REPORT OF THE CHATBOT PROJECT

March 4, 2019

Zou Xiaohan
Tongji University
School of Softerware Engineering

Contents

1 Abstract

This report aims to briefly analyze the working principle and introduce the implementation of the chatbot which is supposed to help people to search stock and weather information.

Keywords: Artificial Intelligence, Natural Language Analysis, Chatbot, Rasa NLU, Iexfinance API

2 Introduction

Natural language analysis is one of the import areas which aims to process large amounts of natural language data, which is also one of the important technologies of chatbot. Chatbot is a computer program designed to imitate human's way of speaking when being a dialogue partner.

Chatbots can be used in many areas for various purposes. Nowadays we can find chatbots when we are accessing to some websites, when we are using some devices (Google Assistant, Apple Siri, Microsoft Cortana), or when we are using some instant messaging applications. More and more teams launch their chatbot platform or API, such as Facebook, Telegram and Wechat. Many companies such as e-commerce companies, banks and hotels can use chatbots to answer simple questions to avoid meaningless duplication of effort and improve efficiency.

Chatbots have wide application prospect, thus I take part in the class and implement a simple chatbot aims to provide people stock and weather information. This report focuses on the working principle of this chatbot and introduce how to implement it.

3 Working Principle

3.1 Intent Extraction

It is important to understand the intents. Program will understand usersâĂŹ intents by learning from the training data. At first, natural language are ransformed into word vectors to make the data readable to computers.

I implement intent recognition part based on Rasa NLU and choose scikit-learn and SVM to train the model. Other algorithms such as CNN and LSTM from deep learning are also optional (Rasa NLU has provided other intent classifiers like Tensorflow), but they need bulk training data and high performance appliance to enable high accuracy. The training data is a set of word vectors with labelled intents.

3.1.1 Nearest Neighbor Classification

Sentence from training data will be transformed to word vectors which can be processed by the program. Word vectors are real numbers which words mapped to. These data have been manually labeled into corresponding intent classifications, such as greet, query about weather and so on. Sentences from test data will also be transformed to word vectors after the same processing. Then calculate the distance between the word vectors of test data and training data. The intent of word vector of training data which is closest to the word vector of test data will be considered as the most likely intent of the test data. This method is similar to KNN, both of them calculate the distance in the same space to obtain the intent.

3.1.2 Support Vector Machine (SVM)

SVM introduces the support vector machine method of machine learning based on the above-mentioned nearest neighbor classification method to recognize intent.

Support vector machine shows a lot of advantages in dealing with small-sample and nonlinear data, and can be applied to other machine learning problems like function fitting. Its main idea is to map a linearly inseparable set of sample data to a high-dimensional space to make it linearly separable, and the kernel function can process data when the data representation of the high-dimensional space is unknown.

The machine learning algorithm which Rasa NLU uses to classify intents is the SVM with GridSearchCV.

3.2 Entity Recognition

For extracting unknown entities, keyword matching is not suitable, because it is impossible to pre-define every entity.

This part of work can also be achieved by regular expression. It is simple, but it is difficult to define a suitable regular expression.

Rasa NLU and spaCy have provided methods to extract entities, but other recognize methods such as spelling, context and after specific words also need to be used. Because similar identifies may play different roles in sentences. For example:

Tell me about the historical open data of Apple from 2018-11-1 to 2018-11-10.

In this sentence, "2018-11-1" and "2018-11-10" are both date, but one is start time, and another is end time.

Synonyms should also be considered in. For example, "Apple" and "AAPL" are the same stock.

3.3 Finite Automation

Finite automation(FA) is a mathematical model of computation which can be in exactly one of a finite number of states at any given time and change from one state to another in response to some external inputs. It can be further classified into deterministic finite automaton(DFA) and non-deterministic finite automaton(NFA).

The transition rules of the chatbot in this project can be considered as a NFA. In a NFA, an input can lead to one, more than one, or no transition for a given state. (However, a DFA only produces a unique computation of the automaton for each input string.)

