HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi-Directional Attention

Qiming Bao*
qbao775@aucklanduni.ac.nz
The University of Auckland
Auckland, Auckland

Lin Ni
l.ni@auckland.ac.nz
The National Institute for Health Innovation(NIHI)
Auckland, Auckland

Jiamou Liu
jiamou.liu@auckland.ac.nz
The University of Auckland
Auckland, Auckland

ABSTRACT

This paper proposes a chatbot framework that adopts a hybrid model which consists of a knowledge graph and a text similarity model. Based on this chatbot framework, we build HHH, an online question-and-answer (QA) Healthcare Helper system for answering complex medical questions. HHH maintains a knowledge graph constructed from medical data collected from the Internet. HHH also implements a novel text representation and similarity deep learning model, Hierarchical BiLSTM Attention Model (HBAM), to find the most similar question from a large QA dataset. We compare HBAM with other state-of-the-art language models such as bidirectional encoder representation from transformers (BERT) and Manhattan LSTM Model (MaLSTM). We train and test the models with a subset of the Quora duplicate questions dataset in the medical area. The experimental results show that our model is able to achieve a superior performance than these existing methods.

KEYWORDS

Hierarchial BiLSTM attention model, natural language processing, knowledge graph, question answering, medical chatbot.

ACM Reference Format:
Qiming Bao, Lin Ni, and Jiamou Liu. 2020. HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi-Directional Attention. In *Proceedings of the Australasian Computer Science Week Multi-conference (ACSW 2020), February 4–6, 2020, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

Difficulty in seeing a doctor, long queuing time, and inconvenience of making appointments have long been hurdles facing patients when they try to access primary care services. To solve these challenges, governments and health care providers around the world are investing in new methods that facilitate more effective use of resources to meet demands. As an example, New Zealand government has issued the "6-hour target" in 2009 aiming to significantly boost the availability of medical resources [10], while more recently, the Precision Driven Health initiative targets a new model that joins force government, commercial, health care providers and researchers in New Zealand, in the hope to better harness the power of digital medical data and information technology to deliver enhanced services [8].

Artificial intelligence plays a crucial role in the advancement of information technology to improve healthcare service quality and efficiency. In particular, chatbots amount to one of the most popular AI technologies for this purpose. A chatbot is a software system that consists of an interactive interface with patients or medical practitioners to provide a range of knowledge extraction tasks and real-time, personalized feedback. Chatbot technologies have been rapidly developed, especially in the medical field. Many medical chatbot systems have been proposed over the years. Typical applications of chatbot include medical assistants that help patients to identify their symptoms, medical service front desks that direct the patient to suitable healthcare service departments, i.e., doctors, and so on.

Our work aligns with the main themes of medical chatbot technology and aims to serve three main objectives: The first objective is to reduce waste on resources and time for users when accessing information with chatbot technologies. We aim to maximally help users to search for the necessary information with a human-like interface. The second objective is to provide more precise answers to ordinary users who have little domain knowledge. In other words, we hope that with AI technologies, the system can understand the meaning of the natural language and be able to reply with high-quality feedback accordingly. The third objective is to make it easier to manage and extend the features and databases. We want to design a system with a flexible and scalable structure to enable efficient management of the functionality and datasets.

To this end, we first design a framework to implement a generic chatbot system. Our chatbot framework contains two main modules. The first module is the user interface, which contains a web-based chatbot front-end, a local GUI, and a back-end to handle database management. The second module serves to respond to user's queries based on our hybrid QA model, which contains a knowledge graph and the *hierarchical BiLSTM attention model* (HBAM).

We build our *Healthcare Helper system with a Hybrid QA model* (HHH) as an instance of the chatbot framework above. The knowledge graph stores more than 600 different kinds of disease records and is able to answer six different types of questions, while the HBAM can query from a big dataset containing 29287 medical questions-and-answer pairs (171 from ehealthforumQAs, 5679 from questionDoctorQAs and 23437 from webmdQAs).

*Corresponding author

ACSW 2020, February 4–6, 2020, Melbourne, VIC, Australia
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7697-6/20/02…$15.00

One novelty of our work lies in the utilization of a *hybrid QA model* that combines a knowledge graph database and an NLP model. A user's question firstly will be queried from the knowledge graph. If it cannot find any result, a text similarity model will be used to find the answers from a large medical QA dataset.

The highlight of this paper within this model involves a novel deep learning-based text-representation and similarity-comparison model: the *HBAM*. HBAM consists of a BiLSTM layer and a word attention layer. The functionality of the BiLSTM layer is to capture the forward and reverse directional information of a sentence. The word attention layer is used to capture the keywords in a sentence. Siamese framework and Manhattan distance are used to compute the medical level semantic similarity. Siamese framework has been widely proposed in the metric learning tasks [24] [4]. Manhattan distance has been utilized to measure sentence similarity, such as cosine similarity [24]. Comparing with MaLSTM [14] and BERT [7], our HBAM gets the highest score in the experiments with different datasets.

Paper organization. The rest of the paper is organized as follows. Section 2 presents the two core problems studied in this paper and presents related works. Section 3 presents the main system architecture of the chatbot framework. Section 4 presents how the knowledge graph is implemented for our medical chatbot. Section 5 describes our HBAM model which is the key to natural language understanding. Section 6 provides some sample output of the system as well as a quantitative analysis of the performance of the system using two sets of experiments. The results show that our system achieves superior performance as compared to existing systems. Section 7 concludes the paper with a discussion of potential future work.

2 PROBLEM FORMULATION AND RELATED WORK

In the following, we define two problems that are at the center of the chatbot system. The first problem aims to realize the ability of natural language understanding, i.e., developing the necessary mechanism for the software system to understand natural language questions as a human would do. The second problem aims to extract the relevant information from a domain-specific database so that answers can be generated to be fed back to the user.

