Research and Design of Chinese Question Answer System for Heart Disease

Hao XU\textsuperscript{1,2,a}, Rui QIAO\textsuperscript{1,2,b,*} and Xiao-lei ZHONG\textsuperscript{3,c}

\textsuperscript{1}College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei, China

\textsuperscript{2}Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, Hubei, China

\textsuperscript{3}Yunnan University Dianchi College, Kunming, Yunnan, China

\textsuperscript{a}2605243601@qq.com, \textsuperscript{b}qiaorui@wust.edu.cn, \textsuperscript{c}664749849@qq.com

*Corresponding author

Keywords: Question answering system, Deep learning, Heart disease.

Abstract. This article is aimed at the current common users to obtain information on heart disease online, citing the current popular question and answer system. This article introduces a reading comprehension model to find answers that match the questions. In order to make the reading comprehension model have better application in the question answering system, the following work is done in this article: (1) Quoted word vectors are used to process Chinese text. Since Chinese is more complicated than English, quoted word vectors are currently a better method for processing Chinese text. (2) Construct a question answering system, which can be used to get answers to questions in a paragraph.

Introduction

Research Background and Significance

Generally speaking, after reading an article, [1,2,3] people will form a certain impression in their minds. The research of machine reading comprehension is to give computers their minds. The research of machine reading comprehension is to give computers the same reading ability as humans. This ability for humans is easy for humans. This is not the case. Machine reading comprehension is the way to solve this problem.

Machine reading comprehension (MRC) refers to the automatic and unsupervised understanding of texts; giving computers the ability to acquire knowledge and answer questions through text data. People believe that this is to build general intelligence A critical step in the body.1.2 Status of research at home and abroad

Around the 1990s, with the advent of advanced search technology, there has been a trend to replace search-based question answering systems. The basic principle of the search-type question answer is to cut the question, then extract the keywords with higher weights, perform full-text search in a large text database, and keep a list of descending documents with the highest relevance to the question. Semantic information of the question key, get the answer corresponding to the question in the content of the document list. Knowledge base-based question answering system (KB-QA) began to appear, that is, for a given natural language problem, some language modeling techniques were used to semantically analyze the problem, so that the back-end knowledge base can be used Semantic matching and reasoning get the answer.

Status of Domestic Research

The Chinese question answering system is still very different from foreign English systems[4]. The most direct is the difference in the corpus used. Chinese and English are very different in grammar. The characteristics and difficulties of Chinese information processing involved in Chinese processing some foreign mature technologies and research results cannot be used directly. On the
other hand, due to the lack of basic language processing resources, such as knowledge bases, corpora and related evaluation mechanisms. Compared with English, the development of Chinese question answering system is more difficult. China's artificial intelligence development is relatively late, but with the strong support of the country, China's artificial intelligence development is also very fast. A large number of Internet companies have joined the trend of artificial intelligence development, and the results are also very many. The Fudan University team proposed the KBQA: Learning Question Answering over QA Corpora and Knowledge Bases [12] method. This method uses neural networks to improve the traditional template-based parsing method and automatically learns templates in the corpus. Today Headline AI scientist Dr. Lei proposed CFO: Conditional Focused Neural Question Answering with Large-scale Knowledge Bases in 2016 [13], but this model can only solve simple factual problems. Chinese Academy of Sciences Automation Liu Kang and others proposed Question Answering over Knowledge Base with Neural Attention Combining Global Knowledge Information in 2016. This method uses a two-way LSTM network and introduces an attention feature alignment mechanism on the question side and candidate answer side to complete the problem. Alignment with candidate answers at different semantic levels.

Research Content of This Article

This article applies the latest research direction of current natural language-machine reading comprehension. Machine reading comprehension (MRC) tasks mainly refer to letting a machine answer text-related questions based on a given text to measure the machine's ability to understand natural language. The origin of this task can be traced back to the 1970s, but limited to small-scale data sets and traditional rule-based methods, the machine reading comprehension system could not meet the needs of practical applications at the time. With the development of modern computers and the emergence of high-intensity computers, it has become a hot spot of research again today. For academic research, machine reading comprehension is one of the most interesting challenges today, and it is the forefront of academic research. At the same time, for industry, machine-to-human communication can increase production efficiency. For today's users, better human-computer interaction, such as voice assistants, can improve the user experience. So a good machine reading comprehension model is of great significance to the development of society.

