Open Access Library Journal
2019, Volume 6, e5817
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Car FAQ Assistant Based on BILSTM-Siamese Network

Bo Jin, Zhezhi Jin

Faculty of Science, Yanbian University, Yanji, China
Email: 1458178828@qq.com, jinzhezhi@sina.com

How to cite this paper: Jin, B. and Jin, Z.Z. (2019) Car FAQ Assistant Based on BiLSTM-Siamese Network. *Open Access Library Journal*, **6**: e5817.
https://doi.org/10.4236/oalib.1105817

Received: September 25, 2019
Accepted: October 6, 2019
Published: October 9, 2019

Abstract

With the development of artificial intelligence, the automatic question answering technology has been paid more and more attention. Along with the technology's maturing, the problem is gradually exposed. This technique has several major difficulties. There is semantic recognition. How accurate are the answers to the questions? Because Chinese grammar is relatively complex. There are many different ways of saying the same thing. There is a powerful challenge to semantic recognition. Due to poor semantic recognition, the accuracy of the corresponding questions and answers is low. Aiming at this kind of questions, we build a Q & A library for common use of cars. Firstly, TFIDF method is used to build a basic FAQ retrieval system. Then we use BILSTM-siamese network to construct a semantic similarity model. The accuracy rate on the test set was 99.52%. The final FAQ system: first retrieves 30 most similar questions using the TFIDF model, then uses BILSTM-siamese network matching and returns the answer of the most similar question. The system has a lot to be improved. For example, how to combine this system with speech recognition? How to realize speech recognition will be a great challenge.

Subject Areas

Computer and Network Security

Keywords

TFIDF, BILSTM-Siamese, FAQ

1. Introduction

With the gradual improvement of people's living standards, cars have gradually entered our door to door. There will also be a variety of confusion. When we have a problem, we usually open "Baidu". However, we live in an era of rapid

development of the Internet, and there is more and more complicated information, so it is difficult to find really effective and valuable information among them. Therefore, the traditional way of searching information can no longer meet people's needs [1]. In the past two years, with the development of artificial intelligence, intelligent questions and answers have gradually received attention. We built an intelligent question and answer system, which can give effective answers to questions in mandarin. It greatly facilitates our life. The most difficult problem in intelligent question answering is to identify the meaning of sentences. For example, many different sentences express the same problem. How to identify them and then accurately correspond to the answers. It is one of the criteria for testing the quality of this question and answer system [2].

The automatic question answering system began to be studied by people in the 1960s. In the 1990s, with the development of natural language processing technology and the application of semantic information, the automatic question answering system has been greatly improved, from about 30% in the past to more than 50%. In 1993, MIT released START on the Internet, which greatly improved the number and types of questions answered, and made a staged breakthrough in the automatic question answering system technology [3]. Other countries have also begun to invest in research, including the CLEF cross-language question and answer system and Japan's NTCIR [4]. There's also a question and answer system for turby. Its name is FREYA. It is specifically designed to train semantic models. The development of our country's intelligent question-answering system began in 1970. China's first human-computer dialogue system was invented in 1980 by the Research Institute of the Chinese Academy of Sciences. Intelligent question-and-answer technology is becoming more and more mature. It gradually replaces human customer service, liberates a large amount of labor, and can solve people's problems all day long.

2. Construct Corpus

We collect the problems that car owners often encounter during their use. And expand it to build a question-answer corpus. The same Id is the same semantic question, and the first one is the standard question. There are 123 common questions in the corpus. The structure is shown in **Table 1**.

Table 1. Question and answer table.

id	question	answer
Q001	How to adjust the seat	Pull up the right seat position adjustment lever and slide the seat back and forth to …
Q001	How to adjust the position	Pull up the right seat position adjustment lever and slide the seat back and forth to …
Q001	How to adjust the chair	Pull up the right seat position adjustment lever and slide the seat back and forth to …
…	…	…
Q123	How to turn on the fog light	Please turn the fog light dial on the left side of the steering wheel to on, or …
Q123	How to open the fog lights	Please turn the fog light dial on the left side of the steering wheel to on, or …

3. TFIDF-Based FAQ System

TFIDF Value

TF-IDF Full name is Term Frequency/Inverse Document Frequency. In 1973 Salton [5] came up with TFIDF. It is a word frequency-inverse document frequency algorithm and is a statistical method. The number of times a given word appears in a document, we define it as the word frequency. In how many documents a word appears in, we define it as a reverse word frequency. Then multiplying the two parts of the word frequency is the TFIDF value [6].

$$TF_{ij} = \frac{n_{ij}}{N_j}$$

$$IDF_{ij} = \log \frac{N}{N_i}$$

$$\omega_{ij} = TF_{ij} * IDF_{ij} \quad [7]$$

We can use "jieba" word segmentation to solve the problem of corpus, and then extract the keywords and part of speech tagging, and then statistical word frequency and reverse word frequency, calculate the TFIDF value. Its value can be approximated to a probability between 0 and 1. Arrange the value of TFIDF from small to large, the larger the value is, the more important it is. We can make the problem the row of the matrix and the keyword the column of the matrix. Form a TFIDF matrix and save it locally. When the user enters a question, the system participles the question and calculates the TFIDF. This value is mapped to the vector space, and then the inner product space cosine is applied to calculate their similarity with the previously saved TFIDF matrix locally. Eventually we return the answer to the most similar question to the user. The structural framework is shown below (**Figure 1**) [8].

