Text generation

Ruilin He

March 2022

Abstract

On September 8, 2020, an article about the harmlessness of AI to humans was written independently by the computer using GPT-3 model GPT-3 (2020). At the same time, some intelligent assistants such as Apple's Siri and Amazon's Alexa are already being used in daily life, which means that as one of the fastest applications of machine learning, natural language generation has made significant progress. There is no doubt that the natural language generation is revolutionizing the way we communicate with machines and is starting to enter our daily lives.

1 Introduction

Researchers hope that machines can be trained to write logical articles like humans in machine learning. Therefore, as one of the virtual fields of natural language processing, natural language generation, namely, text generation, has been widely studied. This report will discuss the historical development of natural language generation, introduce how different neural networks automatically generate text, explore the challenges that remain in natural language generation.

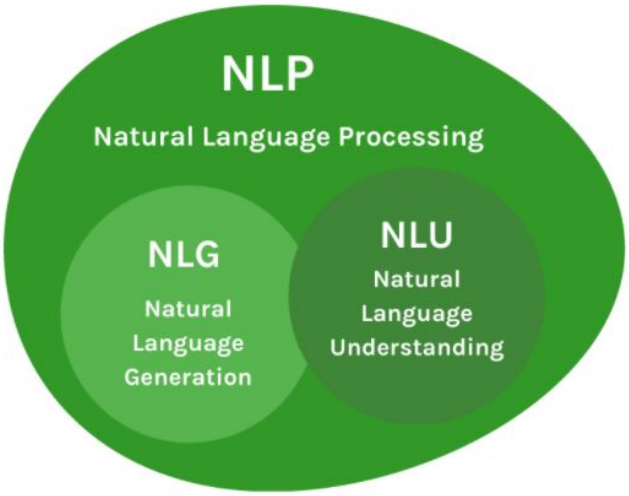

Figure 1: NLP, NLG and NLU
Reiter & Dale (1997)

2 Concept of text generation

In the beginning, before introducing different neural networks, it is significant to give a clear definition of text generation. In general, natural language generation is to let the machine understand natural language through training and then express it logically. Due to different compilation methods, there will be different versions of natural language generation. In addition, text generation can be divided according to different training purposes. Currently, it is mainly text-to-text generation, data-to-text generation, and image-to-text generation.

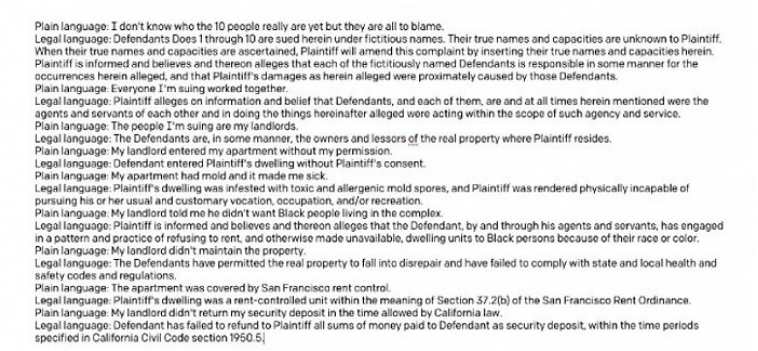

Figure 2: text-to-text generation
Naojiti (2020)

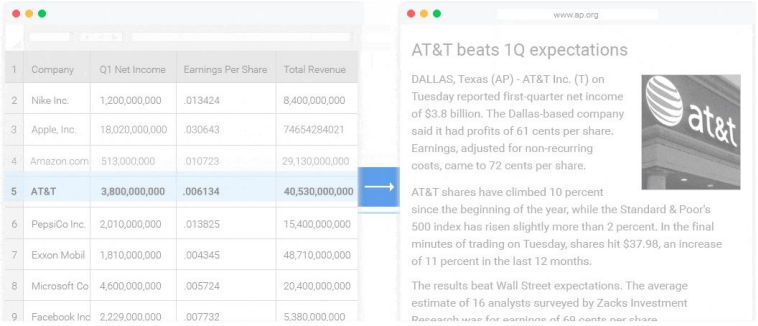

Figure 3: data-to-text generation
Naojiti (2020)