(1) User question understanding (Intent Detection): Natural language understanding (NLU) and natural language processing (NLP) to understand and process a user's question.
(2) Knowledge base storage and retrieval: A domain knowledge database to be able to store and query the medical questions and answers.

We review existing works that are related to the two problems above.

Chatbots. Eliza was the first chatbot in the world developed at MIT Artificial Intelligence Laboratory by Joseph Weizenbaum in 1966 [21]. However, Eliza cannot understand the question from the user. Parry was the first chatbot to pass the Turing Test created by psychiatrist Kenneth Colby in 1972 [3]. Nevertheless, only 48% of the psychiatrists can correctly figure out the real patient from the conversation.

Ni et al. [15, 16] tried to use the multiple-turn dialog decision tree to make a judgment for a patient. Helen et al. [26] found that using transfer learning to transfer common scenarios from SQuAD to Bible QA can effectively improve the accuracy of the model on shorter context conversations. Dai, Z., etc. [6] proposed a "focused pruning method" to reduce the candidate result space and make some improvements by using N-gram methods, which efficiently reduce the data noise. Wang, Y., etc. [20] proposed "APVA" to accurately predict the connection between the question entity and answer entity. Yih, S., etc. [23] proposed a new semantic analysis framework when the question has been transferred and analyzed to query language, the new query will be related to the knowledge base. Yu, M., etc. [25] proposed a hierarchical RNN network by using residual learning to improve the performance in 2017. When there is an input question, it can detect the relation inside the knowledge base. Besides, they developed a simple KBQS system that integrates the entity linking and relation detector.

Knowledge Base Storage and Retrieval. Cui et al. [5] built an open domain knowledge base question-and-answer system in 2017. They tried to design more templates from a billion scale QA corpora to better understand questions. However, they do not consider user intention with a knowledge graph so that the answer is limited by the template itself rather than capture the user intention. Lukovnikov, etc. [11] propose a model to capture useful information from different layers and combine the different characters of RNN and CNN. They have used RNN [12] to capture the semantic level connection and Attention [19] to follow the entity and relationship. However, RNN cannot capture the forward and backward context information.

Siamese based Semantic sentence similarity. Mueller et al. have proposed a Siamese Long Short-Term Memory (LSTM) network to compute the semantic similarity between two variable-length sentences [14]. However, LSTM cannot detect the keywords from a sentence. Baziotis et al. [2] proposed a Siamese architecture with Bidirectional Long Short-Term Memory (LSTM) networks with an attention mechanism. The model uses Bidirectional Long Short-Term Memory (LSTM) to capture both two-direction contexts. However, they consider the fully-connect (tanh) in the final layer to make the classification, which can cause over-fitting.

3 SYSTEM ARCHITECTURE

We solve the two problems mentioned above by including a hybrid QA model in our chatbot framework, which combines a knowledge graph to manage a medical dataset and the HBAM to understand the text. The adoption of such a combined system is driven by the following motivations:

Firstly, a knowledge-based system holds some clear advantages in providing targeted responses to well-defined questions and thus is a convenient and reliable approach in implementing a question-answering system in knowledge-centric domains such as medical fields. A predominant type of medical question seeks explanations of specific symptoms that have rather specialized knowledge, and a knowledge-based system can quickly return the desired results upon requests. Furthermore, certain questions require a certain amount of logical reasoning, and these are, e.g., deriving the cause

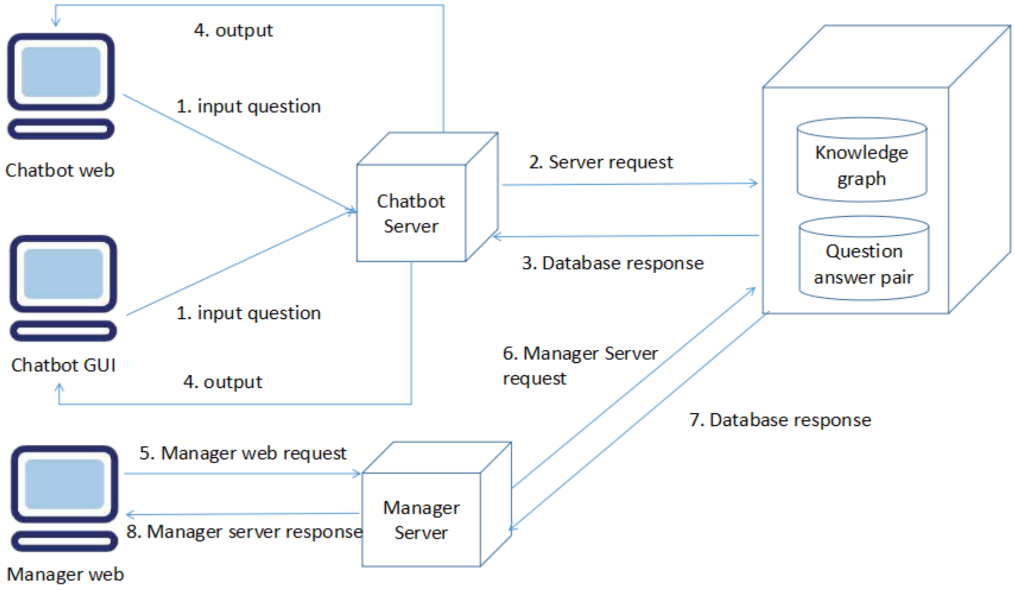


Figure 1: The HHH System Architecture

of certain illness, which can also be solved using a knowledge representation approach such as RDF queries, as RDF triples in a knowledge graph can well represent the complex connections between entities. Therefore, it is natural to adopt a knowledge graph as an integral part of a question-and-answering system.

