Research and Implementation of Seq2Seq Model Chat Robot Based on Attention Mechanism

Weiyao Luo, Di Wu

Department of Computer Science and Engineering, Northeastern University, Shenyang, 110819, China
nlp_neu1024@163.com
luoweiyao@stumail.neu.edu.cn

Abstract. With the continuous development of deep learning technology, computers can not only understand the natural language input by users, but also reply to the sentences used for input. This paper designs an intelligent chat robot based on Attention mechanism and Seq2Seq model. It uses jieba word segmentation tool and Word2vec to convert sentences in corpus into semantic vectors, and then trains and tests the model through TensorFlow platform. The experiment compares whether the Attention mechanism is added or not, and proves that the model added with Attention mechanism has better effect than the traditional Seq2Seq model, and performs better in actual dialogue.

1. Introduction

In recent years, artificial intelligence technologies including natural language processing, deep learning, speech recognition and pattern recognition have made steady progress, which has also promoted the great development of chat robots [1]. Chat robot is no longer an entertainment tool that most people think, but now it is more widely used in the fields of education, e-commerce customer service, public place service, intelligent equipment, etc. At present, there are two methods to realize chat robot dialogue system: retrieval-based dialogue system and generation-based dialogue system.

The retrieval-based dialogue system uses the knowledge base constructed in advance to return the answers corresponding to the questions with the highest similarity according to the similarity between the input questions and the sentences in the knowledge base. Based on the generative dialogue system, the relationship from question to answer is learned, so that the machine can learn a set of rules and generate replies word for word.

With the application of deep learning in NLP field, deep neural networks such as recurrent neural network (RNN) have achieved better results than traditional probability methods in practical application. This paper proposes a generative chat robot based on Seq2Seq model of Attention mechanism. In order to ensure the usability of the dialogue system, this paper adopts the small yellow chicken corpus to train the model. The generated sentences are evaluated by two evaluation methods. The results show that the model has achieved good results.
2. Seq2seq Model Based on Attention Mechanism

2.1. The form of expression of the text
In the data preprocessing phase, we need to segment the text information, because computers cannot understand human language, so we need to vectorize the words.

In this paper, we use the method of Word Embedding to map words or phrases in the vocabulary into vectors composed of real numbers.

2.2. RNN structure
For most neural networks, we all assume that the input data are not related to each other, but in the scene of chat robots, the model should consider the correlation between the data [2]. The output of the recurrent neural network depends on both the current input and the state of the previous moment. The introduction of timing feedback mechanism enables the model to effectively utilize context information. The RNN network structure is shown in the following figure:

![Fig.1 RNN structure](image)

The calculation formula is as follows:

\[ h_t = f(Wx_t + Uh_{t-1} + b) \]

Where \( W, U \) and \( B \) are the weight parameters of the model, \( h_t \) represents the hidden layer state at time \( t \), \( x_t \) represents the input at time \( t \), and \( f \) is the nonlinear activation function. In practical application, there will be some long sequences in the input of chat robots. As the final derivative will include the product of gradients in each step [3] during back propagation derivation, this will lead to the problem of gradient explosion or gradient disappearance in this network structure [4].

2.3. LSTM structure
The algorithm of recurrent neural network (RNN) is effective in dealing with time series problems, but there is a problem of gradient disappearance. Therefore, the emergence of LSTM (Long Short-Term Memory) can solve the problem that RNN inputs long text sequences, and the data at the back end of the sequence will cover up the data at the front end [5], resulting in the inability to fully express the overall information of long text sequences.
LSTM model is composed of a series of timing modules [6]. A LSTM neuron includes three gates: forgetting gate, input gate and output gate. The gates are used to process and filter information, leaving the required information behind and discarding the unnecessary information [7]. LSTM is expressed as:

\[
f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)
\]
\[
i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)
\]
\[
c_t = f_t \ast c_{t-1} + i_t \ast \text{tanh}(W_c \cdot [h_{t-1}, x_t] + b_c)
\]
\[
o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)
\]
\[
h_t = o_t \ast \text{tanh}(c_t)
\]

Where \(h_{t-1}\) is the output information of the previous unit and \(h_t\) is hidden state. \(x_t\) is the input to this cell; \(W_f, W_i, W_c, W_o\) are weight matrices of LSTM; \(b_f, b_i, b_c, b_o\) are the biases of the LSTM; \(f_t\) represents the output of the forgetting gate and \(i_t\) represents the output of the input gate. \(\sigma(\cdot)\) represents the activation function sigmoid.

2.4. Seq2seq Model
In RNN structure, the input sequence and the output sequence are required to be equal in length, but in actual application, the lengths of the input sequence and the output sequence are often different. Such as machine translation, dialogue generation and other fields. Therefore, Seq2Seq model was proposed. The model was initially used in machine translation tasks and achieved more results than the traditional machine translation model. Therefore, the Seq2Seq model is introduced into the chat robot [8].

The Seq2Seq model is also called the Encoder-Decoder model. The encoder module encodes the input sequence into a fixed-length context vector \(C\), while the decoder is responsible for decoding the context vector \(C\) into the target sequence.

