Research and Implementation of Open Domain Question Answering System Based on DuReader Dataset and BiDAF Model

Zechen Guo1,2, Fucheng Wan1,2*, Ning Ma1,2
1 Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China
2 Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou, Gansu 730000, China

*Corresponding author’s e-mail: 306261663@qq.com

Abstract. At present, major search engines emerge in endlessly, and the pop-up result list of search keywords makes people dizzy. In order to improve this problem and make the question query more concise and convenient, this article is based on the DuReader data set and reproduces BiDAF (Bi-Directional Attention Flow) The model builds an intelligent question answering system in the open field. The final model can extract answers to questions more accurately after training. This article embeds deep learning technology into the system and uses intelligent chat to show them.

Keywords- automatic question and answer; machine reading comprehension; DuReader data set; BiDAF model, open domain

1. Introduction
The intelligent question answering system can find the answers accurately to user’s questions from a large number of documents. According to the classification of different fields, the intelligent question answering system can be divided into an open field automatic questioning system and a limited field automatic questioning system. Open domain question answering systems are different from restricted domain question answering systems. This type of system can answer are not limited to a specific domain. When answering questions in the open domain, a certain amount of common sense knowledge or world knowledge and a semantic dictionary are required. For example, WordNet [1] in English is used in many English open domain QA systems. In addition, Chinese WordNet and "Synonym Word Forest" are also often used in open domain question answering systems. Restricted domain question answering system means that the problems that the system can handle are limited to a certain field or content range, such as medicine, chemistry, or the business field of a certain enterprise.

According to the way the answers are obtained, the classification can be simply divided into extractive automatic QA systemand generative automatic QA system. As the name suggests, the
extractive method extracts answers from the text, and the generative method uses an end-to-end generative model to generate accurate answers. This article is based on the DuReader training set for reading comprehension, a fragment extraction type automatic question answering system for open fields.

2. Introduction to related theories and technologies

2.1 Machine reading comprehension

In recent years, machine reading comprehension has become one of the most popular and cutting-edge directions in language processing research. It has important scientific research and practical value that using computers to build models so that computers can read articles, analyze semantics and answer questions like humans. From intelligent customer service to search engines, from automatic composition scoring to intelligent finance, machine reading comprehension technology can automate a large number of time-consuming and laborious manual analysis and improve social productivity greatly.

With the continuous development of deep learning technology, the research of machine reading comprehension has made great progress. In some specific tasks, the answers of computer models can already be comparable to human levels. In 2016, Stanford University launched the machine reading comprehension data set SQuAD (Stanford Question Answering Dataset) [2]. Participating models need to read more than 500 article paragraphs and answer more than 100,000 related questions. Just one year later, the BERT pre-training deep learning model proposed by Google reached the level of accurate matching 87.4% and F1 index 93.2%, surpassing human scores (exact matching 82.3%, F1 index 91.2%), which triggered Hot discussion in the industry [3].

A machine reading comprehension task can be defined as a supervised learning problem, which can generally be formalized as a triplet: T=(Q, P, A) where T is defined as a machine reading comprehension task, Q is defined as a problem, and P is defined as an article or related fragment, A is defined as the answer to the question. In terms of the development of machine reading comprehension tasks, the previous tasks were mostly cloze reading comprehension, which is removing some keywords from the original text and requiring the model to fill in the correct words or phrases. Typical data sets include CMRC2017 and CNN/Daily Mail. [4] But in such tasks, the machine sometimes fails to understand the text. The next stage of machine reading comprehension task is a single-document machine reading comprehension based on encyclopedia, that is, to give question Q, article P and answer A, the answer to the question needs to be extracted from the article, and the answer is often a continuous fragment in the article, so This kind of reading comprehension task is also called single document fragment extraction reading comprehension. With the development of technology, a single document cannot satisfy people's exploration of machine reading comprehension, and a multi-document machine reading comprehension task has been derived. This task is extremely difficult and extremely complex, which is more suitable for machine challenges. Typical data sets include MS-MARCO [5,6] and DuReader, etc.