Secondly, a knowledge-based system can sometimes be too rigid in a conversational context. A patient may not be able to use the vast amount of domain-specific and accurate keywords in formulating a question, but rather, they resort to a casual and even layman's language. The knowledge graph contains a fixed set of knowledge, and when the system fails to match a question with an RDF triple, a knowledge-based system may fail to provide a meaningful answer. Thus, it is beneficial to go beyond merely encoding knowledge explicitly by RDF triples, but preferably using an alternative, data-driven approach. Given this limitation, we propose a neural-based model, namely, HBAM, which provides a more flexible model for various situations. The most frequently-used questions in the conversation model can be filtered first, followed by dialogue understanding. Furthermore, when the knowledge graph cannot be parsed and matched to the appropriate problem, the method of comparing the similarities is used to find the most similar problem in the question-and-answer the knowledge base. The utilization of HBAM is expected to improve the dialogue quality of our system.

Figure 1 summarizes the system architecture of HHH:

- two chatbot clients (website and GUI) connect with a chatbot server;
- one manager client communicates with a manager server;
- a hybrid QA model aims to respond to the messages from the chatbot server with two datasets (a knowledge graph, and a medical QA pair dataset) that are managed through the manager server.

The knowledge graph is developed by Neo4j[1] with data from the Health Navigator New Zealand[2], common illnesses and symptom[3] and common diseases and conditions[4]. The QA pair dataset[5] is generated in 2017, originally from eHealth Forum[6], Question Doctors[7] and WebMD[8] (HealthTap and iCliniq are not used). HBAM (which will be presented in Section 5) will be used to find the best match questions from this QA pair dataset and return the answers to the user.

Figure 2 shows the hybrid QA model in the Chatbot framework. When a user's question is given as input, it can be processed by our two QA retrieval modules.

(1) The information from "Web Interface Interaction" will be transferred into the information retrieval module, which first tries to retrieve the answer from our two datasets. If the answer can be extracted directly from the knowledge graph dataset, the information retrieval module can retrieve and return the answer.
(2) If, on the other hand, the required answer cannot be found from the knowledge graph due to the limitation of the scale of the dataset. In this case, the question will be transferred into the question-answer pair retrieval module. Here we use HBAM to check the semantic similarity of the user's question and the questions from the question-answer pair dataset. The top *k* most similar questions will be returned as the answer set.

[1] https://neo4j.com/
[2] https://www.healthnavigator.org.nz/apps-videos/b/
[3] https://www.nhsinform.scot/illnesses-and-conditions/a-to-z
[4] https://www.medicinenet.com/diseases_and_conditions/article.htm
[5] https://github.com/LasseRegin/medical-question-answer-data
[6] https://ehealthforum.com/health/health_forums.html
[7] https://questiondoctors.com/blog/
[8] https://www.webmd.com/a-to-z-guides/qa

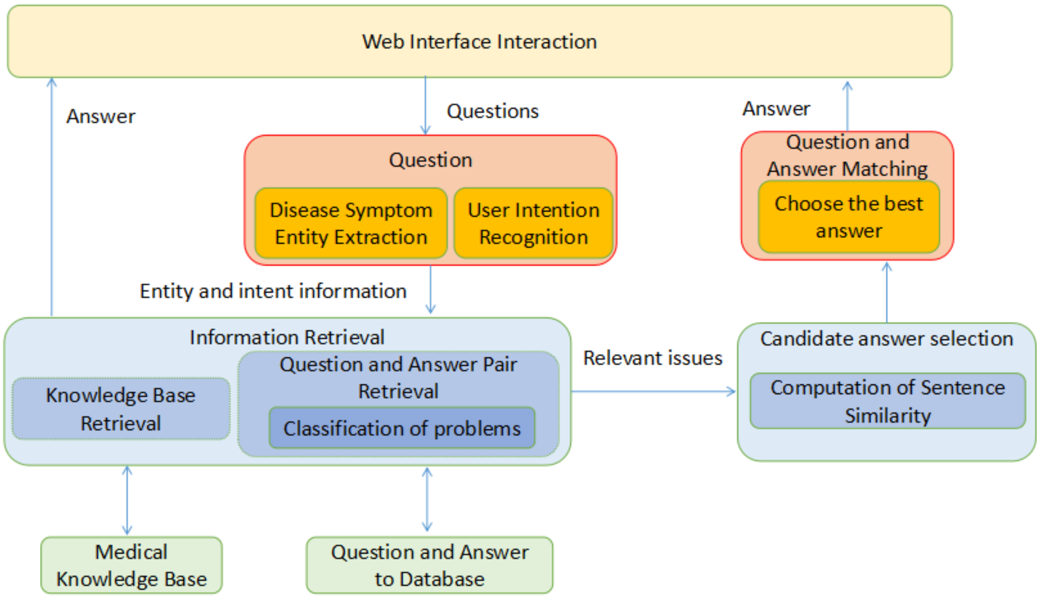

Web Interface Interaction
Answer
Questions
Answer
Question
Disease Symptom Entity Extraction
User Intention Recognition
Question and Answer Matching
Choose the best answer
Entity and intent information
Information Retrieval
Knowledge Base Retrieval
Question and Answer Pair Retrieval
Classification of problems
Relevant issues
Candidate answer selection
Computation of Sentence Similarity
Medical Knowledge Base
Question and Answer to Database

Figure 2: A question-and-answer framework that combines knowledge graph and HBAM

In the next two sections, we describe how we implement the two models for our medical chatbot system.