FA is a quintuple $(\sigma, S, s_0, \delta, F)$:

- σ: a input alphabet (a finite, non-empty set of symbols)
- S: a finite, non-empty set of states
- s_0: a initial state, an element of S
- δ: a state-transition function: $\delta : S \times \sigma \to P(S)$, in other words, δ would return a set of states
- F: a set of final states, a subset of S

If a FA is in a state q, the next symbol is x and $\delta(q, x)$ is not defined, then this FA will reject the input and throw an exception.

3.4 Negation

Negation is a special but common method in natural language. The existence of negation increase the level of difficulty of computer's understanding of natural language.

The intent "deny" will be extracted out when extracting intent. If a sentence's intent is "deny", the program will do some specialized processing on last sentence.

3.5 Regular Expression

Regular expression is used to matches strings that conform to certain rules. When there are information with certain rules appear in sentence such as phone number and zip code, it will be convenience to use regular expression to extract these information.

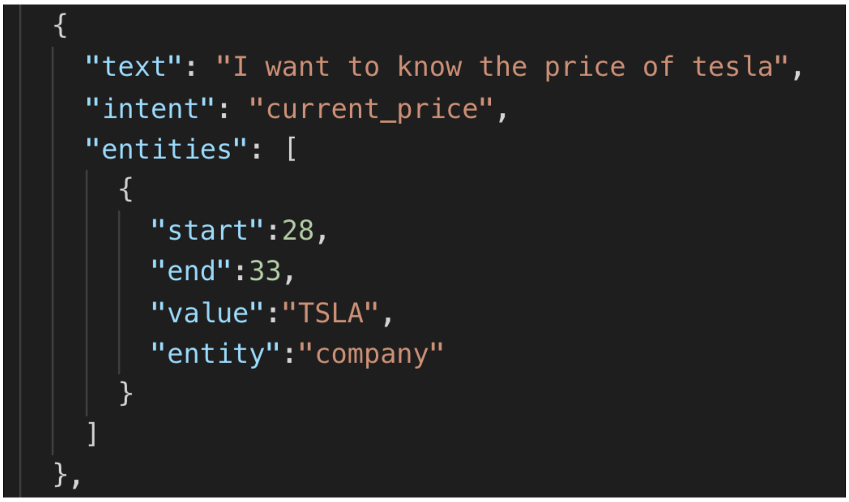
When process query about weather, this program use regular expression to extract the data.

4 Implementation

This project aims to implement a chatbot helping people to search stock and weather information.

4.1 Training Data and Model

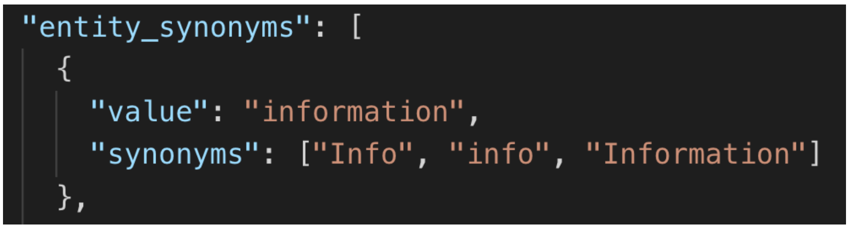
Rasa NLU is used as the main training tool. The training set need to be established first. The data type used in this project is Json, the format is:



```
{
  "text": "I want to know the price of tesla",
  "intent": "current_price",
  "entities": [
    {
      "start":28,
      "end":33,
      "value":"TSLA",
      "entity":"company"
    }
  ]
},
```

Figure 1: Data Type

The synonyms should be considered in:

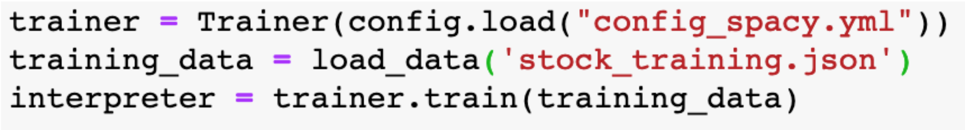


```
"entity_synonyms": [
  {
    "value": "information",
    "synonyms": ["Info", "info", "Information"]
  },
```

Figure 2: Synonyms

In theory, the training data set should as large as possible.