For Seq2Seq model, Encoder and Decoder parts can adopt a variety of different neural network structures, and often choose different neural network models according to different model characteristics.
In the data set, the data is represented by \( \langle H, Y \rangle \), where \( H \) represents the user's question and \( Y \) represents the corresponding reply, so there are:

\[
H = \langle h_1, h_2, h_3, ..., h_t \rangle \\
Y = \langle y_1, y_2, y_3, ..., y_k \rangle
\]

Where \( t \) and \( k \) represent the maximum length of question \( H \) and reply \( Y \) respectively.

(1) Encoder
The encoder converts the input sequence \( H \) of indefinite length into the context vector \( C \) of fixed dimension through a series of nonlinear transformations.

(2) Decoder
When decoding, the output sequence \( y_2, ..., y_{t-1} \) and the context vector \( C \) are used to predict the next output word \( y_k \).

\[
y_k = \arg \max P(y_k) = \prod_{k=1}^{T} P(y_k | y_1, y_2, ..., y_{k-1}, C)
\]

(3) Output Results
According to the maximum likelihood estimation, the conditional probability of the output sequence in a given input sequence \( H \) is maximized, and the loss of the output sequence \( Y \) is obtained at the same time.

\[
P(Y | X) = \prod_{k=1}^{N} P(y_k | y_1, y_2, ..., y_{k-1})
\]

\[
\mathcal{L} = -\frac{1}{n} \log P(y_1, ..., y_m | h_1, ..., h_t) = -\frac{1}{n} \sum_{t=1}^{n} \log P(y_m | y_1, ..., y_{m-1}, C)
\]

In order to determine the specific output at this time, we need to map the output of the decoder to the dimension of the vocabulary and use the Softmax function to calculate the final output, so that the probability distribution of each word can be calculated. Usually, the word with the highest frequency is selected as the output at that time.

However, this Seq2Seq model still has some defects. First, the process of the encoder converting the input sequence into the context vector \( C \) is still a lossy compression, which inevitably leads to semantic loss for information with high compression ratio. Second, the word information at the back end of the input sequence will be covered up by the information at the front end, and the transmission of information is biased, making the information in the sequence more "important". In order to solve these problems, we introduce the attention mechanism into the model.

2.5. Attention mechanism
Attention mechanism simulates the characteristics of human brain attention. The core idea is to allocate more attention to important content and less attention to other parts. Attention mechanism is often applied to sequential learning tasks, such as problems related to natural language processing. For the Seq2Seq model, there are Encoder and Decoder modules, and Attention mechanism can be applied to both modules. Adding Attention mechanism to Encoder module will assign weights to input data. Similarly, adding Attention model to Decoder module will assign weights to output data, which can improve the training effect of Seq2Seq model.

Encoder and Decoder in the traditional Seq2Seq model exchange data through an intermediate semantic vector, and the length of this semantic vector is fixed, which will bring the problem of long-distance dependence to the model, that is, for long sequence text, as the input progresses, the information in the back part of the sequence will cover the information in the front part of the sequence.

Attention by retaining part of Encoder's output results to the input sequence, such as \( c_2 \) in Fig. 4, the training model selectively learns these output results and will eventually be associated with the corresponding output sequence [9]. In other words, the output and input will be selectively related. An abstract diagram of the Attention mechanism based on Seq2Seq is shown in the following figure:
Fig. 4 Seq2Seq model with Attention

Where $s_i$ is the hidden layer state of RNN at time $i$, and the calculation formula of $s_i$ is:

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

For each output $h_l$, different weights are introduced here:

$$c_l = \sum_{j=i}^{T_x} a_{ij} h_l$$

The calculation formula of $\alpha_{ij}$ in $h_l$ in each hidden layer is as follows:

$$\alpha_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{T_y} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_l)$$

The function $a$ has several implementation directions, and the following is one of them.

$$e_{ij} = v^T \tanh(W_s s_{i-1} + W_h h_l)$$

Among them, $v$, $W_s$, $W_h$, various weights, offset terms and embedding layer parameters in the two recurrent neural networks of encoder and decoder are all model parameters that need to be learned at the same time.

3. Test Results and Discussions

3.1. Data set and data preprocessing

In order to verify the model in this paper, the team adopted an open corpus of yellow chickens, covering real dialogue pairs (including question and answer) in different fields.

In this paper, jieba word segmentation tool is used to segment question and answer sentences in the corpus [9], and the word segmentation results are sorted, numbered sequentially according to the word frequency from high to low, and the threshold value of minimum word frequency is set at the same time. Only words reaching the threshold value of minimum word frequency will be stored in the word list.

TensorFlow requires the input parameter encoder_inputs to be in a specific format, so it needs to start with GOID, end with EOS_ID, and be filled with a null value with PAD_ID.

Finally, the word vector training is carried out on the processed corpus by using the Word2Vec method inside the natural language processing tool gensim[10].

