

MINI PROJECT, ITE398

A NEURAL NETWORK BASED CHATBOT

February 13, 2017

Submitted by

Saurabh Mathur, 14BIT0180

Under the guidance of

Prof. Daphne Lopez

Contents

| | |
|--|---|
| Abstract | 2 |
| Literature Survey | 3 |
| Hardware and Software Requirements | 4 |
| Software Architecture | 5 |
| References | 7 |

ABSTRACT

Building Conversational Agents, or Chatbots is one of the most exciting problems in Artificial Intelligence. Researchers have been trying to build and improve chatbots from as early as 1966. However, those chatbots were domain-specific and required a rule-base of hand-crafted rules. This often proves to be costly and limits access to such technology.

Machine Learning is a statistical approach at building AI Agents. The most common set of techniques is supervised learning. It allows us to build a system that can infer or, learn rules from a set of labelled examples. Recent advances in Artificial Neural Network architectures have made it possible to model sequences of variable length. Researchers have applied such architectures to the task of Language Translation with great success. Modern Graphical Processing Units (GPUs) have been found to be well suited to performing matrix operations at a large scale efficiently. This facilitates training of Neural Networks on large sets of data. Neural Network based agents can be trained end-to-end. Thus, they don't depend on hand-crafted rules.

This project is an attempt to build a Neural Conversational Model, with a sequence to sequence Neural Network architecture at its core.

LITERATURE SURVEY

Building chatbots and conversational agents has been pursued by many researchers over the last decades.

One of the earliest attempts at solving the problem was a pattern matching based system called ELIZA. It was built at the MIT Artificial Intelligence Laboratory by Joseph Weizenbaum in 1966. Since then, many similar systems have been proposed. However, most of these systems require a rather complicated processing pipeline of many stages. (Lester et al., 2004; Will, 2007; Jurafsky & Martin, 2009). This project is different from conventional systems as uses an end-to-end approach to the problem which lacks domain knowledge.

This project is mainly inspired by Oriol Vinyals and Quoc V. Le's work titled A Neural Conversational Model. They use a sequence to sequence architecture to model conversations. This architecture was proposed by Sutskever et. al. in 2014. The architecture, originally developed for the task of Statistical Neural Machine Translation, has been used for various tasks such as parsing and image caption generation.

The basic building block of the sequence to sequence architecture is the Recurrent Neural Network (RNN). It is well known that simple RNNs suffer from vanishing gradients making them hard to train. So, most implementations of the sequence to sequence architecture use Long Short Term Memory (LSTM) RNNs, proposed by Hochreiter & Schmidhuber in 1997. The LSTM avoids direct feedback connection of the simple RNN by the use of a gating mechanism.

Iulian Vlad Serban et. al. surveyed available datasets for building Data-Driven Dialog Systems. The Cornell Movie-Dialogue Corpus has over 200,000 short conversations extracted from movie dialogs.

HARDWARE AND SOFTWARE REQUIREMENTS

Training phase

Hardware requirements

- Amazon AWS g2.2xlarge
 - 60 GB SSD
 - 15 GB RAM
 - 8 CPU Cores
 - 1,536 CPU Cores with 4GB Video Memory

Software requirements

- Ubuntu Linux 16.04
- Python 2.7.12
- Tensorflow 0.12.0
- Cuda Toolkit 8.0
- cuDNN v5.1

Deployment phase

Hardware requirements

- 512 MB RAM
- 1 CPU Core

Software requirements

- Python 2.7.12
- Tensorflow 0.12.0

SOFTWARE ARCHITECTURE

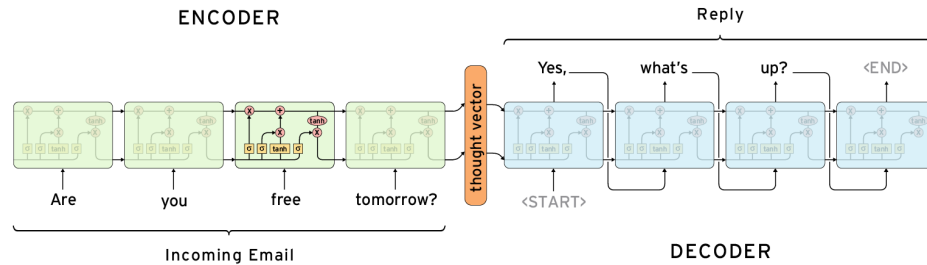


Figure 1: The Sequence to Sequence model Architecture

The project consists of the following components -

- Data Preprocessor.
 - Model: The sequence to sequence model.
 - Sampler: for generating responses
- Web based chat interface.
 - Browser based client application.
 - REST API Server.