4 THE KNOWLEDGE GRAPH ARCHITECTURE

4.1 Storage Scale of the Neo4j Graph Database

The system knowledge graph which contains

- 3 entities (department, disease, and symptom),
- 6 properties (name, description, cause, prevent, accompany, cure_way), and
- 5 relationships (have_symptom, accompany_with, disease_prevent, disease_cause, disease_cureway).

There are about 3,500 entities (which include 675 diseases and 2825 symptoms) and 4,500 relationships. The relationship includes the relationship between the diseases, symptoms, and the other 6 properties.

4.2 The Process of Selecting Answers from the Graph Database

The whole process can be divided into five steps: 1. User input question 2. Extract entity with D&S Extractor 3. Get user intention by Intention Recognizer 4. Answer Selection 5. return the answer.

As an example, Figure 3 shows how the word "cold" is detected as a disease keyword by the Entity Extractor, how the intention "has_symptom" is recognized by Intention Recognizer, and how the answer "fever" is selected.

4.3 Design of Problem Analysis Module

Figure 4 shows the disease symptom entity extraction functionality of the system. This function extracts the disease keywords from a medical keywords dictionary and is performed using the Aho-Corasick algorithm [1]. If the Aho-Corasick algorithm does not identify the disease and symptom entities in the given question, it will enter the semantic similarity calculation module, and search for the most similar k entities in semantics. The user interaction recognition is to predict the user intention by some pre-defined predicate libraries. If the pre-defined predicate libraries do not recognize the intention of the given question, it will prompt the user to ask again, which means the system cannot understand the meaning of the question. Overall, Six typical questions can be answered by our system, based on the five relationships.

5 HIERARCHICAL BILSTM ATTENTION MODEL

The diagram of the new hierarchical BiLSTM Attention model we proposed is shown below in Figure 5.

It is designed for semantic similarity comparison. The whole structure based on a Siamese LSTM framework [14]. We apply one BiLSTM layer and one word attention layer into the Siamese framework. The bottom left, and the right sentences represent user input query and the question from the QA dataset. The two questions will be represented by using word embedding [13] firstly and then using BiLSTM [17] to form the whole sentence embedding based on the context. After that, each BiLSTM encoder will be multiplied by a word attention value, which can be assumed as a weight to highlight the key-point in a sentence. Context vector will be combined with attention to understanding the sentence representation u_w

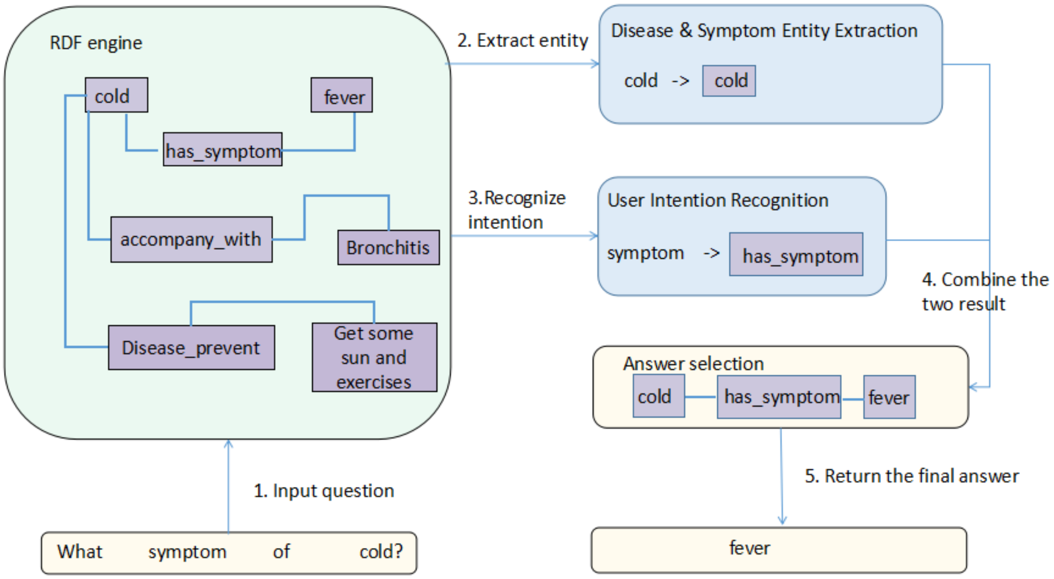


Figure 3: Answer feedback from knowledge graph

[22]. Finally, the similarity value will be computed by the weighted sum of each hidden state value h_{ij} multiply its attention value. The details will be shown in the subsections below.

LSTM-based sequence encoder Long Short Term Memory networks (LSTMs) are proposed by Hochreiter & Schmidhuber [9]. One of the critical things of LSTMs is the cell state. The state of the cell can be regarded as a sort of conveyor belt. With a few linear interactions, it goes straightly down the entire chain. It is straightforward for information to flow along with it unchanged.