In machine reading comprehension tasks, recurrent neural networks are now used for orderly modeling, and all sectors of society have developed a large number of models for modeling, many of which have achieved good results, especially in the single-document answer extraction task, some models The answers made are almost the same as those of humans. The use of multi-layer neural networks for sequence modeling increases the model's training and inference speed, but it also increases machine computing time. The SQuAD (Stanford Question Answering Dataset) text understanding challenge sponsored by Stanford University's Natural Language Computing Group is no stranger. It is also known as “ImageNet for machine reading comprehension”. Many research teams from the global academia and industry are actively participating in it. The leaderboards of the competition are updated every moment, and a large number of models are carefully prepared by each team. Machine reading is like the English reading comprehension questions we do in high school. Let the machine act like the human brain to find the answer to the current question in the given text. There may be more than one text and a large number of search queries are required. For example, the results returned by our commonly used Baidu search engine are calculated to extract answers. However, this method is too cumbersome and the single-document task is too simple. Based on the web log, the machine reading comprehension data set is constructed and a new document machine reading comprehension task is designed. In this type, our machine reading comprehension model needs to combine all the document information to extract the correct answer. The main problem faced by such tasks is due to different or incorrect answers to the same question in different passages. In addition, the quality of Chinese training data is mostly uneven, and it is very common to cause wrong answers, and the output results are greatly affected. Therefore, the choice of document quality is also very important. In the multi-document machine reading comprehension task, we must reconsider how to choose the answer and verify the question.
In the reading comprehension part of this article, the R-net model is applied to the Chinese question answering system using techniques such as text processing. Train the corpus on the word vector. Get machine-ready answers to candidate paragraphs that match the target question. In this way, the Q & A with the traditional knowledge base is more readable.

The research in this article is as follows: For our neural network, we want the data input to be a numerical vector. We use our existing word vector wiki Chinese library and embed the data into the word vector library to get a vector training model. For data processing, since the Chinese data set cannot be used directly for training models, we need to process the existing data into data that we can use.

Model Answer Prediction

Model Structure

Figure 1 outlines R-NET. First, questions and paragraphs are handled by a two-way recursive network (Mikolov et al., 2010). We then match the questions and paragraphs with a recurrent network based on gated attention to obtain a problem-aware representation of the paragraphs.

Question and Passage Encoder

Suppose a problem \( Q = \{w_t^Q\}_{t=1}^m \) and \( P = \{w_t^P\}_{t=1}^n \). We first convert the words to their respective word-level embeddings \((\{e_t^Q\}_{t=1}^m)\) and character-level embeddings \((\{e_t^P\}_{t=1}^n)\). Character-level embeddings are generated by applying the final hidden state of a bidirectional implicit neural network (RNN) to the embedding of marked characters. This method helps to deal with some unclear passages. Then, we use a bidirectional RNN to generate all words in the question and paragraph respectively and express them as \( u_1^Q, \ldots, u_m^Q \) and \( u_1^P, \ldots, u_n^P \).

\[
\begin{align*}
    u_t^Q &= B_t RNN_Q(u_{t-1}^Q, [e_t^Q, C_t^Q]) \\
    u_t^P &= B_t RNN_P(u_{t-1}^P, [e_t^P, C_t^P])
\end{align*}
\]

We chose to use a gated loop unit (GRU) in our model because its function is very similar to LSTM and the calculation is simpler.

Gated Attention-Based Recurrent Networks

This model uses a recurrent network based on gated attention to integrate problem information into paragraph representations. It is a variant of an attention-based recurrent network with an additional gate to determine the importance of information about a problem in a paragraph. Given question and paragraph representation \( \{u_t^Q\}_{t=1}^m \) and \( \{u_t^P\}_{t=1}^n \) Rocktèschel et al. (2015) suggested generating...
sentence pair representation \( \{v^p_t\}_{t=1}^n \) through soft alignment of questions and paragraphs in words, as shown below:

\[
v^p_t = \text{RNN}(v^p_{t-1}, c_t)
\]

Where \( c_t = \text{att}(u^Q_t, [u^P_t, v^p_{t-L-1}]) \) is an attention pool for the entire problem \( u^Q \):

\[
s^P_t = v^T (u^Q_t + w^P_t u^P_t + w^P_t v^p_{t-1})
\]

\[
a^P_t = \exp(s^P_t) / \Sigma^m_{j=1} \exp(s^P_j)
\]

\[
c_t = \Sigma^m_{i=1} a^P_t u^Q_i
\]

\( v^p_t \) for each paragraph dynamically merges summary match information from the entire question. Then introduce match-LSTM, which takes \( u^P_t \) as an additional input to the recursive network:

\[
v^p_t = \text{RNN}(v^p_{t-1}, [u^P_t, c_t])
\]

In order to determine the importance of the channel part and participate in the part related to the problem, another gate was added as an input to the RNN \( [u^P_t, c_t] \):

\[
g_t = \text{sigmoid}(w_g [u^P_t, c_t])
\]

\[
[u^P_t, c_t]^+ = g_t \theta [u^P_t, c_t]
\]

Unlike the gates in LSTM or GRU, the additional gates focus on the relationship between the problem and the current pass word based on the concentration vector of the current pass word and its question. The gate effectively simulates the phenomenon that only some soft passes are related to reading comprehension and answering questions. \( [u^P_t, c_t]^+ \) is used for subsequent calculations instead of \( [u^P_t, c_t] \). We call this attention-based recurrent network.