Data Preprocessor

It uses Natural Language Processing to perform basic data cleaning. It also tags the start and end of each sentence with special tokens. It has the following methods -

- `make_conversations`: pre-processes the raw text and bins the sentences as statement and response.
- `encode_sequence`: converts a sentence to a list of word indices. The method is used to feed the data to the network in training phase.
- `decode_sequence`: converts a list of word indices to corresponding sentence.

Chatbot

This module is the core of the system. It has two sub-components -

- Sequence to Sequence model: learns a probability distribution from a large set of examples and predicts probabilities for words in response.
- Sampler: Uses the predicted word probabilities to draw a sample, thus forming the response.

Chat Interface

The chat interface consists of two subcomponents - A REST API Server that calls the chatbot model and a JavaScript based client application that makes AJAX requests to the Server.

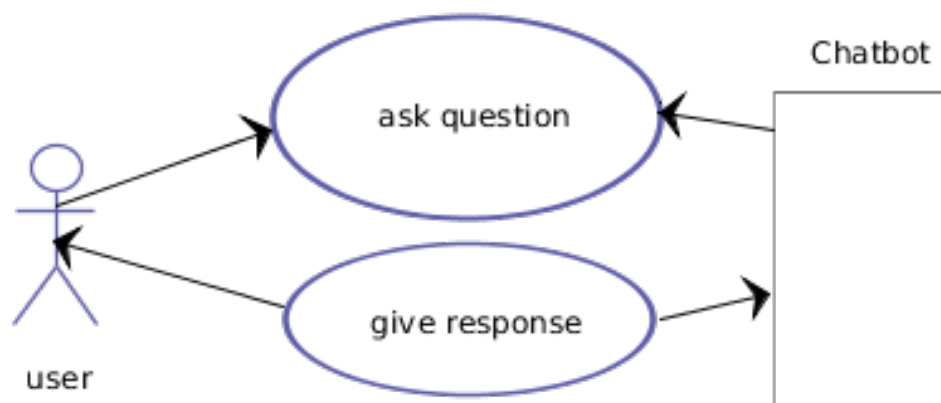


Figure 2: Use case diagram for the chat interface

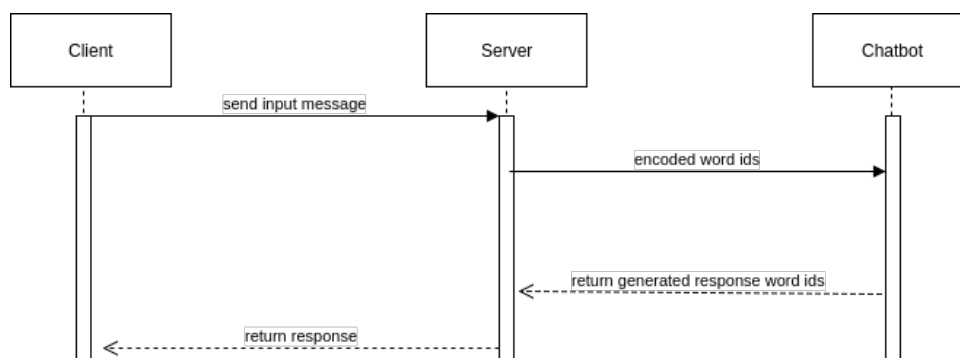


Figure 3: Sequence diagram for the chat interface

REFERENCES

Weizenbaum, J. ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine. Communications of the ACM, Vol 9, 1966.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In NIPS, 2014.

Vinyals, O., Le, Q. V. A Neural Conversational Model. In arXiv:1506.05869v3, 2015.

Lifeng Shang, Zhengdong Lu, and Hang Li. Neural Responding Machine for Short Text Conversation. In ACL-IJCNLP 2015.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Meg Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In NAACL-HLT 2015.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, Joelle Pineau Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In arxiv 1507.04808.

Jurafsky, D. and Martin, J. Speech and language processing. Pearson International, 2009.

Lester, J., Branting, K., and Mott, B. Conversational agents. In Handbook of Internet Computing. Chapman & Hall, 2004.

Will, T. Creating a Dynamic Speech Dialogue. VDM Verlag Dr, 2007.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural Computation, 1997.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, Joelle Pineau, A Survey of Available Corpora for Building Data-Driven Dialogue Systems. In arxiv 1512.05742