Creative and Innovative Project- IT7611

Conversational AI Chatbot

Batch No: **20**

Team Members: Karthikeyan JM (2018506049), Dinesh V (2018506028), Abinaya A (2018506003)

Submitted To: Dr. MR Sumalatha

Aim: To develop a generative based conversational chatbot using Seq2Seq with LSTM Attention mechanism for society and enterprise applications.

Abstract:

Conversational AI Chatbot using Deep Learning is an interesting problem in the field of Natural Language Processing. Chatbot is a Software program that generates response based on given input to mimic human conversations in text or voice mode. These applications are designed to simulate human-human interactions. In the past, methods for constructing chatbot architectures have relied on hand-written rules and templates or simple statistical methods. With the rise of deep learning, these models were quickly replaced by end-to-end trainable neural networks. More specifically, the recurrent encoder-decoder model dominates the task of conversational modelling. Among current chatbots, many are developed using rule-based techniques, simple machine learning algorithms or retrieval-based techniques which do not generate good results. In this project, we have developed a Seq2Seq AI Chatbot using modern-day techniques. For developing Seq2Seq AI Chatbot, we have implemented encoder-decoder attention mechanism architecture. This encoder-decoder is using Recurrent Neural Network with LSTM (Long-Short-Term-Memory) cells. These conversation agents are predominately used by businesses, government organizations and non-profit organizations. They are frequently deployed by financial organizations like bank, credit card companies, businesses like online retail stores and start-ups. Their functioning can range from customer service, product suggestion, product inquiry to personal assistant. Recently, there has been a major increase in interest in the use and deployment of conversational chatbot systems. Though they are primarily questioning answering systems, their adoption by major corporations has peaked interesting customers and seems promising for more advanced conversational agent system in research and development.

Proposed Concept:

The conversational AI chatbot will be a generative model-based approach where the model generates a response, word by word based on the query. In this model we will be using RNN Encoder-Decoder/Seq2Seq LSTM model with Attention mechanism. The model will contain two LSTM-RNNs (Recurrent Neural Network), an Encoder and a Decoder. The encoder takes a sequence(sentence) as input and processes one symbol(word) at each time step. Its objective is to convert a sequence of symbols into a fixed size feature vector that encodes only the important information in the sequence while losing the unnecessary information. Each hidden state influences the next hidden state and the final hidden state can be seen as the summary of the sequence which is called the context. From the context, the decoder generates another sequence, one symbol (word) at a time. Here, at each time step, the decoder is influenced by the context and the previously generated symbols. In the first step the decoder keeps the top K words with the highest probabilities. Then at each time-step this list is expanded by computing the joint probability of the partial sequences in the list and the words in the current time-step and retaining the K most probable partial sequences until the end of the output sequence is reached. This model will be trained on Cornell Movie Dialogs Corpus. As the length of the sequence gets larger, we start losing considerable amount of information so we try to add the attention mechanism, allowing the decoder to selectively look at the input sequence while decoding.

This takes the pressure off the encoder to encode every useful information from the input. The project that we build will be made into a web application using flask web framework in python.

So with this idea behind we will try to build an intelligent conversational chatbot that will be useful to the society.

Literature Survey:

1. **Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation**:

Literature search was carried out using Google scholar, PubMed, Research gate, arXiv, IEEE journals, Elsevier journals and Scihub. The literature review in 2014, K. Cho, B. van Merrienboer, et.al described "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". The authors proposed a novel neural network architecture that can be used as a part of the conventional phrase-based SMT system called RNN Encoder-Decoder consisting of two recurrent neural networks. They qualitatively analyzed the trained RNN Encoder–Decoder by comparing its phrase scores with those given by the existing translation model. The qualitative analysis shows that the RNN Encoder–Decoder is better at capturing the linguistic regularities in the phrase table, indirectly explaining the quantitative improvements in the overall translation performance.

The goal of this literature review aims at better capturing the linguistic regularities using RNN Encoder-Decoder model.

Refernces:

K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio (2014) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation (https://arxiv.org/pdf/1406.1078.pdf)

2. **Neural Machine Translation by Jointly Learning to Align and Translate**:

Literature search was carried out using Google scholar, PubMed, Research gate, arXiv, IEEE journals, Elsevier journals and Scihub. The literature review in 2014, D. Bahdanau, K. Cho, and Y. Bengio described "Neural Machine Translation by Jointly Learning to Align and Translate". The authors showed that their proposed approach of jointly learning to align and translate achieves significantly improved translation performance over the basic encoder-decoder approach. The improvement is more apparent with longer sentences, but can be observed with sentences of any length. On the task of English-to-French translation, their proposed approach achieves, with a single model, a translation performance comparable, or close, to the conventional phrase-based system. Furthermore, qualitative analysis reveals that their proposed model finds a linguistically plausible (soft-)alignment between a source sentence and the corresponding target sentence.

The goal of this literature review aims at better linguistically plausible (soft-)alignment between a source sentence and the corresponding target sentence with their proposed model.