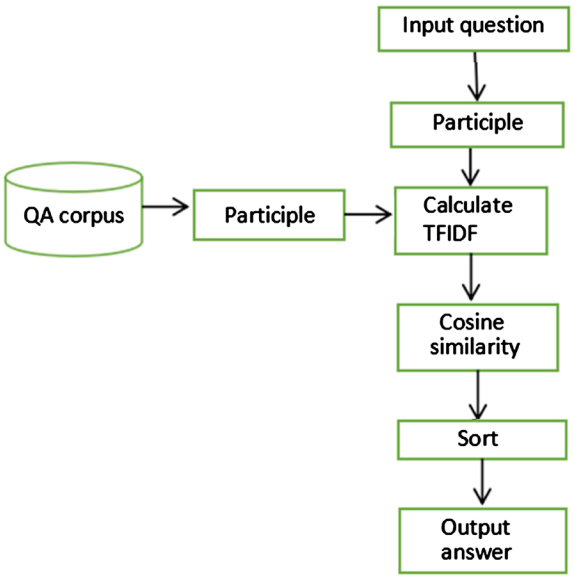

Figure 1. System framework.

4. Semantic Similarity Model Based on BILSTM-Siamese

4.1. Model Architecture

We convert the problem pair (Q1, Q2) into a list of characters of length 15. According to previous research scholars, character-based models perform better. So here, we use a pre-trained 100-dimensional character vector to convert the Q1, Q2 character list into a character vector matrix. Use the BILSTM-32 network to extract the timing characteristics of the statement, then connect the neural node to the fully connected layer of 64 and output the advanced features. Finally, we use cosine similarity to calculate the similarity of advanced features. [9] In the entire model, the left subnetwork shares the weight parameters with the right subnetwork. Therefore, when Q1 is exactly the same as Q2, the model output is 1. The result is shown in Figure 2.

4.2. Sample Construction

1) Under the same problem, different questions are combined between two pairs to form a synonymous statement pair (positive sample set);

2) Using 123 standard problem sets, use tfidf to retrieve the k problems that are closest to each problem (removing themselves), and compose non-synonymous statement pairs (negative sample).

4.3. Positive Sample Expansion

1) Commonly used consultation words to expand, how to swear, how; where → where, where, where, what position; why → how, oh, for example: How to adjust the seat? Expanded to:

(How to adjust the seat? How to adjust the seat?)

Where is the speedometer? Expanded to:

(Where is the speedometer? Where is the speedometer? Where is the speedometer? Where is the speedometer? Where is the speedometer?)

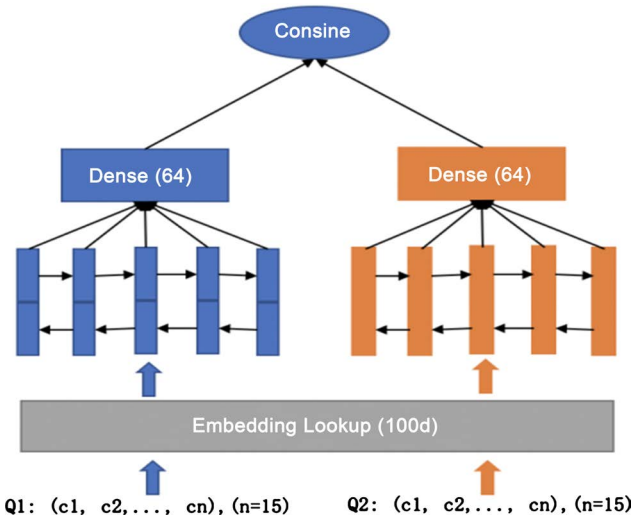

Figure 2. Model framework.

Why is the warning light on? Expanded to:

(Why is the warning light on?, how does the warning light light up? Is the warning light on?)

2) Randomly exchange the position of two words. E.g.:

How to adjust the seat? Expanded to:

(How to adjust the seat, how to adjust the seat, how to adjust the seat, …)

4.4. Negative Sample Expansion

The negative sample is expanded by the unexpanded positive sample, that is, if q1 and q2 are negative, the positive examples of q1 and q2 can form a negative example.

The sample set built is shown in Table 2.

The constructed sample set is divided into a training set and a test set according to a ratio of 7:3, as shown in Table 3.

5. Experiment

The model training batch size is 128, and the epoch (iteration period) is 30. Each layer of the model is processed by dropout, which speeds up the training iteration and prevents over-fitting.

