香 港 科 技 大 學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

NLP CHATBOT

MAFS 6010U Final Project

CHEN Xiaoyu	20651960
SUN Lei	20661604
XU Yuqing	20640234
YIN Li	20670514
ZHU Hanfeng	20644589

May 2020

CONTENT

1 Background

Chatbot is an intelligent software that can communicate with people. As we all know, in the 2020s, Chatbot is an interactive revolution and an innovation platform for multi technology integration. Specifically speaking, chatbots have a wide range of applications in customer interaction, social network marketing and instant sending information to customers because it is a new way to connect users and services. Since people are born with the idea of laziness and want to get services in the most natural and simple way, it has become a hot topic and trend. Many companies want to develop chatbots that are hard to tell true from false. Therefore, we all regard the interactive dialogue service as the entrance of many other services. In fact, most chatbots are a mixture of Natural Language Understanding (NLU) and Natural Language Generation (NLG), such as question answering robot, assistant robot and chatting robot.

In fact, chat robot technology can be divided into four types: 1. Retrieval based technology; 2. Pattern Matching based technology; 3. Natural Language Processing based technology; 4. Statistical Translation Model based technology. These technologies are not all implemented, but we choose the third technology which is based on Natural Language Processing.

Next, we will explain their respective advantages and disadvantages, so it is not difficult to guess why we choose the third one. Retrieval based technology is information retrieval technology, which is simple and easy to implement, but it could not provide answers from syntactic and semantic relations, that is, it could not solve the reasoning problem, that is why we did not choose it. The second technology based on Pattern Matching relies on several patterns that have been sorted out. This kind of reasoning problem is very simple, but the patterns we cover are not complete, so we did not choose it. And the third technology, which is based on natural language processing, is the supplement and improvement of shallow analysis, syntactic analysis and semantic analysis. Finally, the last technology that based on the statistical translation is to leave out the question words in the questions, and then match them with the candidate answer data. If you can align them, you could get the answer. However, if you cannot match them, you could not get the answer. To sum up, our group aims to explore the fundamental theory of traditional artificial intelligence in order to create our own chatbot based on NLP algorithm.

For our project, we have intended to try the chatbot with the comprehensive topic knowledge graph. Originally, we would like to focus on the third kind of chatbot in the list of the project requirements. However, due to the limited conditions of our computers, we could not bear the heavy load of deep learning operation. Therefore, we change our sight to the second kind chatbot: Educating Chatbot, which based on the simple neural network. We also studied Chatbot 1 and Chatbot 3, which will be shown in the second part.

2 Concept and Principle

2.1 LSTM

First, let's understand what is Recurrent Neural Network. RNN is a kind of neural network for processing sequence data. Compared with the general neural network, it could deal with the data of sequence change. For example, let us see the figure below: On the left of the figure is cyclic neural network structure. If we expand it, we could obtain a sequence structure. We could see that the previous output would affect the later input. This kind of chained feature reveals that RNN is essentially related to sequence. That is to say, RNN is very suitable for processing data in the kinds of voice, text and sequence.

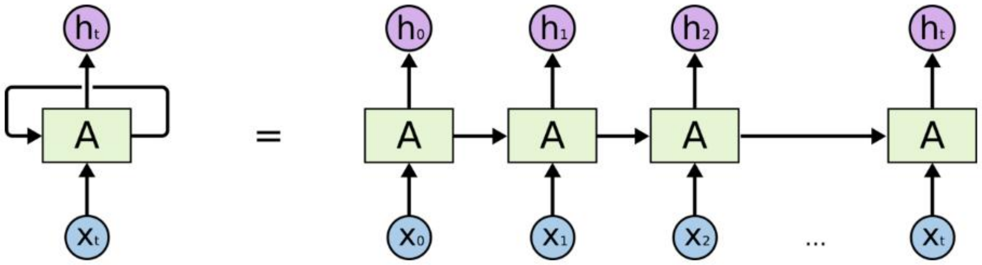


Figure 1 Recurrent Neural Network Processing

The key point of RNN is to be able to connect the previous information to the current task, such as inferring the meaning of the current statement through the previous text. However, when the interval between the related information and the current statement is too large, RNN will be difficult to learn the long-distance information, which is shown in the figure below:

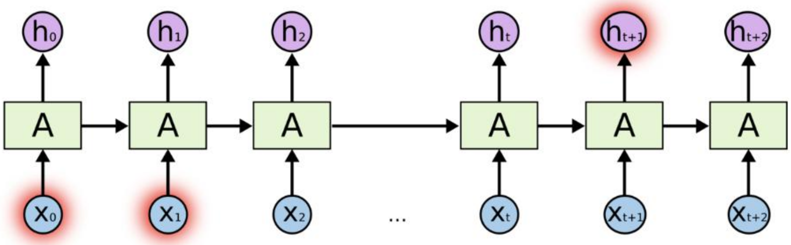


Figure 2 Key point of RNN Processing

When the long-term dependencies problem arises, we introduce a new thing——Long Short-Term memory (LSTM).