References:

D. Bahdanau, K. Cho, and Y. Bengio (2014) Neural Machine Translation by Jointly Learning to Align and Translate (https://arxiv.org/pdf/1409.0473.pdf)

3. **Sequence to Sequence Learning with Neural Networks**:

Literature search was carried out using Google scholar, PubMed, Research gate, arXiv, IEEE journals, Elsevier journals and Scihub. The literature review in 2014, Sutskever, O. Vinyals, and Q. Le. CoRR described “Sequence to Sequence Learning with Neural Networks”. The authors present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. They used a multilayered Long Short-TermMemory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. They obtained a BLEU score of 34.81 by directly extracting translations from an ensemble of 5 deep LSTMs (with 384M parameters and 8,000 dimensional state each) using a simple left-to-right beam-search decoder on WMT’14 English to French translation task. This is by far the best result achieved by direct translation with large neural networks. For comparison, the BLEU score of an SMT baseline on this dataset is 33.30. The 34.81 BLEU score was achieved by an LSTM with a vocabulary of 80k words, so the score was penalized whenever the reference translation contained a word not covered by these 80k. Their result shows that a relatively unoptimized small-vocabulary neural network architecture which has much room for improvement outperforms a phrase-based SMT system. Finally, they used the LSTM to rescore the publicly available 1000-best lists of the SMT baseline on the same task. By doing so, they obtained a BLEU score of 36.5, which improves the baseline by 3.2 BLEU points and is close to the previous best published result on this task (which is 37.0). Surprisingly, the LSTM did not suffer on very long sentences, despite the recent experience of other researchers with related architectures. The authors were able to do well on long sentences because they reversed the order of words in the source sentence but not the target sentences in the training and test set. They introduced many short term dependencies that made the optimization problem much simpler. As a result, SGD could learn LSTMs that had no trouble with long sentences. Their simple trick of reversing the words in the source sentence is one of the key technical contributions of their work.

The goal of this literature review aims at developing a better model for translation between a source sentence and the corresponding target sentence.

References:

I. Sutskever, O. Vinyals, and Q. Le. CoRR (2014) Sequence to Sequence Learning with Neural Networks(https://arxiv.org/pdf/1409.3215.pdf)

4. A Neural Conversational Model:

Literature search was carried out using Google scholar, PubMed, Research gate, arXiv, IEEE journals, Elsevier journals and Scihub. The literature review in 2015, O. Vinyals, and Q. Le described “A Neural Conversational Model”.The authors found that Conversational modeling approach did surprisingly well on generating fluent and accurate replies to conversations. They tested the model on chat sessions from an IT helpdesk dataset of conversations, and found that their model can sometimes track the problem and provide a useful answer to the user. They also experimented with conversations obtained from a noisy dataset of movie subtitles, and found that their model can hold a natural conversation and sometimes perform simple forms of common-sense reasoning. In both cases, the recurrent nets obtain better perplexity compared to the n-gram model and capture important long-range correlations. From a qualitative point of view, their model is sometimes able to produce natural conversations.

The goal of this literature review aims at developing a better model for producing natural conversations.

References:

O. Vinyals, and Q. Le. (2015) A Neural Conversational Model (https://arxiv.org/pdf/1506.05869.pdf)

5. Neural Network Approach to Context-Sensitive Generation of Conversational Responses:

Literature search was carried out using Google scholar, PubMed, Research gate, arXiv, IEEE journals, Elsevier journals and Scihub. The literature review in 2015, A. Sordoni, M. Galley,et.al described "Neural Network Approach to Context-Sensitive Generation of Conversational Responses".The authors proposed to address the challenge of context-sensitive response generation by using continuous representations or embeddings of words and phrases to compactly encode semantic and syntactic similarity. They argued that embedding-based models afford flexibility to model the transitions between consecutive utterances and to capture long-span dependencies in a domain where traditional word and phrase alignment is difficult. To the end, they present two simple, context-sensitive response-generation models utilizing the Recurrent Neural Network Language Model (RLM) architecture. Their models first encode past information in a hidden continuous representation, which is then decoded by the RLM to promote plausible responses that are simultaneously fluent and contextually relevant. Unlike typical complex task-oriented multi-modular dialog systems their architecture is completely data-driven and can easily be trained end-to-end using unstructured data without requiring human annotation, scripting, or automatic parsing.

The goal of this literature review aims at developing a context sensitive model for producing conversational responses.

References:

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, B. Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses.(https://arxiv.org/pdf/1506.06714.pdf)

6. Attention with Intention for a Neural Network Conversation Model:

Literature search was carried out using Google scholar, PubMed, Research gate, arXiv, IEEE journals, Elsevier journals and Scihub. The literature review in 2015, K. Yao, G. Zweig, and B. Peng described "Attention with Intention for a Neural Network Conversation Model". The authors developed a model that consists of three recurrent neural networks (RNNs). The source side RNN, or encoder network, encodes the source side inputs. The target side RNN, or decoder network, uses an attention mechanism to attend to particular words in the source side, when predicting a symbol in its response to the source side. Importantly, this attention in the target side is conditioned on the output from an intention RNN. Their model, which has the structural knowledge of the conversation process, is trained end-to-end without labels. They experimented with this model and observed that it generates natural responses to user inputs.

The goal of this literature review aims at developing a attention based conversational model for producing natural conversational responses.

References:

K. Yao, G. Zweig, and B. Peng (2015) Attention with Intention for a Neural Network Conversation Model(https://arxiv.org/pdf/1510.08565.pdf)