A Disease Interaction Retrieval and Recommendation System based on NLP Technology

Puyuan Yao*

The High School Affiliated to Renmin University of China, Beijing, China

*Corresponding author’s e-mail: yaopuyuan0302@126.com

Abstract. There is a great request for a Chatbot system that uses natural language processing for disease retrieval and recommendation in medical treatment with interactive functions. This paper uses cMedQA2 data for training, developing and testing. The Chinese word segmentation tool jieba to reassemble the text data of consecutive words into the word sequences, and uses Bag of Words to count the frequency of keywords appearing in sentences, and uses N-gram method to predict whether the evaluation and prediction of a sentence is reasonable. This paper proposes an improved TextCNN network framework structure, which uses overlay technology to ensure the contextual coherence of information. The final experimental results show that there are 96.2% on the training dataset and 87.5% on the testing dataset.

1. Introduction

With the popularization of mobile phones, laptops and other devices and the rapid development of the mobile Internet, more and more applications for mobile computing devices have been developed in large numbers, and their use frequency has exceeded their desktop applications. The ubiquitous mobile computing trend is still accelerating. In this very cutting-edge and widely used research field, one of the current challenges is to further enhance the application capabilities of mobile computing devices by using environmental information, which requires the collection of all interactive information between users, applications and the surrounding environment[1]. The simplest and commonly used human-computer interaction methods are undoubtedly the keyboard and mouse. The virtual keyboard and touch press on smart phones are also effective methods of interaction. In addition, they also include voice, image and other interactive forms. However, these interactive forms still have many limitations. For example, for the elderly and people with physical disabilities, the existing human-computer interaction methods still have certain barriers, and it is difficult to help these people. To achieve richer and easier human-computer interaction functions, it is necessary to design and develop applications with context-aware capabilities. However, the current research status in this field has not yet met the requirements, mainly because the definition of context is still vague, and there is a lack of help to promote Models and methods for context-aware applications. For example, a smart phone may predict the user's behavior when a call comes in. If the user is in a meeting, it is likely to refuse to answer the call. This context-sensitive information processing method can help improve the user's experience.

Natural language processing is a technology used to automatically analyze and represent human language[2]. Natural language processing technology is used to perform various tasks related to natural language, from word parsing to machine translation, and human-machine dialogue functions similar to Apple mobile phone Siri, all of which require the support of natural language processing.
Traditional NLP machine learning methods use high-dimensional/sparse features as features, and input them into relatively basic machine learning models such as support vector machines and logistic regression. Such processing methods require manual definition and extraction of features, so the results of the model are deeply affected. The influence of feature selection. With the increasing popularity of deep learning methods, recent NLP research has increasingly focused on the use of new deep learning methods. For example, deep neural networks based on word embedding have produced good results on various NLP tasks.

Chatbot requires complex development methods and skills, because many functions of Chatbot are based on its knowledge base and correspond to words and their correct answers one by one. Therefore, a large database and knowledge base are needed to provide a Chatbot interactive platform that users feel smooth and satisfied. The knowledge base can be completed manually, or it can be learned from some libraries, that is, save some words and provide answers[3].

However, the existing Chatbot is still difficult to meet the needs of users. Part of the reason is that Chatbot design needs to be based on specific principles, that is, generally speaking, you cannot deviate from the script. If the user’s question or communication is outside the script, it is difficult for Chatbot to understand and cause repeated useless communication. And natural language processing models usually require a long time of training and massive data accumulation, and it is difficult to meet these two requirements for general tasks.

This paper proposes an improved TextCNN model, which takes word vectors as input to initialize the neural network and fine-tune it, and initially produces a pre-training model. Model training is then performed on the collected medical text data. This model uses a 1*1 convolution kernel with a step size of 1, and performs operations such as pooling and dropout. In the end, 96.2% was obtained on the training set and 87.5% on the test dataset.

