Conversational Chatbot using Deep Learning Algorithm and Attention Mechanism

Ammar K, Prithvi K, Hafiz Hassan

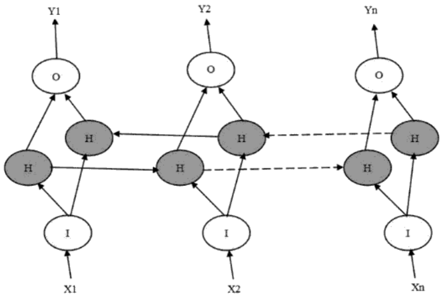
Abstract— This paper identifies the development of Chat-bot using Machine Learning and Artificial Intelligence techniques. A chatbot is quite an interesting problem in Natural Language processing. There are various Machine Learning and Natural Language Processing algorithms for developing a conversational chatbot, even after coming this far the development and research organizations are progressing towards the new age where the chatbots can be given human intelligence. These virtual agents are deployed in Business, Finance, and Telecommunication organizations. They are adopted by banks, mass-media, and businesses in their customer service procedures. There are multiple frameworks available for developing chatbots but these virtual assistants lack accuracy and efficiency in developing real dialogues. Alexa, Cortana, and Google Assistant are some of the most popular virtual assistants. However, the functionality of these assistants is limited and exhaustive, therefore these rule-based chatbots do not provide accurate results hence limiting its capability of continuing a conversation. Therefore, we have developed a conversational chatbot using encoder-decoder architecture and attention mechanisms. The encoder-decoder uses Recurrent Neural Networks to give out results with utmost accuracy.

Index Terms— Deep learning, Chatbot, Bidirectional RNN and Attention model, Tensorflow, Neural Machine Translation

1 INTRODUCTION

THE Chatbot is a language recognition system which helps human to make conversation with a machine using Natural Language Processing. Chatbots can be accessed through multiple mediums like voice, video, and text. In this paper, we provide textual data from the Reddit dataset as input to the chatbot. Due to large input value, the chatbot independently chooses its path of conversation as a natural way of conversation using the question/answer protocol. They are predominantly used in finance and business organizations and are known to replace human-human interaction. A chatbot is a software that helps people to answer their problems or to have someone to converse with for a long duration of time using natural language like English. The conversation between them can be a broad range of conversational topics. Now deep learning has become the top trending technology in the world and chatbot is one of its applications. In this paper, the chatbot we created is an open-Domain chatbot which means it is not fixed to a limited set of questions.

The questions can be asked from any topic and expect a relevant response. But it can be changed to a particular domain also by making changes in the dataset, which means to train the data which is relevant to only a particular domain knowledge.



Since this is an open domain chatbot this can be used anywhere to our day to day life assistance like Siri or Alexa etc. To use the power of deep learning Google has provided with an open-source python library called TensorFlow which consists of all the machine learning and deep learning models and algorithms. Our paper is on the chatbot using Neural Machine Translation (NMT) which is an improvement than sequence to sequence model. The neural network used by our chatbot is Bidirectional Recurrent Neural Network (BRNN). This Bidirectional Recurrent Neural Network is chosen because the input to the chatbot can be Dynamic. This network helps the input to belong and not short.

This BRNN also supports the attention mechanism. This means that the chatbot can remember a long sequence of sentences when given as input. In BRNN the hidden layer has data that can go both down and up which is in both directions.it is a connection of two hidden layers in opposite direction to a single point. Therefore, it can receive information from both past and future states. The Training of BRNN is done the same as the RNN as both the neurons don't interact with each other. When forward and backward propagation is done the weights get changed.

2 RELATED WORKS

2.1 Sequence to sequence(seq2seq)

A sequence to sequence model is the reason behind various running systems in our day to day life. For instance, seq2seq model runs applications like google translate, voice enabled devices and online chatbots. These applications are composed of machine translation, speech recognition, video captioning. In these applications using seq2seq is the best solution. This seq2seq model is the best suited for sequence-based problems, where the inputs and outputs have different sizes and categories.in seq2sesq the length of the input may differ from the length of the output.so how the seq2seq model works is it consists of three parts which are encoder, intermediate vector and a decoder. Encoder is a stack of recurrent units where each accepts a single element of the input sequence collects information for that element and propagates it forward.in a question answering problem the input sequence is a collection of all words from that question. A decoder is a stack recurrent unit where each predicts as output y_t at a time step t. The advantage of this model is that it can map sequences of different lengths to each other.