The basic principle of LSTM can be divided into three steps. The first step shows which information will be forgotten by the cell state. The “forget gate layer” makes this choice by combing the h_{t-1}

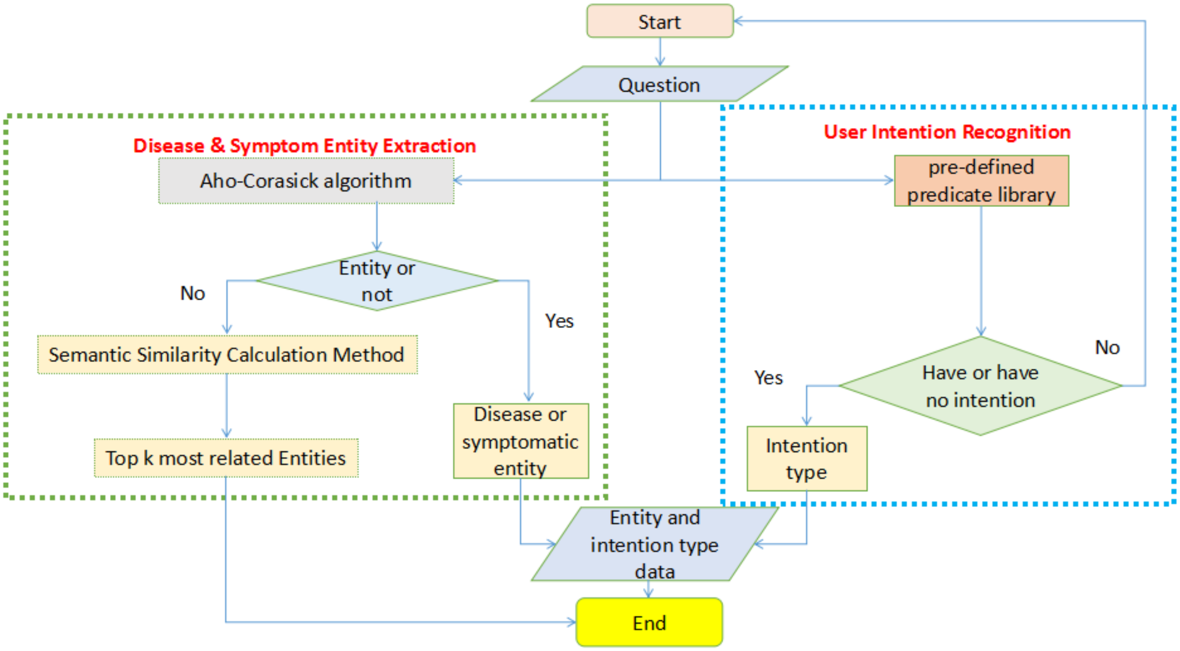

Start
Question
Disease & Symptom Entity Extraction
Aho-Corasick algorithm
Entity or not
No
Yes
Semantic Similarity Calculation Method
Top k most related Entities
Disease or symptomatic entity
User Intention Recognition
pre-defined predicate library
Have or have no intention
Yes
No
Intention type
Entity and intention type data
End

Figure 4: Entity detection and Intention recognition

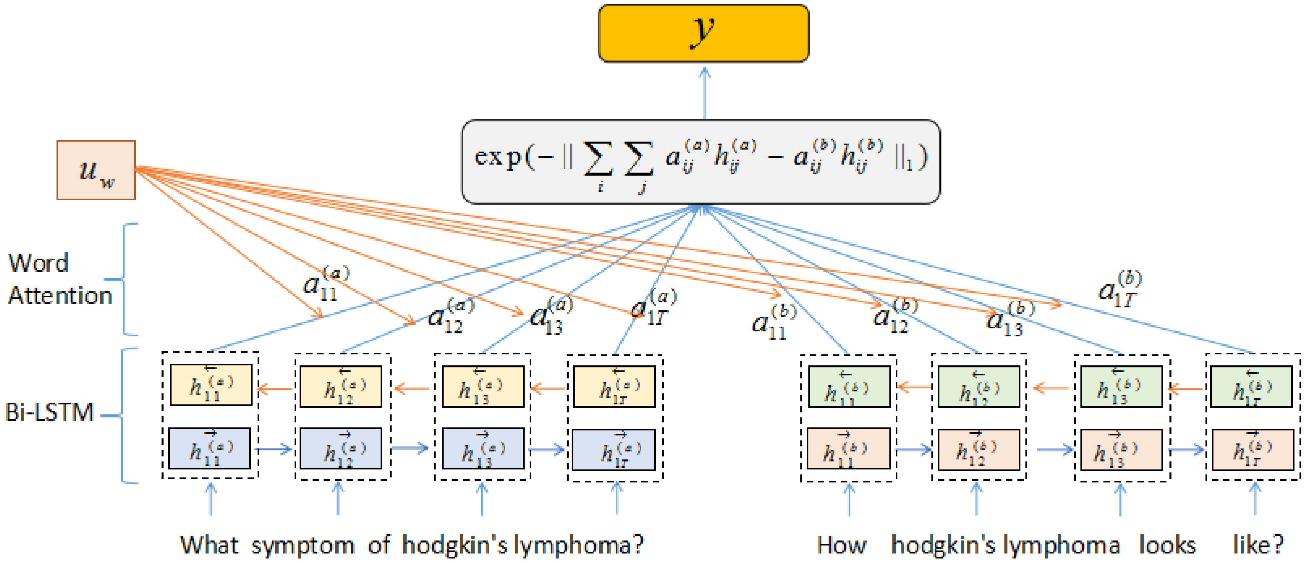

y
$\exp(-\| \sum_i \sum_j a_{ij}^{(a)} h_{ij}^{(a)} - a_{ij}^{(b)} h_{ij}^{(b)} \|_1)$
u_w
Word Attention
Bi-LSTM
$a_{11}^{(a)}$
$a_{12}^{(a)}$
$a_{13}^{(a)}$
$a_{1T}^{(a)}$
$a_{11}^{(b)}$
$a_{12}^{(b)}$
$a_{13}^{(b)}$
$a_{1T}^{(b)}$
$\overleftarrow{h}_{11}^{(a)}$ $\overleftarrow{h}_{12}^{(a)}$ $\overleftarrow{h}_{13}^{(a)}$ $\overleftarrow{h}_{1T}^{(a)}$
$\overrightarrow{h}_{11}^{(a)}$ $\overrightarrow{h}_{12}^{(a)}$ $\overrightarrow{h}_{13}^{(a)}$ $\overrightarrow{h}_{1T}^{(a)}$
$\overleftarrow{h}_{11}^{(b)}$ $\overleftarrow{h}_{12}^{(b)}$ $\overleftarrow{h}_{13}^{(b)}$ $\overleftarrow{h}_{1T}^{(b)}$
$\overrightarrow{h}_{11}^{(b)}$ $\overrightarrow{h}_{12}^{(b)}$ $\overrightarrow{h}_{13}^{(b)}$ $\overrightarrow{h}_{1T}^{(b)}$
What symptom of hodgkin's lymphoma?
How hodgkin's lymphoma looks like?