LSTM is a kind of special RNN, which is mainly used to solve the problem of gradient disappearance and gradient explosion in the process of long sequence training, that is, the problem of information loss caused by long-distance transmission. In short, LSTM can perform better in a longer sequence than RNN. And the following figure explains the differences between LSTM and RNN:

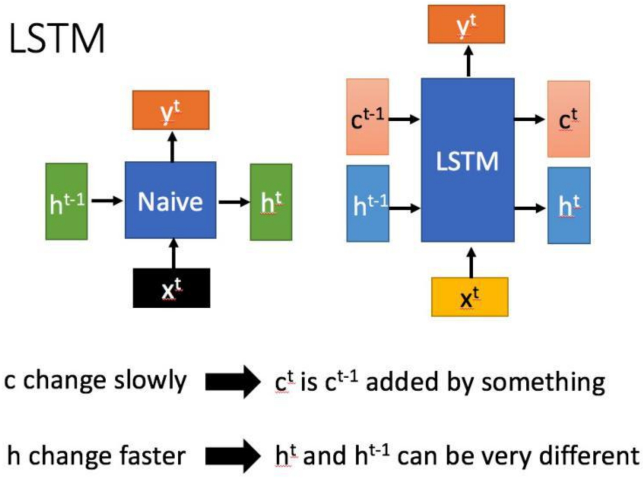


Figure 3 Long Short-Term memory

2.2 Keras and NLTK Packages

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Keras is the official high-level API of TensorFlow or a high-level neural network API. In particular Keras is born to support rapid experimentation, and can quickly transform your idea into results.

3 Introduction of the simple neural network

Our project aims to build a chatbot using deep learning techniques, which will be trained on the dataset contains categories, patterns and responses. The following packages would be used in our project: TensorFlow, Keras, Pickle, and NLTK. We use a special recurrent neural network to classify which category the user's message belongs to and then give a random response from the responses list. The dataset used is a JSON file that contains the patterns we need to find and the responses returning to the user.

Here are the common steps to create a chatbot in Python. First of all, we need to preprocess the data with Tokenizing, which is the process of breaking the whole text into small parts like words, lemmatize each word and remove duplicated words from the list. Next, to create training and testing data where the input would be the pattern and output would be the class our input pattern belongs to. Thirdly, since we have our training data ready, we now build a deep neural network that has 3 layers and apply the predicting process with suitable GUI system.

4 The structure of our own chatbot

4.1 Data and Corpus

In fact, there are many open source corpora on the Internet, some of which can be directly used in NLP model training, and more of which are unprocessed data like video subtitles, popular science websites and so on. After comprehensive consideration of various factors, we chose "Chinese local food introduction" as the main topic of our chat robot. Based on the introduction of related information on Wikipedia, we generate our own training corpus, which is stored in the file "intents1. JSON" after processing, which is a JSON file that contains the patterns we need to find and the responses we want to return to the user.

4.2 Chat Logic

In the corpus, a topic tag, labeled "Zhejiang cuisines", will be used in various forms of questions with the same meaning, such as "Zhejiang cuisines", "what is Zhejiang cuisines?", "what kind of food I could taste in Zhejiang?", as the input here. Similarly, different types of answers with the same semantics, such as "beggar's chicken", "Dongpo braised pork", "stewed spring bamboo shoots", "red stewed duck", "fried sweet and sour pork", are used as the output of responses. That is to say, we focus on transferring the questions and answers into a zero-one identity vector. If the first element of the vector is 1, it means that the question or answer belongs to the first topic tag (the vector will only have one 1, and the rest will be 0). Then the generated identity vector, where the question is x, and the topic tag corresponding to the question is y, is used to train the neural network model. In the end, we can use the trained neural network model to divide and judge the question input by the questioner on the topic tag. If the neural network model judges which topic tag it is, we will output the corresponding answer of the topic tag.