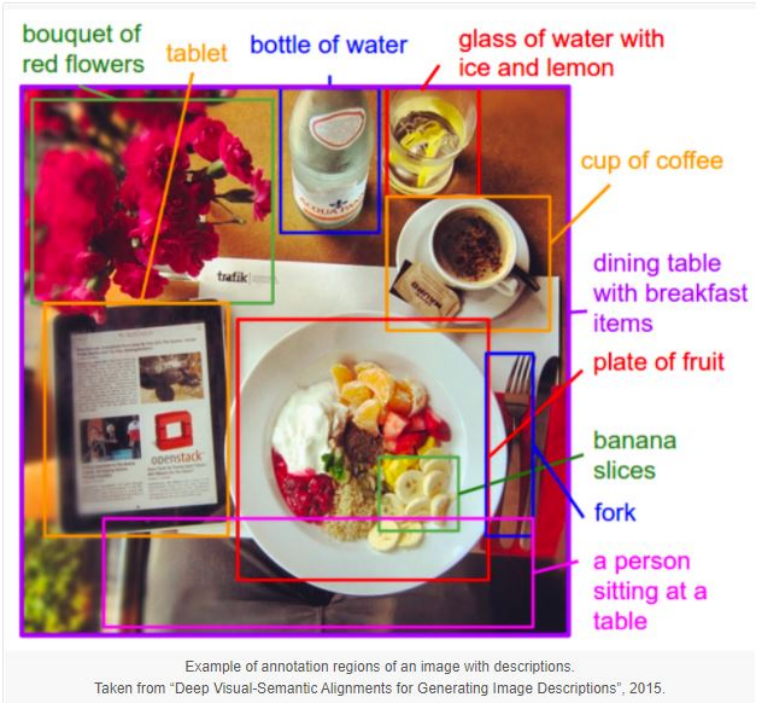

Figure 4: image-to-text generation
Chun et al. (2021)

3 Different models for text generation

Early network models of natural language generation are normally based on rules and templates. It is expected that the generated text can conform to syntactic rules by setting sentence structures or grammar templates. Text generation in these ways was simple and missed messages. However, as researchers continue to develop new models, some innovational networks such as Recurrent Neural Network have been invented and applied in text generation, making the generated text more logical.

Figure 5: neural network
Micheli-Tzanakou (2011)

3.1 Markov Chain

In the early days of text generation, Markov chains were considered a basic algorithm by researchers. A Markov chain is a stochastic process in which we assume that the first few states play a decisive role in predicting the next state. Unlike a coin toss, these events are not independent of each other. If it is sunny today, there is a 90% chance that it will be sunny tomorrow, and there is a 10% chance that it will rain.

Markov model is a text generation method based on a language model, it only predicts events based on the probability of previous events. Since its future state distribution only depends on the present and has nothing to do with the past, it is necessary to get the probability that a word appears to predict the next word.

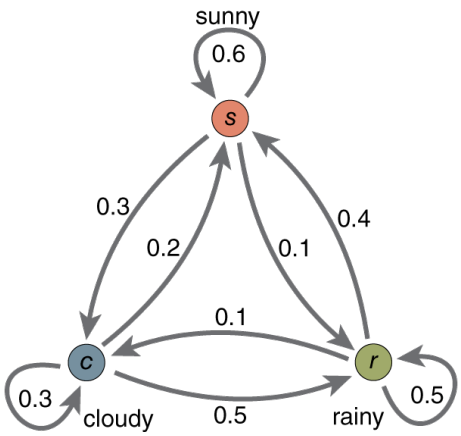

Figure 6: Markov Chain
Borucka (2018)

The process of automatic text generation with Markov is simple. It is worth noting that in the process of text generation using Markov chains, a large number of words need to be connected, if just a few words are connected for text generation, all the context and structure of the previous words in the sentence will be lost, which may lead to the wrong prediction. At the same time, due to the large number of words that need to be associated, the model needs lots of running time if it wants to generate text with more content. Is there a model that can output a sequence of words according to a certain logic? Brilliant researchers thought of using Recurrent Neural Network for text generation.

