Demystifying Text Generation Approaches

Lichi Upadhyay, M. I. Hasan & P. S. Patel

ABSTRACT

Natural Language Processing (NLP) is a subfield of Artificial Intelligence that is focused on enabling computers to understand and process human languages, to get computers closer to a human level understanding of language. The main emphasis in the task of text generation is to generate semantically and syntactically sound, coherent and meaning full text. At a high level.

The techniques has been to train end to end neural network models consisting of an encoder model to produce a hidden representation of text, followed by a decoder model to generate the target. For the task of text generation, various techniques and models are used.

Various algorithms which are used to generate text are discussed in the following subsections. In the field of Text Generation, researcher's main focus is on Hidden Markov Model(HMM) and Long Short Term Memory (LSTM) units which are used to generate sequential text. This paper also discusses limitations of Hidden Markov Model as well as richness of Long Short Term Memory units.

Keywords: natural language processing, HMM, RNN, ANN, LSTM.

Classification: DDC Code: 006.3 LCC Code: Q335

Language: English

London Journals Press

LJP Copyright ID: 975832
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science and Technology

Volume 23 | Issue 1 | Compilation 1.0

Demystifying Text Generation Approaches

Lichi Upadhyay[α], M. I. Hasan[σ] & P. S. Patel[ρ]

ABSTRACT

Natural Language Processing (NLP) is a subfield of Artificial Intelligence that is focused on enabling computers to understand and process human languages, to get computers closer to a human level understanding of language. The main emphasis in the task of text generation is to generate semantically and syntactically sound, coherent and meaning full text. At a high level.

The techniques has been to train end to end neural network models consisting of an encoder model to produce a hidden representation of text, followed by a decoder model to generate the target. For the task of text generation, various techniques and models are used.

Various algorithms which are used to generate text are discussed in the following subsections. In the field of Text Generation, researcher's main focus is on Hidden Markov Model(HMM) and Long Short Term Memory (LSTM) units which are used to generate sequential text. This paper also discusses limitations of Hidden Markov Model as well as richness of Long Short Term Memory units.

Keywords: natural language processing, HMM, RNN, ANN, LSTM.

Author α σ ρ: Department of Computer Engineering, BVM Engineering College, Gujarat, India.

I. INTRODUCTION

Natural Language Processing (NLP) is a subfield of Artificial Intelligence that is focused on enabling computers to understand and process human languages, to get computers closer to a human level understanding of language. Humans have been writing things down for thousands of years. Over that time, our brains have gained tremendous amount of data and experience in understanding natural language. [9] The goal of NLP is to accomplish human like language processing. It is a theoretically motivated range of computational techniques. There are various applications such as Machine translation, Speech synthesis, Automatic summarization, word processing, Text Prediction, Dialogue systems, Named Entity Recognization, Story understanding, Language teaching and assistive computing.

The steps for generating text is divided in to four phases. First is dataset collection, second one is cleaning of that dataset, third one is loading of cleaned text and the final one is generating text.

In 2016, Artificial Intelligence has generated movie script "Sun spring" created by Ross Goodwin and also directed by Oscar Sharp. It was written by program called Jetson which is called Benjamin. Benjamin's other films are "Zone out" and "This wild." In addition, A new chapter of famous series "Harry Potter" by J. K. Rowling had been published by Botnik studios titled as "Harry Potter and the Portrait of What Looked Like a Large Pile of Ash."

There are many songs which are generated by Artificial Intelligence such as "Daddy's car" and "Break free". The other experiment is Wikipedia text generation. The poems can also be generated by Artificial Intelligence. In Chinese literature, poems have been generated by AI. The connectionist models are used to model the aspects of human perceptions such as, cognition and behavior, learning process under such behaviors and storage and information retrieval from memory. The Neural Networks, which are a sub part of connectionist models, are nothing but a model that mimics how human brain works. We will discuss how these neural networks are useful for generating text.