2.2 DuReader data set introduction

DuReader, as you can see from the name (Du and Reader), is a reading comprehension data set compiled by Baidu. The DuRader data set is composed of a series of 4-tuples, each 4-tuple is a sample T=(q,t,D,A), where A represents a series of answers to the questions in the data set (manually labeled), D represents a collection of documents related to the corresponding question, and t represents the corresponding the type of question, q represents a question. Compared with the previous reading comprehension data set, it has three main characteristics: (1) Large scale of data; (2) More types of questions; (3) Data sources are closer to reality.

2.3 Attention mechanism

The attention mechanism [7] was first proposed in the field of vision. In the field of computer vision, humans do not observe every pixel when recognizing images, but instead focus on certain pixels and pay attention when observing other objects. The force will move to other areas. This gives people an
inspiration to apply it to natural language processing. In the field of natural language processing, processing text information can focus the weight of attention on a specific word according to the current needs, and as the processing of text sentences is different, the weight of the change of time of information will also change. [8] In the ASReader model, the results of its attention are directly sent to the output layer for prediction, while in the BiDAF model, the value of its attention needs to be coded and finally sent to the prediction layer for prediction. There is also a multi-hop attention mechanism represented by ReasoNet. The model will calculate the value of attention and make inference predictions after multiple iterations. [9]

3. System Design

In order to improve the user experience, this article hopes to design an intelligent automatic question and answer system that can promptly give concise answers when users search for questions instead of a bunch of list pages. This intelligent question answering system includes three modules, namely a text retrieval module, a reading comprehension processing module and an answer generation module.

![Figure 1. Question and answer system flow chart](image)

From the Figure 1, we can see the general process of the entire system and the required data. The internal data flow of the intelligent automatic question answering system is a process from questions to multiple documents and then to answers. The user asks questions and delivers them to the system. The system first retrieves relevant documents through the text retrieval system and delivers them to the reading comprehension processing module for reading comprehension operations. The processing results are delivered to the answer generation module for final summary adjustments and finally delivered to the user.

There should be a general paragraph in this part, an overall statement of the implementation of the system.

The realization of the intelligent automatic question answering system in this paper is basically to realize two general directions (the automatic question answering system researched and implemented in this paper includes the core work), the first is text retrieval. As the name implies, it allows the system to find the specific source of the user's question and obtain its related documents as the basis for the second step.

4. Detailed design system implementation

4.1 Text retrieval module

In the implementation of this module, we use web crawler technology to search for keywords and obtain related documents. For example, after a user enters a question in the system, we use crawler technology to crawl relevant web pages, and use a comprehensive analysis algorithm to select the best relevant content as an article for machine reading and understanding, paving the way for the next module.
Figure 2. Text retrieval data flow diagram

The Figure 2 shows the system text retrieval module. The overall effect of the module is to return the documents most relevant to the user’s question through the text retrieval algorithm for other modules to extract answers.

Figure 3. Detailed flowchart of text retrieval module

The Figure 3 is a detailed flowchart of the text retrieval module, in which the document selection algorithm is particularly important, and the selection of related documents directly determines the quality of the final question answer extraction result. This algorithm is a major innovation in this article, and the algorithm is explained in detail below.

The search engines used in this article are divided into Baidu Know and Baidu Search. The information and format contained in them are different, so there is no way to deal with them uniformly. Therefore, the documents obtained by these two search engines must be processed separately. Among them, the document obtained by the know engine is finally intercepted by the information retrieval algorithm of the relevant content of the answer part of Baidu. This also uses the HTML analysis in the crawler technology to directly select the user's answer fragment. All content is the text of the relevant answers to the questions raised by the user on the web page, but it may also be irrelevant to the questions raised by the user. Based on Baidu knows the characteristics of search, it can be preliminarily considered that the relatively high value of the content or document is higher. In other words, the document and the question are more relevant and the answer is the most likely. Extracted from the previous document. Therefore, the data processing operation for the known part is to connect the document order, and add a specific symbol to identify the joint position of different documents, but one thing to note is that the document length selected by the algorithm has been specified as a fixed value. The maximum length specified above cannot be changed. If the maximum length is exceeded, the model of the reading comprehension module may issue an exception during training, causing the training to fail. So be careful not to exceed the maximum length when connecting, so that the processing of the known part is completed.

