**Final project MIPT-2018**

Zheglov Dmitry

[demonzheg@gmail.com](mailto:demonzheg@gmail.com)

**Abstract**

In this paper I’d like to tell you about constructing telegram stackoverflow assistent based on intent classifier and chatterbot model for dialogue system. This chatbot will give you link to stackoverflow connected with your question also it will talk with you about anything.

**1 Introduction**

We will construct a dialogue chat bot which will be able to:

1) answer programming-related questions (using StackOverflow dataset[1]);

2) chit-chat and simulate dialogue on all non programming-related questions

For a chit-chat mode we will use a pre-trained neural network engine available from [ChatterBot](https://github.com/gunthercox/ChatterBot).[2] Also we will try to train own models for chit-chat.

For those questions, that have programming-related intent, we will proceed as follow predict programming language (only one tag per question allowed here) and rank candidates within the tag using embeddings. For the ranking part, we will need:

word\_embeddings.tsv — word embeddings, that we will train with StarSpace.

We should obtain the following new objects that we will then use in the running bot:

1) intent\_recognizer — intent recognition model;

2) tag\_classifier — programming language classification model;

3) tfidf\_vectorizer— vectorizer used during training;

4) thread\_embeddings\_by\_tags — folder with thread embeddings, arranged by tags.

## **1 Intent and language recognition**

We want to write a bot, which will not only answer programming-related questions, but also will be able to maintain a dialogue. We would also like to detect the intent of the user from the question. So the first thing we need to do is to distinguish programming-related questions from general ones.

It would also be good to predict which programming language a particular question referees to. By doing so, we will speed up question search by a factor of the number of languages (10 here), and exercise our text classification skill a bit.

**1.1** **Intent recognition**

We will do a binary classification on TF-IDF representations of texts. Labels will be either dialogue for general questions or stackoverflow for programming-related questions. First, prepare the data for this task:

1) concatenate dialogue and stackoverflow examples into one sample

2) split it into train and test in proportion 9:1, use random\_state=0 for reproducibility

Train size = 360000, test size = 40000

3) transform it into TF-IDF features

We will train the intent recognizer using LogisticRegression on the train set with the following parameters: penalty='l2', C=10, random\_state=0.

We have test accuracy is equal to 0.99

### **1.2 Programming language classification**

We will train one more classifier for the programming-related questions. It will predict exactly one tag (programming language) and will be also based on Logistic Regression with TF-IDF features.

First, let us prepare the data for this task.

We will train the tag classifier using OneVsRestClassifier wrapper over LogisticRegression. Use the following parameters: penalty='l2', C=5, random\_state=0.

We have test accuracy = 0.78

## **2 Ranking questions with embeddings**

To find a relevant answer (a thread from StackOverflow) on a question we will use vector representations to calculate similarity between the question and existing threads. We will use weighted mean over words from sentence.

However, it would be costly to compute such a representation for all possible answers in online mode of the bot (e.g. when bot is running and answering questions from many users). This is the reason why we will create a database with pre-computed representations. These representations will be arranged by non-overlaping tags (programming languages), so that the search of the answer can be performed only within one tag each time. This will make our bot even more efficient and allow not to store all the database in RAM.

Let’s train StarSpace[3] embeddings model based on Stack Overflow[1] posts in supervised mode on the same corpus that is used in search. We can account on that these representations will allow us to find closely related answers for a question. It contains more appropriate embeddings.

For each tag we create two data structures, which will serve as online search index:

1) tag\_post\_ids — a list of post\_ids with shape. It will be needed to show the title and link to the thread;

2) tag\_vectors — a matrix where embeddings for each answer are stored.

**3 Chit chat model**

We will train seq2seq model with LuongAttention mechanism and dropout on Cornell movies dataset[4]. We will use symbol level to train our model because it allow us to reply on slang and emotion messages like “gooooood morning”. Encoder is LSTM, Decoder is also LSTM with greedy policy.

General parameters:

embedding\_size = 100, hidden\_size = 500, batch\_size = 60, n\_epochs = 10, learning\_rate = 0.001, dropout\_prob = 0.5, max\_len = 50

We have sequence loss is equal to 1.2.

X is request, Y is answer, O is output of our NN.

Some examples:

X: frank uh you know ill try to contribute but uhfre$

Y: i dont care what it costs this is when ya come to$

O: i dont know what i said i dont know what i said$^^

X: its saturday arnie you work too hard$#############

Y: carl im running late im coming right now$#########

O: i dont know what i said i dont know what i said$^^

X: its battle i want not talk$#######################

Y: but now that i am here will you speak with a woma$

O: what do you want to do to me$^^^^^^^^^^^^^^^^^^^^^

X: hold on give us half a chance to count it$########

Y: what about all the gear$##########################

O: what do you mean the same the police to the pol$^^

X: fuck off back to essex$###########################

Y: fucking mad$######################################

O: what do you mean the same the same the police to t

X: id hate to get lost on that freeway$##############

Y: they all lead to the same place the lungs$########

O: what do you mean the same the police to the pol$^^

X: you have to use the steering wheel to turn the pe$

Y: got it ejection seat$#############################

O: i dont think so$^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

X: i know it wasnt a nice thing to do but$###########

Y: yeah no shit$#####################################

O: what do you want to do to me$^^^^^^^^^^^^^^^^^^^^^

X: have i ever let you down before$##################

Y: you stand there and tell me it is my duty after a$

O: i dont know what i dont know what i dont know whe$

X: they said thats not such a good idea the shock al$

Y: give me the god damn mirror$######################

O: what do you want to say that$^^^^^^^^^^^^^^^^^^^^^

Result of chatterBot is better than our model so we will use chatterBot.

**4 Conclusion**

So we have model for intent recognition, model for tag classifier, database for ranking questions and chat model for dialogue system. Bot was run in Amazon and you could use it in telegram (@DeGoRuKa\_bot).

**References**

[1] <https://github.com/hse-aml/natural-language-processing>

[2] <https://github.com/gunthercox/ChatterBot>

[3] <https://github.com/facebookresearch/StarSpace>

[4] <https://github.com/Conchylicultor/DeepQA/raw/master/data/cornell>