Then create an interpreter to extract the intent and entities of different sentences.



```
trainer = Trainer(config.load("config_spacy.yml"))
training_data = load_data('stock_training.json')
interpreter = trainer.train(training_data)
```

Figure 3: Interpreter

4.2 Policy Rules

The state rules of this chatbot(NFA) is:

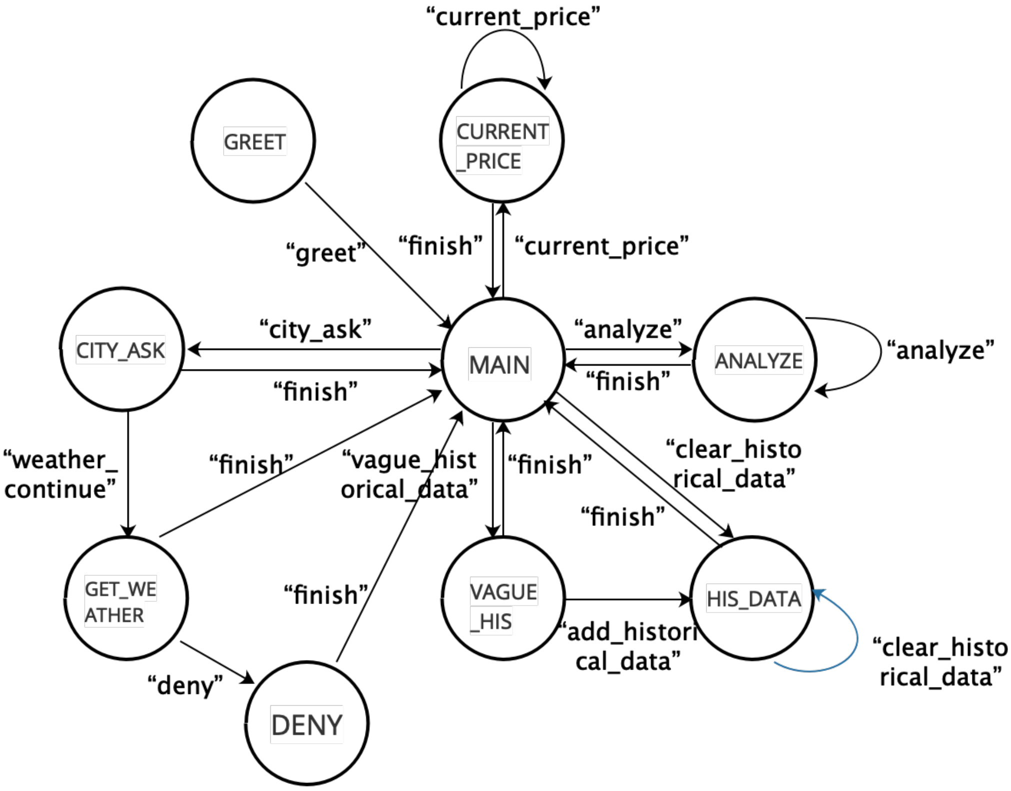


Figure 4: NFA

The chatbot will give a corresponding response to each sentence from the other side. If the chatbot receive a vague query and cannot get enough information, it will ask for additional information. Then the next step will be do the query based on the information from more than one rounds or quit this query.

4.3 Respond

It would be weird if the chatbot always respond the same sentence pattern to the other side. So I define a response group consist of two or three sentence pattern for each intent, chatbot will choose a corresponding sentence randomly to respond to others after it obtain the intent.

4.4 Extract Intent and Entities

interpreter.parse(message) can extract intent and entities of a sentence after interpreter is trained. It choose the intent with the highest confidence to be the intent of the sentence. Then the program will further process the entities to specify the role each entity playing.

4.5 API

This project use two APIs.