2.2 How and why Google's neural machine translation (GNMT) took over.

NMT systems uses a huge amount of cpu and gpu power for both in training and in translation inference. Also NMT always has a difficulty with rare words' attempts to solve many of these problems causes by NMT.GNMT is also used for dialogue generation. But machine translation is not solved yet GNMT can still make errors that a human translator can never make, like missing words or rare terms. A large amount of translation errors is reduced using GNMT on language pairs. GNMT is capable of translating the whole sentences at a time rather than piece by piece.it has a large end to end framework, the system will learn by itself over time to get better.it uses an artificial neural network to increase fluency and accuracy in google translate.

2.3 Deep reinforcement learning

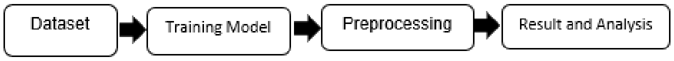
Deep reinforcement learning is the combination of reinforcement learning and deep learning and it's is also the most used machine learning technique now it is because it is able to solve wide range of complex decision-making tasks that was very difficult previously. Deep reinforcement learning is been created for long conversations with the chatbot. The disadvantage with seq2seq is that it can generate consistent response but it produces the same response most of the time regardless of the input asked to the chatbot. Seq2seq can sometimes produces same output due to unavailability of the data and hence that would produce same outcome as the sentence "I Don't Know". This is mainly because of the generic response used in the training set, where most of the responses ends with "I Don't Know".

3 Limitations

The current number of chatbots being developed every day is very high. But all those chatbots have one thing in common which it lacks the ability to respond to long conversations. They often fail to respond accurately to the question or statement asked to the chatbot. Most of the chatbots developed are structured based chatbots which are restricted domain. The majority of the chatbots use simple rule-based techniques. They will perform well in question and answering problems like knowing the information which rarely changes therefore their accuracy is high. They perform in a structured conversation rather than a flexible or creative conversation these kind of chatbots can be used by business for developing automated customer support and for an any enquiry information like in college admissions. These kind of chatbots fail to emulate the real human conversations and lacks flexibility in functioning.

4 System architecture

In this project, we have developed an intelligent conversational chatbot which follows the state of the art techniques.in this GNMT is used for constructing dialogue generator. Normally the GNMT is used for machine translation mechanism but GNMT has also shown good results in using NLP tasks also including dialogue generation and text summarization. GNMT seq2seq module has additional features for dialogue generation. It has a strong seq2seq module but it is not working with the latest version of Google's machine learning framework TensorFlow. It is compactable only with the TensorFlow V1.0. GNMT has a various method inside it for dialog generation agent development.



4.1 Data collection

The dataset we have used is the Reddit dataset for the chatbot. we use a tree-like structure to represent the Reddit where everything is in no-linear format, but the comments are in a linear format. Multiple comments of the same type branch out vastly.

-Top level reply 1
--Reply to top level reply 1
--Reply to top level reply 1
--Reply to reply...
-Top level reply 2
--Reply to top level reply 1
-Top level reply 3

So, we'll take the Reddit dataset and produce input-output pairs. Now there should be only 1 reply per comment. Although some comments have many replies we need to choose only 1 among them. So, we can go with most appropriate reply or the reply which shows up first. More on this later. We first need to get the data. This dataset contains comments of January, year 2015. In most of the Machine Learning algorithms we need to provide the input and the output data. This means for the chatbot we need to provide that the input is this and for this particular input, the following is the output.

4.2 Data pre-processing.

The format of our data:

{"author":"Arve","link_id":"t3_5yba3","score":0,"body":"Can we please deprecate the word \"Ajax\" now? \r\n\r\n(But yeah, this _is_ much nicer)","score_hidden":false,"author_flair_text":null,"gilded":0,"subreddit":"reddit.com","edited":false,"author_flair_css_class":null,"retrieved_on":1427426409,"name":"t1_c0299ap","created_utc":"1192450643","parent_id":"t1_c02999p","controversiality":0,"ups":0,"distinguished":null,"id":"c0299ap","subreddit_id":"t5_6","downs":0,"archived":true}

Each line is like the above. We don't have to take all of the data as input, we only need certain parameters like the parent_id, comment_id and the body of it. Just a single month of comments can cost a storage of 32GB, which however cannot be fit into the RAM. Therefore, our idea here is to go ahead and buffer through the comment files, and then store the data we're interested in into an SQLite 3 database. The most appropriate idea we found is to insert the data of comments to the database. All comments will come chronologically, so all comments will be marked as parents and therefore parents do not have parents. Sometimes there will be certain replies having many comments and therefore we will store that "reply" which will have a parent in the database that we can also pull by id, after which we can search them by parent_id in rows, hence we'll get the replies. Hence we can find certain replies which are better in comparison to the other. So, when this happens, we will change the row with the modified information that is most widely used.