**Self-Matching Attention**

The self-Matching Attention layer is similar to the above layer, except that the attention of the article P and the question Q is turned into the attention of the article P and the article P (self-attention). Then, the output \( c_t \) obtained through self-attention is connected to the output \( v^p_t \) of the previous layer, and is subjected to a bidirectional RNN encoding. If you need to add gating, you can do a gating like the upper layer. The purpose of this section is to use the current article words containing question information to match the answer evidence (Evidence) in the current article.

\[
s^P_t = v^T \tanh(W^P_t v^P_t + W^P_t v^p_t)
\]

\[
a^P_t = \exp(s^P_t) \Sigma^n_{j=1} \exp(s^P_j)
\]

\[
c_t = \Sigma^n_{i=1} a^P_i v^P_i
\]

\[
h^P_t = \text{BiRNN}(h^P_{t-1}, [v^P_t, c_t])
\]

**Output Layer**

The R-NET model outputs the starting position of the answer in the article. In this process, the idea of pointer-network is used. The R-NET model first calculates the distribution of the starting position in the article, \( p^1 \), and then uses this. A distribution \( p^2 \) weights the entire article as input to obtain the distribution \( p^1 \) of the termination position in the article. So this is actually a process of seq2seq, except that there are only two items in the final seq, namely the starting position \( p^1 \) and the ending
position $p^2$. For a seq2seq process, R-NET uses the $u^Q_t$ obtained from the attention-pooling of the problem as the starting key $r^Q$.

$$s_j^t = v^T \tanh(W_u^Q u_j^Q + W_v^Q v_r^Q)$$  \(12\)

$$a_i^t = \frac{\exp(s_i^t)}{\sum_{j=1}^m \exp(s_j^t)}$$  \(13\)

$$p^t = \text{argmax}(a_1^t, \ldots, a_m^t)$$  \(14\)

Here $h_{t-1}^a$ represents the last hidden state network (pointer network) of the answer, and the input of the recursive network of the answer is based on the currently predicted attention vector probability $a^t$

$$c_t = \sum_{i=1}^n a_i^t h_i^b$$  \(15\)

$$h_t^a = \text{RNN}(h_{t-1}^a, c_t)$$  \(16\)

When predicting the starting position $h_t^a$, a represents the initial hidden state network where the answer repeatedly appears. We use the problem vector $r^Q$ as the initial state of the answer recurrent network. $r^Q = \text{att}(u^Q, V_r^Q)$ is a parameter-based attention vector $V_r^Q$ for the problem.

For the initial state $h_0^a = r^Q$, the amount can be obtained by using Attention = Pooling on the problem representation. The specific calculation formula is as follows:

$$s_j = u^T \tanh(W_u^Q u_j^Q + W_v^Q v_r^Q)$$  \(17\)

$$a_i = \frac{\exp(s_i)}{\sum_{j=1}^m \exp(s_j)}$$  \(18\)

$$h_0^a = r^Q = \sum_{i=1}^m a_i u_i^Q$$  \(19\)

To train the network, we minimize the sum of the negative log probabilities of the ground truth start and end locations by predicting the distribution. Through the model network, we predict the answer to "What is heart disease?" In the paragraph "Heart disease is a common type of circulatory system disease. The circulatory system consists of the heart, blood vessels, and neurohumoral tissues that regulate blood circulation. Circulatory system diseases are also called for cardiovascular disease, epidemiological studies have shown that the morbidity and mortality of cardiovascular disease continue to rise, of which coronary heart disease has gradually become the most common heart disease. "7 to 14 (the number represents the number of words) position. This way we can predict the answer.

**Design and Implementation of Heart Question Answering System**

**Related Technologies for Problem Understanding**

Machine reading comprehension (MRC) is a task used to test how well a machine understands natural language by requiring it to answer questions based on a given context. Early MRC systems were rule-based and performed very poorly. [7,8,9]With the rise of deep learning and large-scale data sets, deep learning-based MRC is significantly better than rule-based MRC. Common MRC tasks can be divided into four types: cloze filling, multiple selection, fragment extraction, and free answer. The general MRC architecture consists of the following modules: Embedding, Feature Extraction, Context-Question Interaction, and Answer Prediction. In this article, we use the r-net model to process Chinese documents. Assuming that the answers are extracted from different documents, the correct rate is highest, and the wrong answers and other deviation answers are different. A two-layer neural network is used to design a reading comprehension. A model that can automatically identify segments to find answers.
Preprocessing of Q&A Test

The source of question and answer data is often from the corpus of the community question and answer system in the Internet, so when obtaining the question and answer data, we often get some non-text information or meaningless information, such as link URLs, special symbols, special expressions, etc. The data of the model we trained is using the Baidu dataset dureader and the heart disease data labeled by ourselves, but the data itself cannot be directly applied to the model, we need to process it manually.