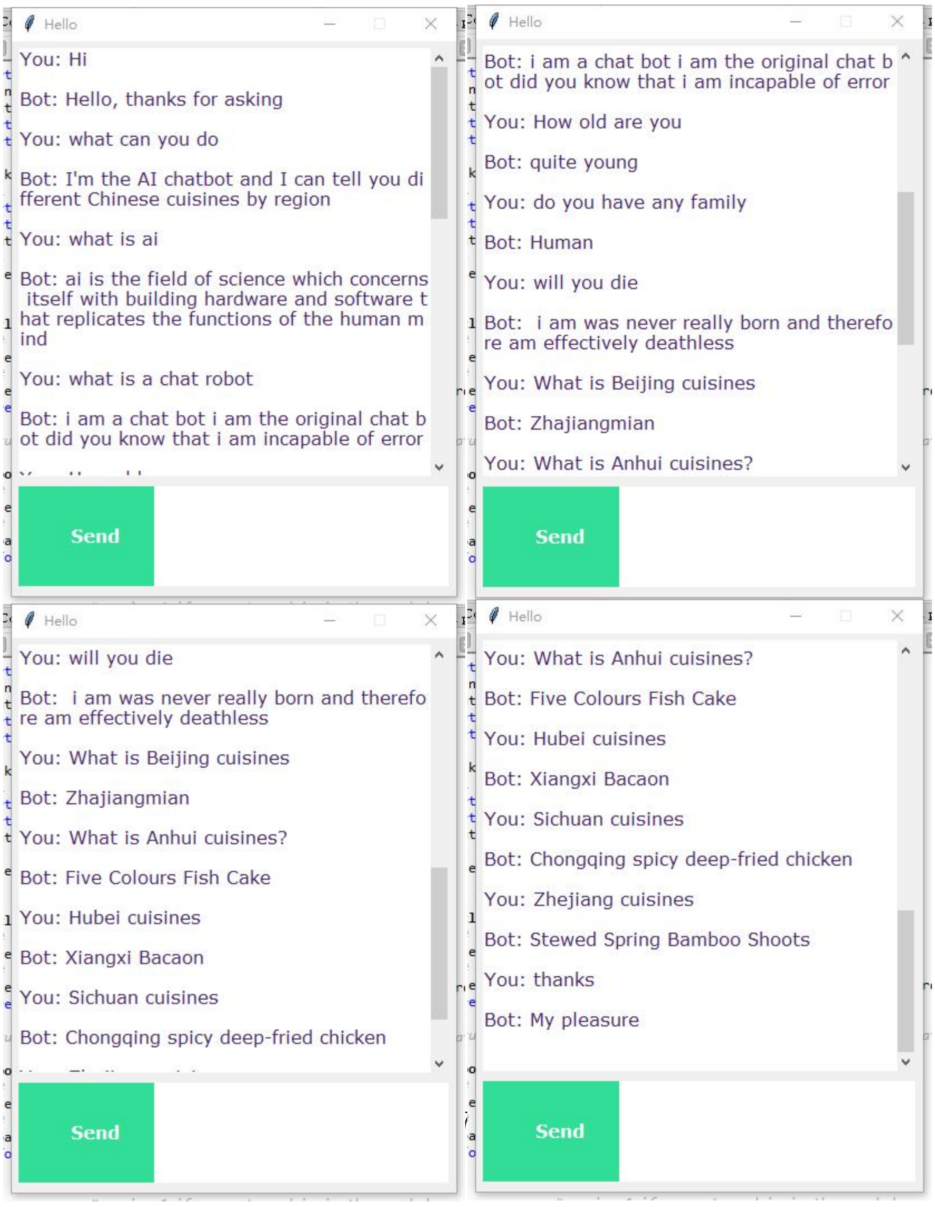
4.3 Construction Process

In this project, we tend to utilize the NLTK python package to do the natural language processing and a python package, Keras, to establish our neural network for training our chatbot model based on the stochastic gradient descent with Nesterov accelerated gradient.

Above of all, the structure of our neural network count on 3 layers with 128 neurons in the first layer, 64 neurons in the second layer and for the third output layer, its number of neurons is equal to the number of intents that predicts the output intent with soft-max.

5 Output Presentation

In this project, we have developed an understanding of the chatbots and implemented a deep learning version of a chatbot in Python. The outputs are shown as follows:



```
You: Hi

Bot: Hello, thanks for asking

You: what can you do

Bot: I'm the AI chatbot and I can tell you di
fferent Chinese cuisines by region

You: what is ai

Bot: ai is the field of science which concerns
 itself with building hardware and software t
hat replicates the functions of the human m
ind

You: what is a chat robot

Bot: i am a chat bot i am the original chat b
ot did you know that i am incapable of error
```

```
Bot: i am a chat bot i am the original chat b
ot did you know that i am incapable of error

You: How old are you

Bot: quite young

You: do you have any family

Bot: Human

You: will you die

Bot:  i am was never really born and therefo
re am effectively deathless

You: What is Beijing cuisines

Bot: Zhajiangmian

You: What is Anhui cuisines?
```

```
You: will you die

Bot:  i am was never really born and therefo
re am effectively deathless

You: What is Beijing cuisines

Bot: Zhajiangmian

You: What is Anhui cuisines?

Bot: Five Colours Fish Cake

You: Hubei cuisines

Bot: Xiangxi Bacaon

You: Sichuan cuisines

Bot: Chongqing spicy deep-fried chicken
```

```
You: What is Anhui cuisines?

Bot: Five Colours Fish Cake

You: Hubei cuisines

Bot: Xiangxi Bacaon

You: Sichuan cuisines

Bot: Chongqing spicy deep-fried chicken

You: Zhejiang cuisines

Bot: Stewed Spring Bamboo Shoots

You: thanks

Bot: My pleasure
```

Figure 4 Chatbox result

APPENDIX

Dataflair, T. (2020, Jan 07). Python Chatbot Project – Learn to build your first chatbot using NLTK & Keras. Retrieved from https://data-flair.training/blogs/python-chatbot-project

Xianjiang, S. (2019, Feb 20). The Introduction of RNN and LSTM. Retrieved from https://www.jianshu.com/p/7e6e55c48972

Video Link:
https://www.youtube.com/watch?v=Syhz17WZ9_w

Code Link:
https://github.com/SimonsChu/6010U_Chatbot