Figure 5: Hierarchical BiLSTM Attention Model

and x_t, which means the value of the hidden layer at time $t-1$ and the value of the input layer at time t. W_g means the weight matrix between the hidden layer and output layer, and b_g means the bias vector. We can get the g_t through this formula, which decides to filter out that unimportant information.

$$g_t = \sigma\left(W_g \cdot [h_{t-1}, x_t] + b_g\right) \tag{1}$$

The second step is to choose which new information will be stored in the state of the cell. The input gate layer determines which values are updated. A tanh layer produces a vector that can be added to consider whether the new candidate values $\tilde{C}_t$ should be updated in the state. Then, these two values will be merged to generate a state update.

$$j_t = \sigma\left(W_j \cdot [h_{t-1}, x_t] + b_j\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{3}$$

The time that decides whether the old cell state C_{t-1} will be updated by the new cell state C_t is depending on the previous steps. The old state multiplied by g_t, and the forgetting things will be chosen to forget earlier. After that $j_t * \tilde{C}_t$ will be added. So, there will be a new candidate value, ranged by how much each state value has been chosen to update.

Finally, the output will be decided. The output will be filtered by the cell state. Then a sigmoid layer will be operated that is chosen which components of the cell state will be output. The cell state will be put through tanh ranging from -1 to 1 and multiplied it by the sigmoid gate output so that the output will be decided by the chosen components.

$$q_t = \sigma\left(W_q\,[h_{t-1}, x_t] + b_q\right) \tag{4}$$

$$h_t = q_t \times \tanh\left(C_t\right) \tag{5}$$

Word Attention Given a sentence $w_{it}, t \in [0, T]$. Firstly, each word of the sentence will be embedded by using a embedding matrix W_e.

$$x_{it} = W_e w_{it}, t \in [1, T] \tag{6}$$

We use Bidirectional LSTM [18] to capture both forward and reverse direction information of each word. The bidirectional LSTM contains forward LSTM $\vec{f}$ and reverse LSTM $\overleftarrow{f}$.

$$\vec{h}_{it} = \overrightarrow{\mathrm{LSTM}}\left(x_{it}\right), t \in [1, T] \tag{7}$$

$$\vec{h}_{ti} = \overrightarrow{\mathrm{LSTM}}\left(x_{ti}\right), t \in [1, T] \tag{8}$$

In order to represent those keywords in a sentence. We try to use Attention. Firstly, we feed the h_{it} into the tanh function to get u_{it} as a hidden representation of h_{it}. Secondly, we calculate the importance of each word u_{it} and get a normalized importance weight α_{it} by using a softmax function. Then, we calculate the sentence vector $\boldsymbol{s}_i$ as a weight sum of each word with its weight.

$$u_{it} = \tanh\left(W_w h_{it} + b_w\right) \tag{9}$$

$$\alpha_{it} = \frac{\exp\left(u_{it}\right)}{\sum_t \exp\left(u_{it}\right)} \tag{10}$$

$$s_i = \sum_t \alpha_{it} h_{it} \tag{11}$$

Similarity function

$$f\left(s_i^{(a)}, s_j^{(b)}\right) = \exp\left(-\left\|\sum_i \sum_j a_{ij}^{(a)} h_{ij}^{(a)} - a_{ij}^{(b)} h_{ij}^{(b)}\right\|_1\right) \in [0, 1] \tag{12}$$

The formula is based on Manhattan distance. From this formula, the representation from two sentences can be represented by $s_i^{(a)} = \sum_i \sum_j a_{ij}^{(a)} h_{ij}^{(a)}$ and $s_j^{(b)} = \sum_i \sum_j a_{ij}^{(b)} h_{ij}^{(b)}$. $a_{ij}^{(a)}$ and $a_{ij}^{(b)}$ mean the

