the Markov method are found below.

> It had been a long time's opposite nine of sentencing contents i shoemaker book to lately people a smear the blackheath. It had been a long time loom book wine-shop thread book 's their honest picture the dim preparation first scarcely belief book picture golden the lies thread shadows the mrs. It had been a long time boffin was. It had been a long time book recalled mail thread to sight five five promises recalled book.

And GPT-2 generated the following text for this experiment.

> It had been a long time. Mr. Carus, with his wife and myself here, was the first one who looked at me, and wanted to take me home by my side, and I was much more than he could have expected. If I had not been able in that time to be with my father and my mother, they would have all taken their leave of me.

> It had been a long time, to be sure, the lady who had taken the paper on was an old woman herself, of a very high complexion and very long hair. But there were certain features of that hair, that were not noticed, but there were certain features that were. They were always there. Even though their faces were not at all alike, when the women who were looking in on her looked up at her in the morning, and said, "I was looking down the street," she went on: "I was looking at a woman who looked down the street for herself."

Discussion

The Markov Chain does fairly well at generating sentences, but they lack punctuation and capitalization, making them difficult to interpret. It is also interesting to note that most of the time, the Markov Chain language generation makes sense if only a few words are considered. The longer the sentence is, the worse the overall cohesion is. It seems that the Markov Chain does well at moving word to word but it fails to incorporate the intricacies of language. Specifically, how a sentence can change and adapt based on modifiers located anywhere in the sentence.

This failure of the Markov Chain makes sense in the context of how this statistical structure operates. The Markov Chain only records the probability from passing from word to word and contains no information about previous words. This points to a possible improvement of the Markov Chain: utilize sentence information for n-grams rather than just unigrams. Further research could be done to create probability distributions for bigrams, trigrams, and so on, and then weigh them into language generation. This extension could be thought of as a 3-dimensional Markov Chain where each layer is a different n-gram probability distribution. Then as one iterates through, generating text, the probability of cohesion could be calculated and be the deciding factor to accept the proposed next word or resample.

In our experiments the text generated by GPT-2 is far more cohesive and impressive than the text generated by the Markov Chain. This is interesting because the objective GPT-2 was trained on was the same one that drove the Markov Chain. The most likely explanation for this is that the structure of GPT-2 allows the weighing in of words that are more than one step away, which was one of the shortcomings of the Markov Chain. Another is that GPT-2 was trained on more data than the Markov Chain and thus had a larger vocabulary.

Regardless, the text generated by GPT-2 is almost human-like in quality. In longer sentences, cohesion begins to degrade, but GPT-2 is able to generate stories that make sense and are interesting to read as a whole. With GPT-2 language generation, one could easily generate stories, responses, or other text that would need little to no tweaking to make sense. Further research could be done to determine if keywords could be used to generate text such as news articles, research articles, or entire books with this tool.

Improving upon the GPT-2 method shown here is simple: use one of the larger models with the corresponding pre-trained weights. But beyond that is unheard of to this day. There have been releases of other models that perform similarly to GPT-2 such as BERT and CTRL, but are much larger and cumbersome [7] [8].

Ethics

Being able to generate text in another person's voice raises a question of morality. Applications like this could easily be used to harm or defame persons with reputation. This could also easily become a source of forgeries and fake news. It is unclear what could be done to stop this unethical behavior, especially as these NLG applications become flawless. Discussions need to be held to better protect and secure a person's original material in the future.

Conclusion

The field of Natural Language Generation has come a long way since Markov Chains were introduced in 1906. Markov Chain variants still have impact in fields such as Game Theory and Economics, but from the experiments done here, we see that they perform poorly when tasked with generating text [5]. Modern solutions such as GPT-2 require large amounts of data and computational resources and the NLG results are astounding. While Markov Chains struggle to create coherent sentences in every experiment, GPT-2 far exceeds expectations, producing paragraphs that are not only coherent, but captivating and sometimes humorous. With little effort, some data, and a lot of time, any casual user can generate small corpuses in the voice of any author, blogger, or songwriter. Other applications include recipe generation, chat bots, and even an essay assistant. The field of NLG is complex and is evolving quickly, and there are many ethical implications to be considered; but with models such as GPT-2 paving the way, Artificial General Intelligence is one step closer.

References

[1] S. Santhanam and S. Shaikh, "A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions," arXiv:1906.00500 [cs.CL], 2019.

[2] T. Mikolov, M. Karafiat, L. Burget, J. y. Cernock and S. Khudanpur, "Recurrent neural network based language model.," *Interspeech, volume 2,* p. 3, 2010.

[3] "Better Language Models and Their Implications," OpenAI, 14 February 2019. [Online]. Available: https://openai.com/blog/better-language-models/. [Accessed 24 June 2020].

[4] "Project Gutenberg," [Online]. Available: http://www.gutenberg.org/ebooks/author/37.

[5] in *Introduction to Probability*, p. 459.

[6] nshepperd, "gpt-2," GitHub, [Online]. Available: https://github.com/nshepperd/gpt-2. [Accessed 24 6 2020].

[7] minimaxir, "ctrl-gce," GitHub, [Online]. Available: https://github.com/minimaxir/ctrl-gce. [Accessed 25 6 2020].

[8] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 24 5 2019. [Online]. Available: https://arxiv.org/abs/1810.04805. [Accessed 25 6 2020].