The document obtained by Baidu search engine is somewhat different from the previous known part. The search part will contain all relevant information of the entire web page. The entire web page searched by Baidu in this article plays a positive role in extracting the final answer. Therefore, this article does not use crawler technology HTML to parse out the body part of each web page. Therefore, the search part must be processed separately. The specific steps are as follows. First connect the title...
and the document in order. If the total length does not exceed the maximum length preset at this time, this will be the result of the processing. If the length exceeds the preset maximum length, the next 5 steps will describe this matter. First, you need to calculate the correlation between each paragraph and this article. You can use the F1 (fuzzy matching value) value to calculate. The related algorithm of the F1 value will not be repeated in this article. Next, select the earliest paragraph from the top k paragraphs with the highest score, select its title and the corresponding paragraph and the next paragraph, and for the third and subsequent paragraphs, only the first sentence is selected, and the total period is maintained. The length does not exceed the preset maximum length. Finally, all the processed results are connected with the partition number, and the final and processed results are obtained.

As for the reason for this treatment, this is because this article has made the following two assumptions. First, this article believes that the answer will be in a paragraph similar to the title or a paragraph after the article, and the closer the paragraph is, the higher the value. The purpose of introducing the title is to judge whether the paragraph and the article are similar.

4.2 Reading comprehension module

Figure 4. Reading comprehension module model training flowchart

The Figure 4 is the reading comprehension processing module supported by the deep learning-based natural language processing algorithm in the second part of the intelligent automatic question answering system. This module uses the reproduced BIDAF model to predict the article where the answer to the download process is described above and finally obtains the concise answer that the user needs. In this article, we reproduce the BIDAF model and use the DuReader data set to train the model and finally generate a reading comprehension model, which can accurately extract the answers to the questions in the article.

4.2.1 Definition dictionary class

The purpose of building the vocabulary is to map words to an int index and create an index, so that each word has an integer index number, which is also a row in the word vector matrix, such as the index number corresponding to "I" If it is 0, then the 0th row in the word vector matrix is the vector representation of the word "I". The index number corresponding to the dictionary should correspond to the word vector that has been trained, that is to say, the index number corresponding to the same word should be saved in the same vocabulary, so that the index index of the same word is the same every time the same word is input, the saved model In order to reproduce the results, the vocabulary and the word vector must be mapped one to one. If there is no one-to-one correspondence, the specific position of the word embedding in the word embedding table will be confused when the pre-training model is loaded again.

4.2.2 Training word vector

When we process natural language, text should be represented by word vectors, that is, a word should be represented by vectors. The reason for this is that words are traditionally treated as discrete atomic symbols in the process of natural language processing, so we think "Emperor" can be represented as 16200, and "Queen" can be represented as 16355. In fact, these codes are arbitrary, and he does not carry the relationship between words and words. For example, the emperor and the queen are both royal
families, the emperor is male, and the queen is Favoring women, if we can find a representation method to carry this information together in the model training, it will have a great effect on improving the model. We are letting words carry more information, so word vectors are born. This article will use the word2vec tool in gensim to train text word vectors.

In the model parameters, size is the dimension of the output word vector. Here we choose a fixed dimension of 300 dimensions. Min count is the minimum number of occurrences in the text, that is, when a word appears at least 5 words in the text, we will add it to the training. The Workers parameter is the number of threads working at the same time. The saving method here is to emphasize that the model we saved is the last trained word vector, this is the format of which readers need to pay attention to. There are models in numpy format and models in w2v format. Readers need to choose according to their needs. The vocab attribute of the model.wv.vocab model stores the word list used in the entire training.

4.2.3 BIDAF model prediction algorithm

Figure 5. BiDirectional Attention Flow Model model architecture

In Figure 5, we can see that this is the structure diagram of Bi-Directional Attention Flow (BiDAF). It is a multi-level reading comprehension model published by Professor Minjoon Seo and others. Different levels of granularity are introduced into the model. In short, it introduces character-level vector features, word-level vector features and article-level vector features. Among them, the character-level vector and the word-level vector are based on word2vec, and the article-level features are made using two-way LSTM, which finally aggregates the semantics of the article. Attention mechanism is also added to the model, there are attention from article to problem, and attention from problem to article, to obtain more accurate results. Finally, the pointer prediction layer is used to predict the final S and E positions.