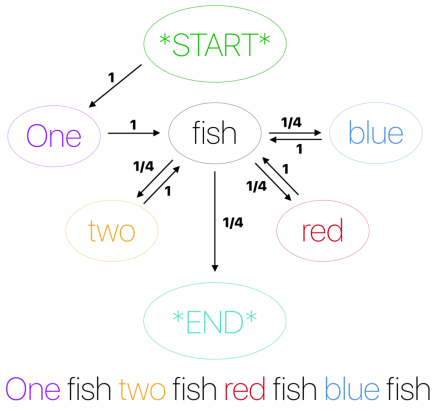

Figure 7: NLG using Markov Chain
Arvindpdmn (2020)

3.2 Recurrent Neural Network

From the Markov chain, it is known that the order of words plays an important role in text generation. Besides, with the development of machine learning, researchers have tried to use deep learning methods to train machines to generate text.

To capture the relationship between different words, Recurrent Neural Network was born. Recurrent Neural Network has an additional box to store the previous information. When the next output is generated, the model must consider the information stored in the box. As data is continuously entered, the information in the box is constantly updated. This box is called the hidden state.

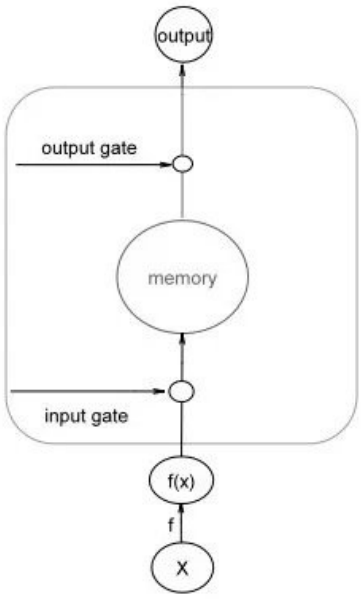

Figure 8: Recurrent Neural Network
Xiaoqiang (2019)

Compared with Markov chains, Recurrent Neural Networks are more efficient in representing high-dimensional data. At the same time, different modeling methods also lead to different learning methods. The Recurrent Neural Network model is solved by the gradient descent method, while the HMM cannot be solved by the EM algorithm by gradient descent.

But there is still a non-negligible missing in processing Recurrent Neural Network, that is, the hidden state can only hold short-term memory. In other words, if a sentence is long enough, then Recurrent Neural Network may forget early words. Just as people only remember important things and forget unimportant things, in this case, Jürgen Schmidhuber, the father of Recurrent

Neural Network, added a gate mechanism to the hidden state of the small box that stores previous information, using 0 to discard unnecessary information and 1 to Keep useful information. Imagine that this small box now has three gates. The forget gate is used to save the original information, the input gate determines how much of the current information is input into the box and the output gate determines the output of the small box. This improved Recurrent Neural Network is called Long short-term memory Network Hochreiter & Schmidhuber (1997).

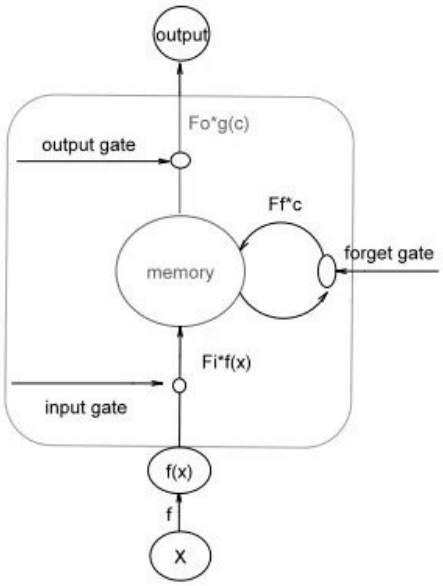

Figure 9: LSTM Neural Network
Xiaoqiang (2019)

3.3 Transformer

In the process of natural language generation, it is worth noting that the number of generated words may not be the same as the number of original input words. In the structure of Recurrent Neural Network, it can only achieve N to N, 1 to N, or N to 1 task, and cannot achieve N to M tasks. Therefore, Google proposed Transformer in the article "Attention Is All You Need".

