1. **Methodology**

During the selection phase, we set out to learn about different types of chatbots, The chatbot is trained to use training data in the form of customer service talks with customers in several business fields. The type of business taken is the type of transportation business, more specifically in the area of ticket booking for train and plane transportation. Training data used for the learning process were 50 conversational data including greetings, congratulations, and greetings to customers.and we chose a seq2seq chatbot, to our underlying principles and improved on it to try to give it the ability to understand natural language and generate richer responses. In this section, we will elaborate on our project preparation and coding process.

**2.1 Research Approach**

Our purpose is to build a chatbot that can make them respond to our questions and generate conversation. The two main challenges are processing data and generating meaningful responses. As Python was chosen to be our programming language, the SequenceToSequence (Seq2Seq), is one of our basic methods to realize our goal. In addition, the recurrent neural network was also applied to reach the same.The training data is then used to build learning models in both the Simple RNN algorithm and the LSTM algorithm. Then the two algorithms are compared to the accuracy of the answers manually carried out by researchers. The second test is done by comparing the two algorithms regarding the learning process time and the process time to respond to the questions given.

# Research Process

After a detailed introduction to the principles and techniques we used, in this section, we will elucidate the general procedure of our experiment.

# Dataset

Corpus usually refers to the fact that it is impossible to observe large-scale language instances in statistical natural language processing. So people simply use text as a substitute, and regard the context in the text as a substitute for the context of the language in the real world. The term corpus in linguistics means a large number of texts, usually after collation, with established formats and tags. It has three notable features: The corpus stores language materials that have actually appeared in the actual use of the language. The corpus uses electronic computers as the carrier to carry the basic resources of language knowledge, but it is not equal to language knowledge. Real corpus needs to be processed (analyzed and processed) before it can become a useful resource. The content and quality of the corpus determines the ceiling that the model can finally reach. The cleaning of the corpus is also very important. It directly determines the effect of the model, and even affects the convergence of the model, unanswered questions, grammatical errors, etc., so the selection and processing of the corpus is very important . We are a dialogue robot project, so the selected corpus is mainly some QA pairs, and a lot of open data can be downloaded. The format of storage is the first behavior question, the second behavior answer, the third row is the question, and the fourth behavior answer. , And so on. Considering the quantity and quality of corpus, we choose Xiaohuangji 50w word corpus.

# Encoding

1. Config.py parameter #configuration file

The configuration of model super parameters and related file path is mainly carried out.

2. Dataprocessing.py #preprocessing files

Mainly processing the corpus, including corpus processing, encoding index, generation of the word vector file EMB of the corpus, etc.

3. Read vecor.py #to modify the word vector file

The original word vector is trained by the Wikipedia corpus word2vec. Now we need to modify the original word vector to some extent.

Mainly added PAD = '</PAD>' # padding ,UNK = '</UNK>' # Unknown ,Start = '</SOS>' # Start ,END = '</EOS>' # END.

These four word vectors are randomly generated (set to random seeds).

-wiki.zh. Text. Vector corresponds to the original word vector

-Word vec.pkl corresponds to the modified word vector

4. SequenceToSequence. Py #Seq2Seq model

5.Train. Py #training file

The operation simply needs to run this file.

6.RestfulAPI.py

Run this file, and then open index.html to start the man-machine conversation.