The performance of the model on the test set is shown in Table 4.

Table 2. Expand problem.

Number	Q1	Q2	Label
1	How to adjust the seat	How to adjust the position	1
2	How to adjust the seat	How to adjust the position	1
3	How to adjust the seat	How to adjust the seat	1
…	…	…	…
95,404	How to open the fog lights	How to raise the seat a little	0
95,405	How to open the fog lights	How to raise the seat a little	0
95,406	How to open the fog lights	How to raise the seat a little	0

Table 3. Sample collection.

	0/Semantic difference	1/Same semantics	Total
Training set	46,961	19,823	66,784
Test set	20,032	8590	28,622
Total	66,993	28,413	95,406

Table 4. Performance table.

Real value\predicted value	0	1
0	19,894	138
1	0	8590

Accuracy (precision): (19,894 + 8590)/28,622 = 99.52%.

Recall rate (recall): 8590/8590 = 100%.

F1:2 * precision * recall/(precision + recall) = 99.20%.

6. FAQ System Structure

Use the data in Figure 1 to build a QA corpus. As shown in Figure 3, for a user input, we first use the TFIDF to recall k questions as candidate sets, and then use the BILSTM-Siamese semantic similarity model constructed above to calculate the semantic similarity between the input and k candidate sets. Eventually return the answer to the semantically most similar candidate question.

We feed the problem into the system. Run according to the program in our system. Can quickly output the most similar answer. The corresponding accuracy rate is high, as shown in Figure 4.

7. Summarize

This paper introduces the automatic question answering system model in the

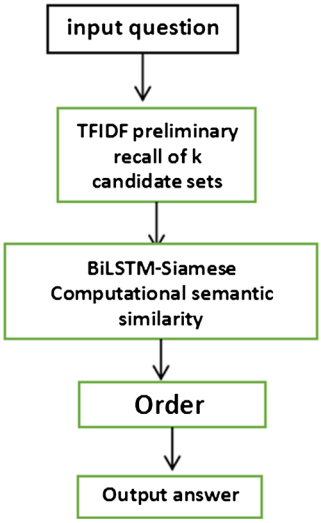

Figure 3. System framework.

```
>>>Hello
Hello
>>>How to adjust the seat
Pull up the right seat position adjustment lever, slide the seat back and forth to the desired position, and release the adjustment lever.
>>>Where is the fire extinguisher?
Dry powder fire extinguisher on the front row member seat bracket
>>>How long does it take to change the filter
Warranty period: 36 months or 80,000 kilometers
>>>How to adjust the seat backwards
Pull up the right rear seat position adjustment lever, slide the seat back to the desired position, and release the adjustment lever.
>>>
```

Figure 4. Actual presentation.

field of mathematical computing, and the specific operation process improves the accuracy of answering questions. In this paper, there are still some problems in expanding the vocabulary of automobile proper nouns and solving various names of the same parts. The characteristics of interaction between the two sentences of BILSTM are not perfect. These are the places that need to be improved next.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Guo, H. (2008) Semantic-Based Online Book Automatic Question Answering System Research. Taiyuan University of Technology, Taiyuan.

[2] Dong, Z.T., Bao, Y.Q. and Ma, X.H. (2010) Method for Calculating Similarity of Question Sentences in Intelligent Question Answering System. *Journal of Wuhan University of Technology* (*Information and Management Engineering Edition*), **32**, 31-34.

[3] Shi, J.X. (2010) Research on Automatic Question and Answer Technology Based on FAQ. Nankai University, Tianjin.

[4] Sen, Z., Wang, B. and Jones, G.J.F. (2007) ICT-DCU Question Answering Task at NTCIR-6. *Proceedings of the 6th Ntcir Workshop on Research in Information Access Technologies*, Tokyo, 15-18 May 2007, 154-167.

[5] Salton, G. and Clement, T.Y. (1973) On the Construction of Effective Vocabularies for Information Retrieval. *Proceedings of the* 1973 *Meeting on Programming Languages and Information Retrieval*, Gaithersburg, 4-6 November 1973, 48-60. https://doi.org/10.1145/951762.951766

[6] Liu, Z. (2019) Application of TFIDF Algorithm in Article Recommendation System. *Computer Knowledge and Technology*, **15**, 17-20.

[7] Yang, Y., Yan, D.B., Xu, M., Wan, L., Li, Q. and Qiu, D. (2019) Classification of 95598 Complaint Work Order Based on Improved TFIDF Feature Weighting Algorithm. *Electricity and Energy*, **40**, 205-207+226.

[8] Sun, H.F. and Hou, W. (2014) Application Research of Improved TFIDF Algorithm in Potential Cooperative Relationship Mining. *Modern Book Information Technology*, No. 10, 84-92.

[9] Lin, S.M., Chen, T.Y. and Liang, W. (2019) DGA Domain Name Detection Method Based on BILSTM Neural Network. Network *Security Technology and Application*, No. 1, 15-17.