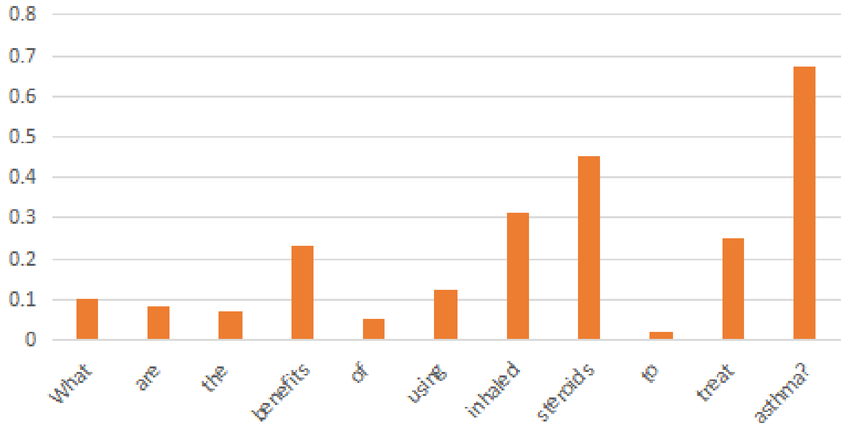
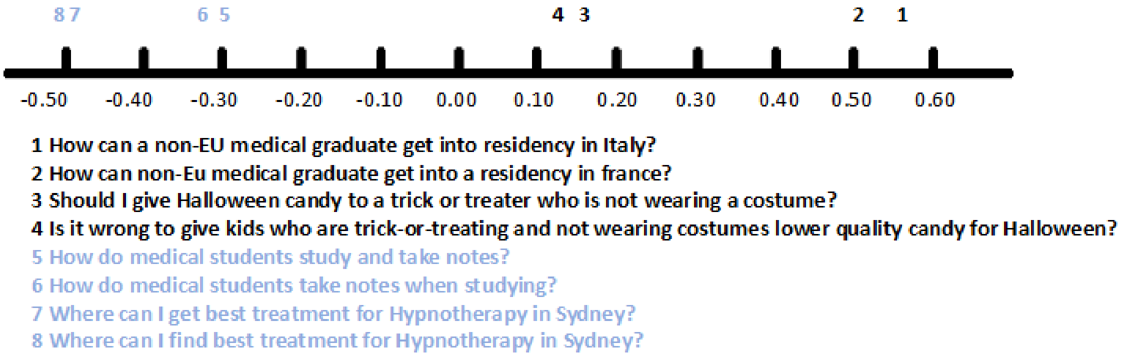


Figure 6: The word attention distribution in the two sentences



1 How can a non-EU medical graduate get into residency in Italy?
2 How can non-Eu medical graduate get into a residency in france?
3 Should I give Halloween candy to a trick or treater who is not wearing a costume?
4 Is it wrong to give kids who are trick-or-treating and not wearing costumes lower quality candy for Halloween?
5 How do medical students study and take notes?
6 How do medical students take notes when studying?
7 Where can I get best treatment for Hypnotherapy in Sydney?
8 Where can I find best treatment for Hypnotherapy in Sydney?

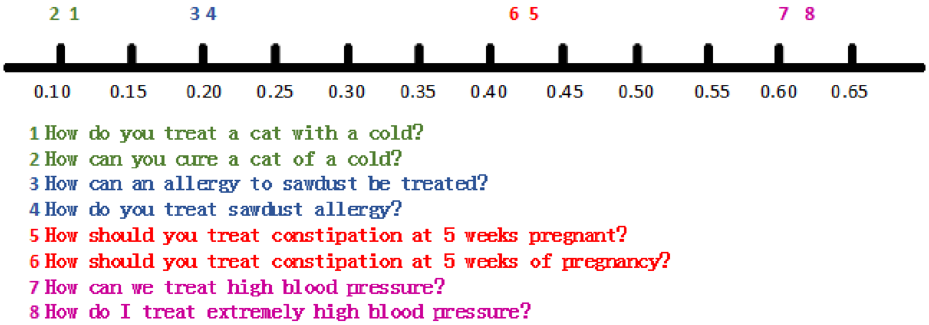
Figure 7: The comparison of positive representation and the negative representation

attention value in both direction. $h_{ij}^{(a)}$ and $h_{ij}^{(b)}$ mean the hidden state value in both direction.

6 RESULTS AND ANALYSIS

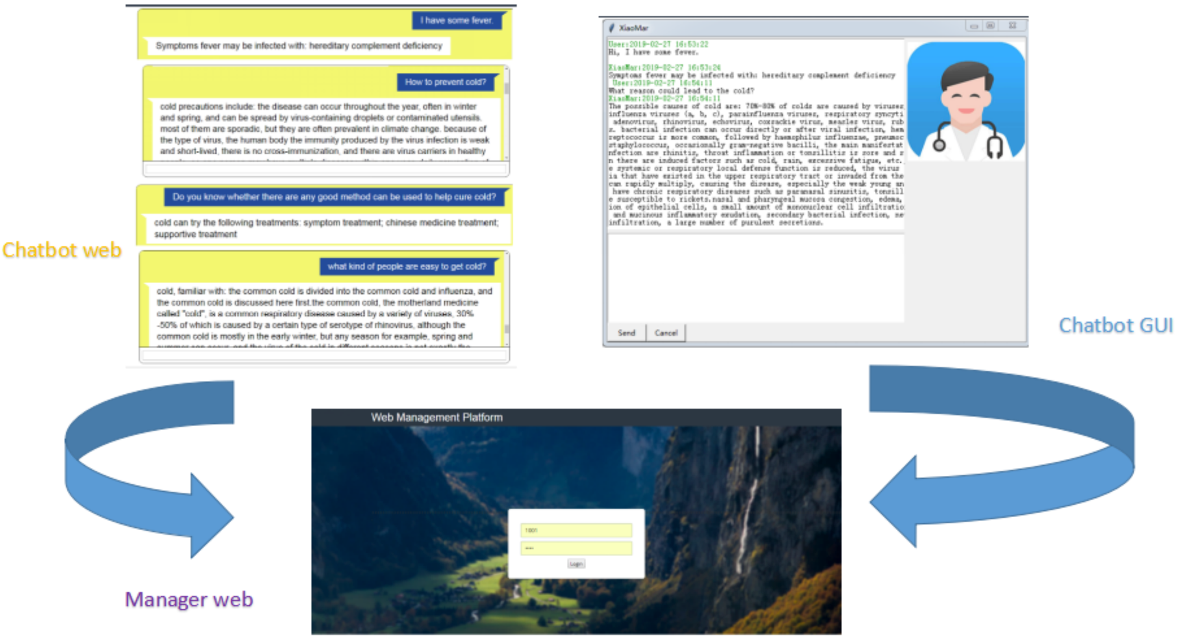
6.1 Sample Results

In order to explain the process of calculating the similarity of medical questions in HBAM model, we found that HBAM has an impressive text representation in understanding the word meaning



1 How do you treat a cat with a cold?
2 How can you cure a cat of a cold?
3 How can an allergy to sawdust be treated?
4 How do you treat sawdust allergy?
5 How should you treat constipation at 5 weeks pregnant?
6 How should you treat constipation at 5 weeks of pregnancy?
7 How can we treat high blood pressure?
8 How do I treat extremely high blood pressure?