London Journal of Research in Computer Science and Technology

II. RELATED WORK

Alex Graves (2014) [1] emphasized on demonstrating that LSTM can use it's memory to generate complex, realistic sequences containing long range structure. In this paper, Alex Graves has taken an approach for generating sequence is for text. He had also shown that how recurrent neural networks can be used to generate complex sequences with long range structure, simply by predicting one data point at a time. In this paper, he had shown that how Recurrent Neural Networks can be trained for sequence generation by processing real data sequences one step at a time, and predicting what comes next. Here predictions are assumed probabilistic and it is also assumed that sequences can be generated from a trained network by iteratively sampling from the network's output and then feeding in the sample as input at the next step. It has been stated in paper that in practice, standard Recurrent Neural Networks are not able to store information about past inputs for very long.The word level Recurrent Neural Network performed better than character level network but that gap appeared close when regularizations are used.

Lipton et al. (2015) [2] has given a review about recurrent neural networks regarding how they learn sequences. The Recurrent Neural Networks are connectionist models. The connectionist models are used to model the aspects of human perception, cognition and behavior, learning process under such behaviors and storage and their retrieval of information from memory.

The neural networks are powerful learning models that give the state-of-the-art results in a wide range of supervised and unsupervised machine learning tasks. But standard neural networks are having limitations, too. In that, there is no dependency between the concurrent states or layers. So when data is related through time or space, these network models are not useful. The examples of such data are frames from video, audio snippets, words pulled from sentences.

Thus, Recurrent Neural Network's requirement came in to picture. Because they are connected through time, all the data that is related through time can be modeled. The recurrent neural network is depicted in figure(1).

Zhengdong et al. (2014) [3] has proposed two convolutional neural networks models for matching two sentences, by adapting the convolutional strategy in vision and speech. The proposed models not only depicts the hierarchical structures (structure of sentences in which phrases are nested in phrases) of sentences with their layer-by-layer composition and pooling, but also can capture the rich matching patterns at different levels. A successful sentence-matching algorithm needs to capture the whole structure including the internal structures of sentences and also rich patterns in their interactions.

Kalchbrenner(2014) et al, have described convolutional architecture dubbed the Dynamic Convolutional Neural Network for semantic modeling of sentences. The network uses Dynamic K-max pooling, a global pooling operation over a linear sequences. The main aim of this paper is to analyze and represent thesemantic content of a sentence for a purpose of classification or generation.

Manurang et al. (2012) [4]has implemented system, McGonagall which uses genetic algorithm to construct text. In this paper, the main goal of authors is to generate texts which are syntactically well formed, meet certain pre specified patterns of metre and convey some meaning. They have proved that if some constraints on metre were relaxed, then their experiments can generate relatively meaningful text. The poetry generation involves many aspects of languages so automatic generation of such poetic text is challenging. They have set some restricted definition of poetry as a text that embodies meaningfulness, grammaticality and poeticness.

Malmi et al. (2016) [5]focus on generating rap lyrics They have given model which is based on two machine learning techniques: 1). The RankSVM algorithm 2).Deep neural Network model with a novel structure. They have taken dataset containing over half a million lines from lyrics of 104 different rap artists, and then new

lyrics are constructed line by line. They have described typical rap lyrics, different rhyme types that are often used. They have considered parameters: 1). Rhyming 2). Song Structure 3).

Automatic Rhyme Detection. They have divided next line prediction problem in to three groups that are rhyming, structural similarity and semantic similarity. They have generated tool called Deepbeat.org which generates rap lyrics by giving a key word as input.

Wei et al. (2018) [6]have tried to generate Classical Chinese poetry, which often incorporates expressive folk influences filtered through the minds of Chinese poets, which consistently has been held in extremely high regard in china. In this paper, they have proposed a Poet based Poetry generation method which generates poems by controlling not only content selection but also poetic style factor. They have done studies that improves the content quality issues of poetic generation system. PoetPG framework takes the content of current line and poet's name as input and then generates a poem in the following two stages: Poetic Style Model, Poem generation.