The organization structure of this article is: The second part introduces deep learning and Chatbot technology based on natural language processing. Followed by the detailed introduction of the model proposed in this article; the fourth part gives the experimental results on the cMedQA2 dataset; the last part is conclusion and future.

2. Deep learning and NLP-Based Chatbot technology

Machine learning uses mathematical models to convert information into data, and is often used for data classification. Machine learning is divided into supervised learning and unsupervised learning. Common methods include KNN, Decision Trees, SVM, etc. Although different algorithms and models have their own advantages and disadvantages, they all have similar and scientific implementation methods in the process of solving practical problems such as collecting data, analyzing data, training, testing, and using. Analyze the collected data and transform the original data into data that can be used by the model; training and testing enable manual or algorithmic use of errors to continuously adjust the weights and parameters to achieve a state suitable for the actual scene; finally, the actual model is used to solve the problem. Among them, machine learning is the most widely used in classification problems, and it is also one of the most basic uses in machine learning.

Deep learning is a specific subfield of machine learning. Different methods are often used for different datasets. Neural networks are often used for natural language processing. The process of deep learning is roughly inputting data, data transformation, calculating loss, and updating weights. From the proposal of the perceptron model in the 1960s to the concept of deep learning proposed by Hinton et al. in 2006, deep learning has continued to develop, and various neural network models have been derived, such as SNN, CNN, and RNN. Different neural networks are used in different fields, such as CNN for image recognition and RNN for audio analysis. Almost all applications such as speech recognition and image recognition use neural networks to transform information into data. The neural network has high controllability, and the adjustable parameters make the neural network have high accuracy.

NLP models based on deep learning use these embedding vectors to represent words, phrases and even sentences. Before the advent of deep learning, traditional NLP often used simple methods based
on word count. Practical deep learning NLP technology has begun to play a role in fields such as emotion classification, chat robots, machine translation, and continuous speech recognition.

Chatbot is a computer program application that can communicate with humans using natural language by imitating intelligent conversations. The most famous chatbots include Apple’s Siri and Microsoft’s XiaoIce. Figure 2 shows XiaoIce dialogue interface. Chatbot are responsible for assisting the process of human-computer interaction, and can confirm tasks or interact with users by asking and responding to user questions. Chatbot usually inputs natural language text or voice, then gives the most appropriate answer to the input, and gives feedback to the outside world in the form of text or voice.

In addition to providing users with general services, there are also some Chatbot entering the professional field. Academia and industry often encounter such problems. If the dataset is small, is the BERT large pre-training model effective [4]. For example, recruitment Chatbot assist companies in the recruitment process and use artificial intelligence technology to complete some monotonous tasks in the recruitment process to reduce labor costs. Employers can directly talk to Chatbot and inform them of employment requirements. Chatbot will automatically retrieve and provide employers with suitable job candidates. At the same time, it will also manage the entire recruitment process. Written examinations, interviews, and uploading materials can be completed automatically. The personal shopping assistant Chatbot is responsible for bill payment, mobile phone charging, and handling services such as taxi, meal ordering, housekeeping, and public transportation. It will soon launch services such as flight tickets, hotel reservations, express service reservations, and insurance purchases, and it can also automatically execute for users. Various procurement tasks.

Table 1. Companies information of medical Chatbot.

|                | Your.MD | Sensely | Buoy Health | Infermedica | Florence |
|----------------|---------|---------|-------------|-------------|----------|
| Established    | 2012    | 2013    | 2014        | 2012        | 2016     |
| Location       | London  | SFO     | Boston      | Wroclaw    | Dortmund |
| Target customers| Patient | Patient | Patient     | Patient     | Patient  |
|                | Physician| Physician| Physician   | Physician   |          |
| Data processing| text    | Text    | text        | Text        | text     |
|                |         | Image   | Image       | video       | video    |
| Funds          | 17.3M   | 11.8M   | 9 M         | 5 M         | Un       |
|                |         |         |             |             | disclosed|