First is iexfinance, it can return information about stocks. It is simple to install it into the environment and import it into the project(Documentation: https://addisonlynch.github.io/iexfinance/stable/stocks.html). This project obtain the current price, historical price, recent news and TTM of a specific stock from this package.

Another is Juhe weather API, which will return information about weather in seven days. The type of the data returned is Json. A key need to be applied before use: https://www.juhe.cn/docs/api/id/39. (Documentation can also be found in this site.) This API need user to post a request with the code of city or province which user want to query about. In order to try some database operations, I establish a database consist of all provinces with their code in China. In fact this database is not necessary, you can find the API which can provide you each province and city's code in above mentioned site.

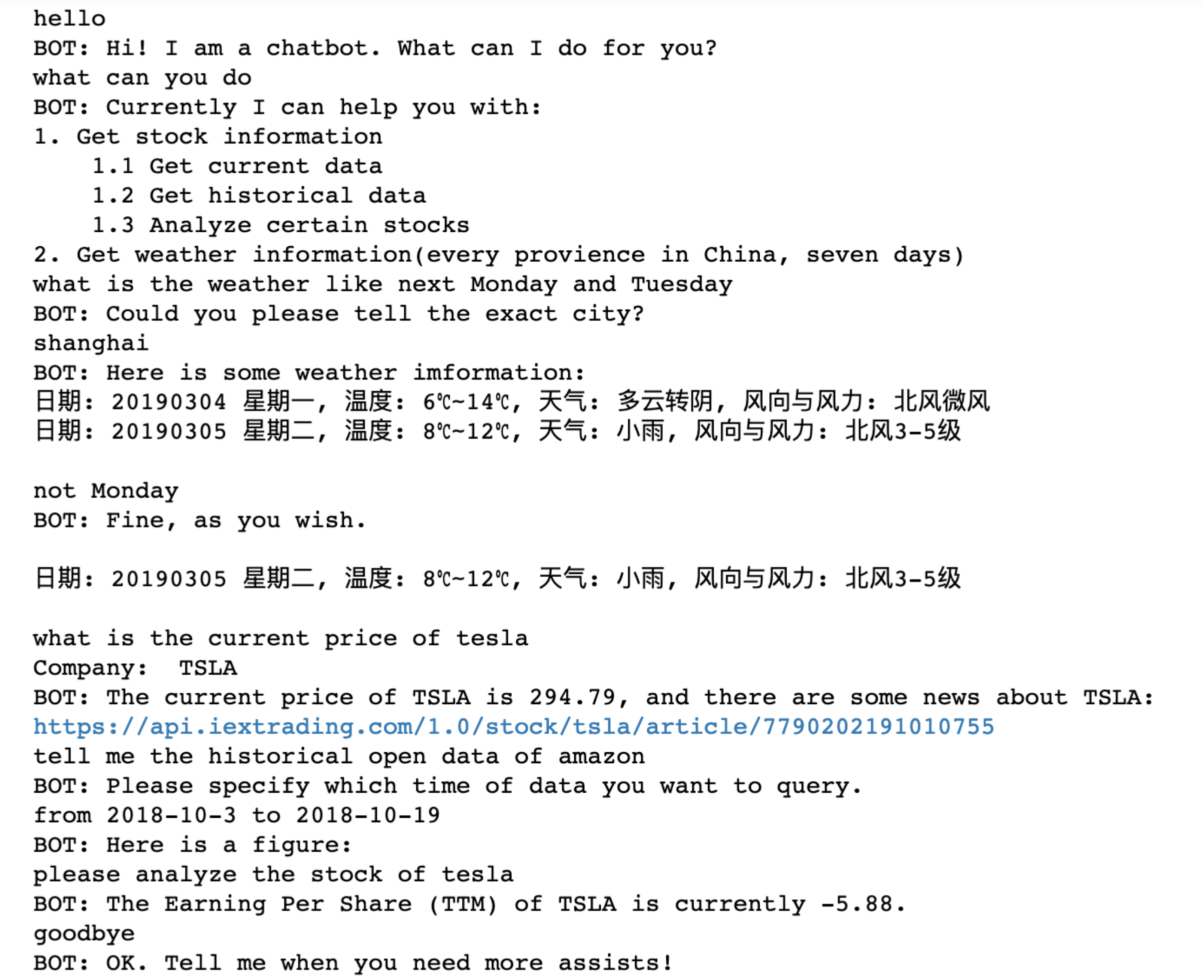
4.6 Connect to Wechat

Wechat is one of the most popular instant message applications in China. It provide a package of Python called WXPY, which can connect the program to a Wechat account so that this account can apparent to be a chatbot with auto-reply function.