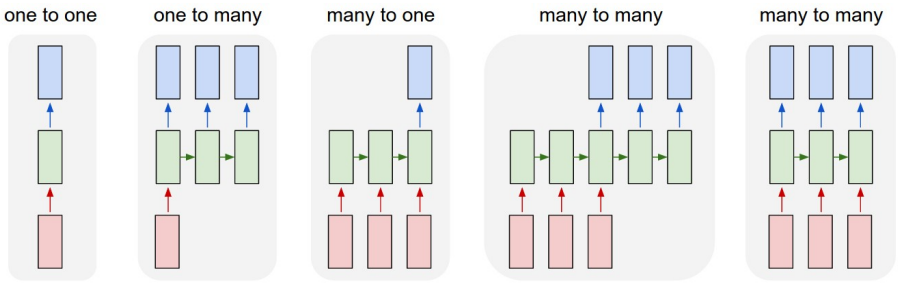

Figure 10: different types of Recurrent Neural Network
Karpathy (2015)

The transformer is a model with an Encoder and Decoder, Encoder and Decoder are still Recurrent Neural Network structurally. The Encoder extracts the meaning of the input text, and the Decoder converts the meaning into natural language, which solves the problem of information asymmetry. In addition, the Self-Attention mechanism is used to extract important information in the sentence from the input text to prevent the memory disappearance problem in Recurrent Neural Network, and then select the required information according to the generated order. Such a structure not only supports parallel computing but also is more efficient. It is worth mentioning that the most advanced natural language generation model GPT-3 model is improved from the transformer model.

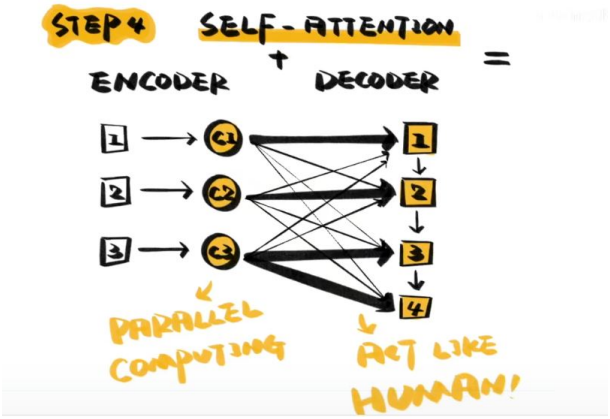

Figure 11: Transformer
Karpathy (2015)

4 Challenges and Trends in Text Generation

As researchers delve deeper into natural language generation, from generating simple sentences using simple Markov chains to more logical text using models with self-attention mechanisms, which means that natural language generation is becoming increasingly dramatic. However, there are still many challenges in the field of natural language generation.

It is undeniable that the development of text generation can help newspapers publish the latest news more quickly, but there is also the possibility of newspapers publishing fake news. At the same time, if enough personal articles of others are collected and trained in the model, these models can imitate the writing habits of others and publish some inappropriate remarks to damage the reputation of others. In addition, ethical issues are also one of the problems worth noting in text generation. In the model generated by GPT-2, Naojiti (2020), there is a description of discrimination against women. The discriminatory content in the training data has a significant effect. However, due to the big data, it is difficult to check related immoral content by people.

Therefore, in the use of natural language generation, specific instructions need to be given for the model to make predictions, and manual modification is required. It is conceivable that some news reports in the future. It is conceivable that in the future, some news reports will be completed by natural language-generated articles and journalists so that the fastest news can be published, and errors in natural language generation can be reduced through manual review.

To sum up, training with enough data to build a model with sufficient depth can achieve a state-of-the-art network to generate text that meets the requirements of the task. Although there are still some obstacles to overcome, natural language generation, that is, text generation has amazing potential in machine learning. It is believed that people will apply this technology in more aspects in the future.

References

Arvindpdmn (2020), 'Natural language generation', `https://devopedia.org/natural-language-generation`.

Borucka, A. (2018), 'Three-state markov model of using transport means', *Business Logistics In Modern Management* pp. 3–19.

Chun, P.-J., Yamane, T. & Maemura, Y. (2021), 'A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage', *Computer-Aided Civil and Infrastructure Engineering* .

GPT-3 (2020), 'A robot wrote this entire article. are you scared yet, human?', *The Guardian* .

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Karpathy, A. (2015), 'The unreasonable effectiveness of recurrent neural networks', `http://karpathy.github.io/2015/05/21/rnn-effectiveness/`.

Micheli-Tzanakou, E. (2011), 'Artificial neural networks: an overview', *Network: Computation in Neural Systems* **22**(1-4), 208–230.

Naojiti (2020), 'Gpt-3, which has been praised in the sky, how to go to commercialization?'.

Reiter, E. & Dale, R. (1997), 'Building applied natural language generation systems', *Natural Language Engineering* **3**(1), 57–87.

Xiaoqiang (2019), 'How to understand lstm', `https://easyai.tech/blog/understand-lstm/`.