III. METHODOLOGY

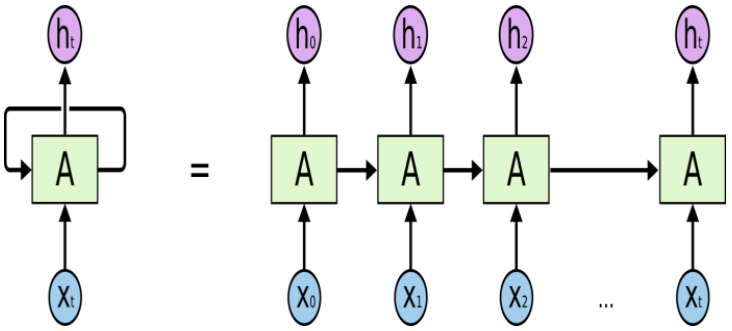
When any writer or poet determines to write about any particular topic, he/she has to gather abundant knowledge about that topic. That knowledge will work as raw material for building a new block. So, from that raw material he/she will be able to write new things about that topic, which will be proprietary. This process of generating new text will be same for the computer as of humans. Text Generation is a part of Natural Language Generation. The Neural Networks are used to model these facilities in the computers.

The connectionist models are used to model the aspects of human perceptions such as, cognition and behavior, learning process under such behaviors and storage and information retrieval from memory .The Neural Networks , which are a sub part of connectionist models, are nothing but a model that mimics how human brain works.

Basically, in a supervised learning ANN (Artificial Neural Network) plays an important role. If we compare it to the human brain, then we can assume that ANN works as temporal lobe, CNN (Convolutional Neural Network) works as a occipital lobe and RNN(Recurrent Neural Network.) works as a frontal lobe of the brain.

The ANNs are very powerful tool to learn machine perception tasks and gives various state-of-the-art results in a wide range of supervised and unsupervised machine learning tasks. But the standard neural networks have a major shortcoming i.e. the current output is independent of previous output. Which is not advantageous to our definition. Humans have context about things, so he/she can get the meaning of new things. When we are reading text book of any subject, if we have understood previous paragraph, then and then only we are able to understand the current paragraph. So we can reach to the conclusion that our current output is dependent on the previous one. So for our definition, RNNs are very helpful which address this issue. In these, networks have many loops which allow information to remain in it.

Basically, in these networks, neurons are connecting to themselves through time. So that they have memory which is short-term, but they can remember what was just happened in the previous neuron or layer. Which helps our definition to generate the sequences? The representation of RNN is as following.



(a) Representation of RNN

LSTMs (Long Short Term Memory units), which are called memory cells of RNN which work as memory units of RNN and also overcome the limitation of traditional Artificial Neural Network. The other techniques used to generate text are Markov chain, Recursive neural networks, Long Short Term Memory ,etc. [10] We will see LSTMs and HMM in depth in following subsection of the paper.

LSTMs:

The Long Short Term Memory unit is a memory unit of Recurrent Neural Network as discussed above. The traditional recurrent neural network is having a shortcoming of vanishing gradient, which is overcome by Long Short Term Memory.

The LSTM captures long range dependencies that means it can remember what has happened just previously. These LSTMs can implemented in various ways such as word level and character level. It observes sequence and then gives output according to input. The standard representation of LSTM is as shown below.

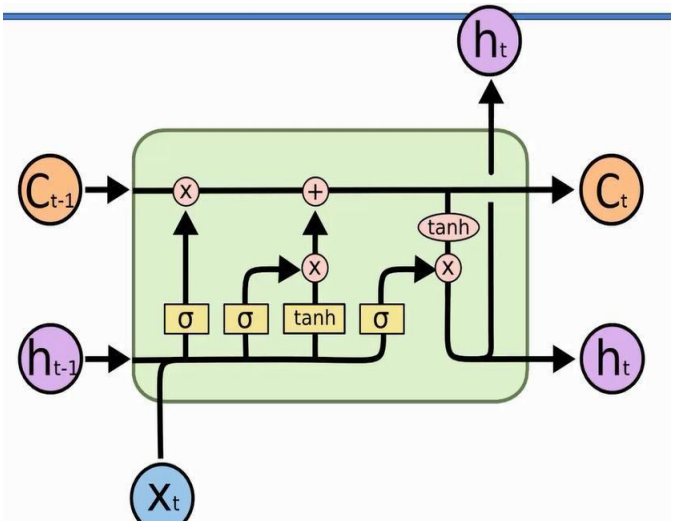


Fig. 2: LSTM[8]