With the growth of the number of people in today's society and the continuous increase of working hours, the demand for mobile medical care is constantly expanding. AI online consultation is a branch of mobile medicine, which is a cross-combination of integrated artificial intelligence, natural language processing and analysis, mathematical modelling and numerical algorithms. With the establishment and breakthrough of AI algorithms and models, the recognition of natural language by artificial intelligence has reached a high degree of accuracy, and the classification of AI triage is more detailed, so that patients can more accurately learn their own disease and pathology. Medical and health-based Chatbot for consultations have appeared in the market. When people cannot go out to the hospital for consultations, Chatbot tools are usually used for online consultations, which reduces the probability of infectious diseases and Waiting time in line. As shown in the Tab.1, the main company information of the current medical Chatbot from the Internet search. Among them, Your.MD collects medical literature covering more than 1,000 medical conditions for machine learning, mainly learning common symptoms in medical literature to provide advice to patients. The Sensely platform integrates the avatar and voice chat interface into existing applications.
Figure 1 shows the general form of word embedding, which a very long vector is used to contain and represent all words in the vocabulary. For example, the word King occupies several positions in the vector belong to Figure 1, same with other words. The difference is that the positions of King and Man on the vector are similar, while the positions of Woman and Queen are similar. That is to say, the method of word embedding puts words with similar meanings in the position of the vector of similar words. By using the similarity measurement method, the similarity between these vectors can be measured. It is conceivable that by pre-training this kind of word embedding on a large word library, the basic syntax and semantics of the universal human language can be obtained.

![Figure 1. Word embedding form.](image)

3. An improved Structure based on Human-computer NLP
The basic function of the convolution operation in the convolutional neural network is filtering, the purpose is to further extract the features of a certain significance from the basic pixel features. The general convolutional neural network structure does not need to design the convolution value, and it is obtained through learning. It can also be operated in parallel when the convolution layer is required to extract multiple features. In text type data, the local feature extraction corresponding to CNN is actually the sliding window feature filtering for word. Describe high-level and primary information by combining and filtering multiple contextual features[5].

![Figure 2. An improved TextCNN structure.](image)

This paper proposes an improved TextCNN network framework structure, which is shown in Figure 2. The basic function of the convolution operation in the convolutional neural network is filtering, and the purpose is to further extract the features with a certain meaning from the basic pixel features. The general convolutional neural network structure does not need to design the convolution value, and it is obtained through learning. It can also be operated in parallel when the convolution layer is required to extract multiple features. In text type data, the local feature extraction corresponding to CNN is actually the sliding window feature filtering for word. The network structure in this paper overlaps the use of contextual word information, describes advanced and primary information by combining and filtering multiple context features, and uses overlap technology to ensure the contextual coherence of information.

For the existing dataset, first use the Python-based Chinese word segmentation tool jieba to recombine the text data of consecutive words into a word sequence[6], which is the basis for further Chinese text analysis. In English, there will be spaces between each word to distinguish, which is much simpler than Chinese. The results are shown in Figure 3. To classify sentences before deep learning, Naive Bayes or support vector machines are usually used to represent the input model as bag of words, such as "Today we visit the Summer Palace, the Summer Palace has the famous Buddha
Incense Pavilion" and "The Summer Palace" It is a famous garden scenic spot. There are 10 words in total, "Today we are visiting the Summer Palace, there is a famous Buddha Xiangge is a garden scenic spot." Then the corresponding vector of the first sentence is [1 1 1 2 1 1 0 0 0], the vector corresponding to the second sentence is [0 0 0 1 0 1 0 1 1 1], which is to count the frequency of words in the sentence, and then use the N-gram method, which is the Nth The occurrence of this word is only related to the previous N-1 words, and not related to any other words. The disadvantage of this method is that if there are tens of thousands of words in total, each text needs to be represented by a corresponding tens of thousands of dimensional vectors, and the order of words in the sentence is also thrown away.

Figure 3. The example of text data to word vector.