4.3 Training the Model

After the completion of the database now rows have to be divided into training data and test data. For both two files are created (Parent comment and Reply comment). There are also list of protected sentences like blacklisted words, to not let it into the learning network. The training files are fed into the multiprocessing tokenizer, as they are CPU intensive. The sentences which are given as input are divided into tokens on basis of space and punctuation. For each step vocabulary size is 15000.which is sufficient for a virtual ram having 4 Gigabytes. RegX is used to mainly for checking whether a string containing a specific search pattern.

Once the training starts the main hyperparameters to look for are bleu score(bleu),perplexity(ppl) and learning rate(Lr). BLEU, or the Bilingual Evaluation Understudy, is a score for natural language processing tasks where it evaluates the text generated. Bleu score is generated for translation.it is used for evaluating text from one natural language to another. The score of BLUE always lies between 0 and 1. BLEU has set a benchmark for any new evaluation metric. Perplexity is the predictive rate at which the model predicts the output based on the input. Another meaning of perplexity means variable probability distribution, which refers to total probability of the model. Learning rate refers to the speed at which the model learns about the input. Slow learning rate means that the model will have to go more training and more number of epochs and therefore the efficiency and accuracy is affected. Higher learning rate means that there is plenty of data and the model learns at the high rate and does not require much training before testing.

5 Result

Here are some responses given by the chatbot after running the inference using the reddit dataset.

Input from User	Reply from Chatbot
How was the weekend?	The weekend?
Would you like to come to the cinema?	Yes, Definitely.
How are you?	I'm doing great

Dataset	Perplexity	Bleu
eval dev:	39.44	11.0
eval test:	41.44	11.3

6 Conclusion

The chatbot using neural machine translation has been developed successfully. The architecture Implemented is the BRNN. This responses from the chatbot is learned form the Reddit dataset used where the dataset is of 1.5 billion comments. It took about 5 epochs for the chatbot to give an accurate response. Later this chatbot can be used for any other purposes also like for Health care which can help the patients having any diseases like Covid-19. By giving the relevant data about the Covid-19 disease. Since it's a huge amount of data to store and process we made it run using the Nvidia 1050 graphics on a Desktop. So the process was smooth and fast.

7 Future work

Deep learning techniques and Attention mechanism has been used while developing the chatbot. Reddit dataset of comments has produced good results. However, it needs further improvement by removal of vague and inappropriate comments and addition of more quality comments. This will improve its performance ratio. Deep Reinforcement Learning (DRL) can be used to produce quality results and enhance sentiment analysis. DRL approach can also be used to a broader aspect to implement a chatbot for various domains such as education, science, healthcare and banking. Subreddit data can be used as a dataset for a domain specific chatbot development.

References

[1] Dhyani, Manyu and Rajiv Kumar, "An Intelligent Chatbot using deep learning with Bidirectional RNN and Attention Model."

[2] Ali, Amir and Muhammad Zain Amin," Conversational AI Chatbot Based on Encoder-Decoder Architectures with Attention Mechanism."22 Nov 2019. https://www.researchgate.net/publication/338100972

[3] Sequence-to-Sequence learning and Neural Conversation model 2 Aug 2017. https://isaacchanghau.github.io/2017/08/02/Seq2Seq-Learning-andNeuralonversationalModel

[4] Bani, Balbir Singh and Ajay Pratap Singh , "College Enquiry Chatbot Using A.L.I.C.E ",International Journal Of Computer Sciences And Engineering ,pp.21-25 ·, 2018.

[5] Nicole Radziwill and Morgan Benton, "Evaluating Quality of Chatbots and Intelligent Conversational Agents", Computing Research Repository (CoRR) 2017

[6] Vibhor Sharma, Monika Goyal, Drishti Malik, "An Intelligent Behaviour Shown by Chatbot System", International Journal of New Technology and Research • 2017.

[7] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate ", last revised 19 May 2016 (this version, v7))

[8] A Formalization of a Simple Sequential Encoder-Decoder https://mc.ai/a-formalization-of-a-simple-sequential-encoder-decoder