The model is divided into 6 layers: [1]. Character embedding [2]. Word embedding [3]. Context embedding attention flow layer: [4].context -to- Query Attention [5].Query-to-context Attention [6]. Output layer

In this paper, after the model is trained, it is added to the reading comprehension module of the intelligent automatic question answering system to predict the start and end positions of the document where the answer is for the user.
In Figure 6, summary of reading comprehension processing module: After the above operations, the overall machine reading comprehension processing module architecture has a basic framework. It is divided into a training module and a prediction module. In the training module, the content of the text retrieval is divided into words and then the dictionary mapping is performed on the dictionary respectively, and the word embedding of the article and the question is queried according to the trained word vector table. Finally, the two are sent to the Bi-DAF model for answer extraction. The final model will generate S and E fragments of the relevant article, and use the answer reduction algorithm to get the last extracted answer fragment. Continue to store the article questions and answers in the answer database. This answer database is the same database as the one mentioned above.

4.3 Answer generation module

The Figure 7 is the flow chart of the answer generation module. Through the reading and comprehension processing module to deal with the questions and articles, the answers to the questions are basically decomposed and extracted completely, but after many experiments, we found that the extracted answers cannot meet the delivery of the end user use. The answer generation module mainly solves problems such as irregularities and confusion of sentences. The regular answer prompt can also be called regular answer filtering. That is, using regular sentences to delete and filter irregular answers again. Garbled, filter out other languages for some Chinese problems. Answer quality evaluation is a scoring mechanism for answers. The final answer is scored by calculating the correlation between the answer and the question and the degree of fuzzy matching. Deliver concise answers for users.

5. Summary

Aiming at the DuReader data set and Bi-DAF model, this paper builds an open field question answering system. The text retrieval module uses crawler, keyword expansion, document selection and other technologies to accurately obtain relevant documents. In the machine reading comprehension module, the model obtains different characteristic information of the answer to the question through multiple iterations and focuses the information, which can predict the final answer more accurately. The regular extraction and answer evaluation techniques in the answer generation module can ensure the reliability of the final answer. In the final evaluation of the system, the question and answer system can answer users’ questions more accurately.
Acknowledgement
This research is supported by The National Natural Science Fund (NO. 61762076). Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China.

References
[1] Hyejin Jang, Yujin Jeong, and Byungun Yoon 2021 TechWord: Development of a technology lexical database for structuring textual technology information based on natural language processing Expert Systems with Applications 164
[2] Rajpurkar P, Zhang J, Lopyrev K, et al. 2016 SQuAD: 100,000+ Questions for Machine Comprehension of Text pp 2383-2392
[3] Raziye Zare Hoseini, Ebrahim Mohammadi Goltapeh, Seyed Ali Mohammad Modarres-Sanavy, and Ali Heidarzadeh 2021 Effect of the bio-fertilizers on the steviol glycosides (SGs) content and biomass in Stevia rebaudiana (Bert.) Bertoni at vegetative and flowering stages Scientia Horticulturae p 275.
[4] Bingtian ZHANG 2017 Pursuit of Power: A Critical Discourse Analysis of the CNN and China Daily South China Sea Reports Cross-Cultural Communication 13 (8)
[5] Nguyen T, Rosenberg M, Song X, et al. 2016 MS MARCO: A Human Generated Machine Reading Comprehension Dataset
[6] Rangan Majumder 2016 MS MARCO: Microsoft Machine Reading Comprehension Dataset [EB/OL], http://www.msmarco.org/
[7] Vaswani A, Shazeer N, Parmar N, et al. 2017 Attention Is All You Need
[8] Cui Y, Chen Z, Wei S, et al. 2016 Attention-over-Attention Neural Networks for Reading Comprehension
[9] Kadlec R, Schmid M, Bajgar O, et al. 2016 Text Understanding with the Attention Sum Reader Network pp 908-918