Compared with the bag-of-words model, Word2Vec extracts word vectors to map each word a high-dimensional vector by the word vector model maps. It obtains the cosine through a lot of training, which can preserve the order of words, and there is no bag-of-words model. The code is shown in Figure 4.

4. Experimental results

After searching and comparing on the Internet, the text data suitable for triage in cMedQA2 was selected as the basic research data. At the same time, some text data related to disease queries were collected by a web crawler as a supplement. cMedQA is the data of the mobile medical online forum http://www.xywy.com/. In the forum, doctors will answer questions raised by Internet users, and will give diagnoses and suggestions based on the symptoms described by users.

Table 2. Companies information of medical Chatbot.

| DataSet | #Ques | #Ans  | Ave. #characters per Question | Ave. #characters per Answer |
|---------|-------|-------|--------------------------------|-----------------------------|
| Train   | 100,000 | 188,490 | 48                             | 101                         |
| Dev     | 4,000   | 7,527  | 49                             | 101                         |
| Test    | 4,000   | 7,552  | 49                             | 100                         |
| Total   | 108,000 | 203,569 | 49                             | 101                         |
We assume that the answers answered by qualified doctors are true answers to the original question. In order to protect the privacy of users, data is anonymized and all possible personal information is deleted. The dataset contains a total of 54,000 questions and about 100,000 answers. cMedQA2 is an extended version of cMedQA, which contains approximately 100,000 medical-related questions and approximately 200,000 answers. The ta is the data distribution of the dataset.

The dataset consists of three parts: training set, developing set and testing set. The training set is used to train the model, the developing set is used to adjust the hyperparameters in the model, and the testing set is used to evaluate different model levels. The average number of characters for the question is about 49, and the average number of characters for the answer is about 101. In the process of development and testing, each question has 100 candidate answers, including truth values, and the goal of training is to select the most suitable truth answer from the candidate answers[7].

Subsequently, the word vector is used as input to initialize the neural network and fine-tune it to initially generate a pre-training model. Then we conducted model training on the collected medical text data. The word count data format was similar to "Oncology: There are many pustules on the tongue. Should I see the Department of Oral Mucosa of the Ninth Hospital or the Department of Frontal Oncology?" The content of the consultation text with detailed triage information. The loss function used cross entropy, and the last layer used the fully connected layer. Finally, 96.2% was obtained on the training set and 96.2% on the test set. 87.5% accuracy rate.

5. Conclusion
The online disease inquiry and recommendation system provides patients with a convenient and safe method and reduces the burden on medical staff. This paper proposes an improved TextCNN framework based on cMedQA2 data, which uses Word2Vec to extract word vectors from a large number of text word data collected on the Internet and build a dictionary. The final experimental results show that the network structure proposed in this paper achieves an accuracy of 96.2% on the training set and 87.5% on the test set. The next step is to use this method to obtain feedback on satisfaction scores and whether they have appeared in recommended hospitals, so as to further enhance the convenience and effectiveness of this method.

References
[1] Dey, Anind K., Gregory D. Abowd, and Daniel Salber. “A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications.” Human-Computer Interaction 16 (2): 97–166, 2001.
[2] Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. “Recent Trends in Deep Learning Based Natural Language Processing [Review Article].” IEEE Computational Intelligence Magazine 13 (3): 55–75, 2018.
[3] Abdul-Kader, Sameera A., and John Woods. “Survey on Chatbot Design Techniques in Speech Conversation Systems.” International Journal of Advanced Computer Science and Applications 6 (7), 2015.
[4] Aysu Ezen-Can. A Comparison of LSTM and BERT for Small Corpus. arXiv:2009.05451v1 [cs.CL] 11 Sep 2020
[5] Kim, Yoon. “Convolutional Neural Networks for Sentence Classification.” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.
[6] https://www.cnblogs.com/zhbzz2007/p/6084196.html
[7] Zhang, Sheng, et al. “Multi-Scale Attentive Interaction Networks for Chinese Medical Question Answer Selection.” IEEE Access, vol. 6, 2018, pp. 74061–74071.