Package WXPY need to be installed and import into the project first. A QR code will be generated after creating a bot. Then register this bot and it will be able to receive messages from specific friends and use the program to respond them.

5 Evaluation

A example of dialogue between use and chatbot is:



```
hello
BOT: Hi! I am a chatbot. What can I do for you?
what can you do
BOT: Currently I can help you with:
1. Get stock information
    1.1 Get current data
    1.2 Get historical data
    1.3 Analyze certain stocks
2. Get weather information(every provience in China, seven days)
what is the weather like next Monday and Tuesday
BOT: Could you please tell the exact city?
shanghai
BOT: Here is some weather imformation:
日期：20190304 星期一，温度：6℃~14℃，天气：多云转阴，风向与风力：北风微风
日期：20190305 星期二，温度：8℃~12℃，天气：小雨，风向与风力：北风3-5级

not Monday
BOT: Fine, as you wish.

日期：20190305 星期二，温度：8℃~12℃，天气：小雨，风向与风力：北风3-5级

what is the current price of tesla
Company:  TSLA
BOT: The current price of TSLA is 294.79, and there are some news about TSLA:
https://api.iextrading.com/1.0/stock/tsla/article/7790202191010755
tell me the historical open data of amazon
BOT: Please specify which time of data you want to query.
from 2018-10-3 to 2018-10-19
BOT: Here is a figure:
please analyze the stock of tesla
BOT: The Earning Per Share (TTM) of TSLA is currently -5.88.
goodbye
BOT: OK. Tell me when you need more assists!
```

Figure 5: Dialogue

This chabot can answer the queries about stocks and weather, it can also recognize the intent of "greet" and "finish" and choose suitable sentence to respond. When use's intent is "deny", it perform a fine response too. This chatbot completed all the basic functions.

However, because of the lack of trainning data and the rough process of language, there is still many limitations for this chatbot. It cannot recognizes the intent and entitise of some sentence and its responses are still a little weired.

6 Conclusion

This report analyze the working principle and introduce the implementation of the chatbot providing stock and weather information. This is a very simple chatbot, but it can still give user a lot of convenience. The result point out chatbot's great potential, we can expect it to make further increase on people's working efficiency.

But it also shows the difficult to improve the accuracy of chatbot. Some chatbot created by high technology companies such as Siri and Cortana do well in practical application, but they are still not smart enough. Companies such as Apple and Microsoft hold large data set and high performance appliance, but the performance of their chatbot still need to be improved. The large training data set is important, by not all. When we are busying with perfecting the data set and improve the parameters of exist algorithms, we should also remember to explore new methods.

I learned much from this class. Before this class, I only know deep learning method for processing natural language. I have tried to use LSTM to process data before, the performance is not so good. I learned about sklearn and SVM in this class, they are good at dealing with small-scale data. I hope I can learn more knowledge in artificial intelligence and natural language processing area.

Thanks to the help and tolerance from my mentor Zhang Fan.

Appendix

Results

- Multiple selective answers to the same question and provide a default answer;
- Answer questions through regular expressions and some other methods;
- Extract users' intent through support vector machine;
- Identify entities through pre-built training data;
- Implement a local basic chatbot system based on Rasa NLU;
- Process natural language and explore database;
- Single-round multiple incremental query for multiple times based on incremental filter and technology of recognize negative entity;
- Deal with Multiple rounds query for multiple times and handle pending state transitions and pending actions;

Source Code

URL: `https://github.com/Renovamen/StockBot`