Because the R-net model ultimately predicts the start and end positions of the answer at the segment, processing the text in this way will help the model.

Annotate answers based on questions and documents. Since a question may correspond to multiple documents, so there may be multiple answers to a question, but one for similar answers. But the average data length of the dureader dataset is 4.8, the average word length of the answer is 69.6, and the average word length of the document is 396.0, which is 5 times that of MS-MARCO. Due to the large scale and complex problem types, the analysis based on the DuReader dataset is much more difficult than previous datasets. Baidu judges the difficulty of answering questions by calculating the minimum edit distance between human answers and documents. The greater the editing distance, the more editing changes to the document, and the more complicated the answer to the question. For datasets whose answers are directly derived from the original text (such as SQuAD), their edit distance should be 0. So it is very bad to use dureader to directly train R-net. So it can only be handled manually.

System Detailed Design and Operation Effect

After starting the software, look at the interface in Figure 2. Enter the segment data and questions in Figure 3 and click Start to see the model's estimated answer.

This article mainly studies the sentence test for heart disease in Chinese. We will use the actual effect to show the effect of Chinese reading comprehension applied to the question and answer of medical knowledge, and judge from some of the actual results. Figure 1 is a test of a simple problem. It still has a good effect on common problems. This is also the benefit of reading comprehension research. It can be used in various application fields. The introduction of the answer content model, which is applied to the double-layer cyclic God frequent network processing module, still has a very high accuracy rate on commonly used simple questions, especially if the question is some "what is what", the answer can basically be correct. Compared with other Chinese reading comprehension models, it still has certain advantages.

Experimental Analysis

| Model       | BLEU-4 | Rouge-L |
|-------------|--------|---------|
| Match-LSTM  | 32.0   | 39.2    |
| BiDAF       | 31.8   | 39      |
| Our Model   | 46.0   | 48.04   |
| Human       | 56.1   | 57.4    |
The dataset provides two baseline systems, Match_LSTM and BiDAF, and evaluates the model using two indicators, BLEU-4 and Rough-L. ROUGE-L is the current universal detection standard for machine reading comprehension in China. The specific implementation is ignored. From the table, it can be seen that the effect of this model has better effects on other models, but there are still many gaps with humans.

Conclusion

In this study, we introduced R-NET for reading comprehension and question answering. We introduce a recurrent network based on gated attention and a self-matching attention mechanism. Get representations of questions and paragraphs and use pointer networks to locate answer boundaries. Our model uses the latest Chinese data set Ru-Reader, which has certain advantages over other Chinese reading comprehension models. However, there are still many problems that we need to solve. We need to work hard to solve the problems. I believe that we will make greater progress in the future.

References

[1] Liu Kang, Zhang Yuanzhe, Ji Guoliang, Lai Swei, Zhao Jun, Progress and Prospect of Knowledge-based Question Answering Research Based on Representation Learning.

[2] Liu Kang, Zhang Yuanzhe, Ji Guoliang, Lai Swei, Zhao Jun, Research Progress and Prospect of Knowledge-based Question Answering Based on Representation Learning.

[3] Xie Jinzhi: Research and Design of Dean’s Mailbox Question Answering System.

[4] X. Yao, B. Van Durme. Information extraction over structured data: Question answering with freebase [C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, 2014, 1: 956-966.

[5] R-NET: Machine Reading Comprehension with Self-Matching Networks.

[6] csdn R-net Analysis of the principle of machine reading comprehension.

[7] A. Bordes, S. Chopra, J. Weston. Question answering with subgraph embeddings [J]. ArXiv preprint, 2014, arXiv: 1406.3676.

[8] L. Dong, F. Wei, M. Zhou, et al. Question answering over freebase with multi-column convolutional neural networks [C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, 2015, 1: 260-269.

[9] S. W. Yih, M. W. Chang, X. He, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base [J]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, 2015, 1: 270-283.

[10] A. Bordes, N. Usunier, S. Chopra, et al. Large-scale simple question answering with memory networks [J]. ArXiv preprint, 2015, arXiv: 1506.02075.

[11] BAG: Bi-directional Attention Entity Graph Convolutional Network for Multi-top Reasoning Question Answering.