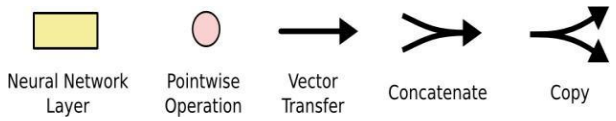


Fig. 3: LSTM notations[8]

Hidden Markov Model

The Markov chains are also capable of showing time dependencies. the data which are related through time (e.g. sequences) can be modeled through Markov models. The Hidden Markov Model is a finite set of transition states, each of which is associated with a probability distribution. These transitions among states are

governed by a set of probability called transition probability. According to associated probability distribution, the observations can be generated. It is called hidden markov model because only outcome is visible to external world, not the internal state transitions are visible. But when the set of possible hidden states grows large, HMM are infeasible. In addition , HMM cannot captures long range dependencies that means HMM cannot remember what has just happened previously.

IV. RESULTS AND DISCUSSION

The Hidden Markov model can model time dependent data. But it cannot capture long range dependencies. As we have to generate text, it can be in form of sentences also. And the sentence can be long. So, these Hidden Markov Model are not useful for generating text. The other model is LSTM. It can be implemented as word level and character level. in character level LSTM , it will observe sequence of characters, and according to that, it will give output. We can generate songs, poems, rap lyrics, etc by giving it's dataset as input.

V. CONCLUSION AND FUTURE SCOPE

from going through various methods used in various papers, we can conclude that, there are different methods available for modelling sequence of words for making sentence. It has been found from the survey that Long Short Term Memory(LSTMs) are best suited for generating text. These Long Short Term Memory units are of the Recurrent Neural Network. The Recurrent Neural Networks are not used for generating meaningful. Long Short Term Memory unit can remember what was just happened in previous layer. The traditional Recurrent Neural Network is having a problem called vanishing gradient, in which weights become smaller and smaller while the network is back propagating weights for training purpose. When these weights are small, and then we forward it in to the network and then we back propagate these weights back in to network for training purpose, then those small weights becomes smaller. This problem is known as vanishing gradient. Thus traditional Recurrent Neural Network faces problem of vanishing gradient. The Long Short Term Memory units of Recurrent Neural Network mitigate this problem of vanishing gradient. So they are having memory. We can use Long Short Term Memory for generating text.,which is best suited for generating text.

REFERENCES

1. A. Graves, "Generating Sequences with Recurrent Neural Networks," Computing Research Repository- CoRR ArXiv, 2014.
2. J. B. C. E. Zachary C. Lipton, "A Critical Review of Recurrent Neyral Networks for Sequence Learning," Computer Research Repository- arXiv, 2015.
3. Z. L. H. L. C. Baotian Hu, "Convolutional Neural Network Architectures for Matching Natural Language Sentences," Neural Information Processing Systems Foundation, 2014.
4. G. R. H. T. Ruli Manurang, "Using genetic algorithms to create meaningful poetic text," Journal of Experimental & Theoritical Artificial Intelligence, vol. 24, pp. 43-64, 2013.
5. P. T.,. H. T. T. R. A. Eric Malmi, "DopeLearning: A Computational Approach to Rap Lyrics Generation," Knowledge Discovery and Data Mining,Association for Computer Machinery , pp. 13-17, 2016.
6. Q. Z. Y. C. Jia Wei, "Poet-based Poetry Generation: Controlling Personal Style with Recurrent Neural Networks," 2018 workshop on computing, Networking and Communications(CNC), 2018.
7. H. O. M. M. Naoko Tosa, "Hitch Haiku : An Interactive Supporting System for Composing Haiku Poem," International Fedaration for Information Processing, pp. 209-216, 2008.
8. christopher c. olah's blog.