3.2. Setting of experimental parameters

For the main parameters in the model, we have made the following settings: the initialization learning rate is 0.1, the input sequence length and output sequence length are 25, the minimum frequency is 1, the number of LSTM neurons is set to 8, and the number of training (epoch) is 10000 times.
3.3. Experimental results and analysis

In the process of training, we mark the corpus in two ways, taking [QUESTION] as the input of Encoder and [ANSWER] as the input of Decoder. However, during the actual operation of the chat robot, there is no [ANSWER] tag, because at this time the output of Encoder is the input of Decoder.

Through experiments, this paper records the loss decrease of Seq2Seq model and Seq2Seq model with Attention mechanism [11], as shown in Figure 5 and Figure 6 respectively.

As can be seen from the figure, the loss of the model with Attention mechanism decreases faster and even lower when iterating 10,000 times at the same time. Although the loss suddenly increases during the training process, the overall loss decrease degree is still very ideal.

In addition, in order to evaluate the quality of the model, we introduce two automatic evaluation mechanisms: BLEUs [3] and Distinct-n. BLEUs index is derived from the evaluation index in machine translation. It expresses the quality of the reply by judging the lexical overlap between the reply sentence and the reference reply sentence. However, Distinct-n measures diversity by judging the number of difference-ngram and the size of entropy value.

The value range of BLEUs is from 0 to 1, and the closer to 1, the better.

| Evaluation method        | Seq2Seq  | Seq2Seq+Attention |
|--------------------------|----------|-------------------|
| BLEUs                    | 0.359    | 0.428             |
| Distinct-n               | 0.015    | 0.032             |

From the experimental results, it is concluded that the two indexes have been improved after the introduction of attention mechanism, of which BLEUs index has been slightly improved, while Distinct-n index has been greatly improved, because BLEU index pays attention to the lexical repetition rate between sentences, and the sentences generated after adding attention have higher matching degree with the sentences in the corpus. On the other hand, Distinct-n is used to evaluate the diversity of statements. After adding the attention mechanism, the diversity of statements has been significantly improved.
Judging from the training loss change diagram and the actual operation effect of the chat robot model, the chat robot model based on attention mechanism proposed in this paper can better understand the problems and produce more diversified answers.

| Form.2 Some experimental results |
|----------------------------------|
| **Question** | Answer(Seq2Seq) | Answer(Seq2Seq+Attention) |
| Hello | Hello | Hello ~ |
| Who are you | I'm Xiao Xi | My name is Xiao Xi |
| Does it rain at night | It may rain | It may rain |
| I hope my efforts will pay off | I'm sorry. | I also hope |
| It seems that you replied automatically | I'm sorry. | No, it's not |
| What is this reply | What's the matter | This is a reply from Xiao Xi. |

4. Conclusion
This paper proposes a chat robot implemented by Seq2Seq model based on Attention mechanism. After preprocessing, QUESTION and ANSWER in the corpus are converted into semantic vectors, and Attention mechanism is introduced to optimize the model by assigning weights to sentence sequences. Finally, the feasibility and effectiveness of the model are verified by experiments and comparison with chat robots implemented by Seq2Seq model without Attention mechanism.

In this paper, the Seq2Seq model based on LSTM is used as the basis for research. In recent years, with the development of reinforcement learning and GAN, some scholars are trying to solve the problem of reply generation by generating countermeasure network GAN. We would like to see that through this technology, the intelligence of chatbots can be enhanced and the interest of replies can be improved.

Acknowledgments
This work S202010145171 was supported by the National Training Program of Innovation and Entrepreneurship for Undergraduates and supported by the Fundamental Research Funds for Central Universities (N182410001).

References
[1] Turing A M. Computing machinery and intelligence[J]. Mind, 1950, 59(236): 433-460.
[2] Lei Yang. Research and Implementation of Chatbot Based on Deep Learning [D]. Beijing University of Posts and Telecommunications,2019(in Chinese).
[3] Sepp Hochreiter, J&uuml, rgen Schmidhuber. Long Short-Term Memory. 1997, 9(8):1735-1780.
[4] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult.. 1994, 5(2):157-66.
[5] Weizhen Wu. Research on Chat Robot Dialogue Based on Seq2seq Model[D]. Nanjing University of Posts and Telecommunications,2019(in Chinese).
[6] Lu Cheng. A Study of Two-way LSTM Model Based on Attention Mechanism in Emotion Classification of Chinese Commodity Comments [J]. Software Engineering,2017,20(11):4-6+3(in Chinese).
[7] Pennington J, Socher R, Manning C D. Glove:Global Vectors for Word Representation[C]EMNLP.2014, 14:1532-1543
[8] Sutskever I, Vinyals O, Le Q V, et al. Sequence to Sequence Learning with Neural Networks[J]. neural information processing systems, 2014: 3104-3112.
[9] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: data mining, inference and prediction. Springer-V erlag,2001.
[10] Qianmin Wang, Yin Li. Research on Personalized Chat Robot Based on Deep Learning [J]. Computer Technology and Development,2020,30(04):79-84(in Chinese).
[11] Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large- Scale Machine Learning on Heterogeneous Distributed Systems. arXiv: Distributed, Parallel, and Cluster Computing, 2015.
[12] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Philadelphia: Association for Computational Linguistics, 2002: 311-318.