Figure 8: The distribution of a different group of sentence representation



I have some fever.

Symptoms fever may be infected with: hereditary complement deficiency

How to prevent cold?

cold precautions include: the disease can occur throughout the year, often in winter and spring, and can be spread by virus-containing droplets or contaminated utensils. most of them are sporadic, but they are often prevalent in climate change. because of the type of virus, the human body the immunity produced by the virus infection is weak and short-lived, there is no cross-immunization, and there are virus carriers in healthy

Do you know whether there are any good method can be used to help cure cold?

cold can try the following treatments: symptom treatment; chinese medicine treatment; supportive treatment

what kind of people are easy to get cold?

cold, familiar with: the common cold is divided into the common cold and influenza, and the common cold is discussed here first.the common cold, the motherland medicine called "cold", is a common respiratory disease caused by a variety of viruses, 30% -50% of which is caused by a certain type of serotype of rhinovirus, although the common cold is mostly in the early winter, but any season for example, spring and

```
User:2019-02-27 16:53:22
Hi, I have some fever.

XiaoMar:2019-02-27 16:53:24
Symptoms fever may be infected with: hereditary complement deficiency
 User:2019-02-27 16:54:11
What reason could lead to the cold?
XiaoMar:2019-02-27 16:54:11
The possible causes of cold are: 70%-80% of colds are caused by viruses
influenza viruses (a, b, c), parainfluenza viruses, respiratory syncyti
 adenovirus, rhinovirus, echovirus, coxsackie virus, measles virus, rub
s. bacterial infection can occur directly or after viral infection, hem
reptococcus is more common, followed by haemophilus influenzae, pneumoc
staphylococcus, occasionally gram-negative bacilli, the main manifestat
nfection are rhinitis, throat inflammation or tonsillitis is sore and s
n there are induced factors such as cold, rain, excessive fatigue, etc.
e systemic or respiratory local defense function is reduced, the virus
ia that have existed in the upper respiratory tract or invaded from the
can rapidly multiply, causing the disease, especially the weak young an
 have chronic respiratory diseases such as paranasal sinusitis, tonsill
e susceptible to rickets.nasal and pharyngeal mucosa congestion, edema,
ion of epithelial cells, a small amount of mononuclear cell infiltratio
 and mucinous inflammatory exudation, secondary bacterial infection, ne
infiltration, a large number of purulent secretions.
```

Figure 9: Chatbot Website, GUI and Manager Website

and weight distribution of a sentence, as well as the distribution of the meaning of the sentence.

Each word in a sentence is represented by weight according to the word attention mechanism. As an example, Figure 6 shows a sentence "`What are the benefits of using inhaled steroids to treat asthma?`" The words "`asthma`" and "`steroids`" are given higher weight.

The sentence representation can be seen in Figure 7 and Figure 8, respectively. The former shows the sentence vector distribution comparison between the positive representation and the negative representation. The latter shows the distribution of a different group of sentence representation. The 1st and 2nd sentences both represent the meaning related to the cold. The 3rd and 4th sentences both represent the meaning related to the allergy. The 5th and 6th sentences both represent the meaning related to the pregnant, and the 7th and 8th sentences both represent the meaning related to the high blood pressure.

Figure 9 shows a single-turn conversation example in the HHH. The image on the top-left displays the online chatbot interface, and the one on the top-right shows the chatbot GUI. They are both managed by the manager website (image at the bottom) - the Github link includes code and data [9]. In the following, we give further quantitative analysis on the system performance using two sets of experiments.

6.2 Experiment 1

6.2.1 Train and Test Dataset. The HBAM is trained with the data from Quora duplicate questions dataset [10]. To filter out the medical subset from the dataset, we create a disease and symptom dictionary which contains medical keywords such as cold, obesity, weight loss, and low temperature according to two New Zealand medical website [11] [12]. The number of disease and symptom keywords in the dictionary is 668 and 2367, respectively. With the dictionary, we collect nearly 70,000 medical-related records from the Quora dataset. For training the models faster and easier, we randomly select 10,000 records (positive: negative = 1:1) as the experiment data. The results of Experiment 2 in Section 6.3 will demonstrate that the performances of the models for the remaining records are similar.

The Quora duplicate questions dataset is an open domain sentence pair dataset. It has more than 400,000 tagged sentence pairs formatted like "`text1 text2 is_duplicate`" means whether the two sentences are semantically similar. If they are semantically equal, the tag will be "1", otherwise "0". Some examples are list in table 1.

6.2.2 Environment. We have experimented the deep learning models on Google Colab[13] (Tesla K80 GPU, 12 GB RAM) to validate the semantic similarity between two sentences. The hyperparameters of HBAM includes the batch_size is 1024, the n_epoch is 9, the n_hidden is 100, the embedding_dim is 300 and the max_seq_length

[9] https://github.com/14H034160212/HHH-An-Online-Question-Answering-System-for-Medical-Questions

[10] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

[11] https://www.nhsinform.scot/symptoms-and-self-help/a-to-z

[12] https://www.healthnavigator.org.nz/health-a-z/

[13] https://colab.research.google.com/notebooks/welcome.ipynb

Table 1: Some examples in Quora medical subset

id	qid1	qid2	question1	question2	is_duplicate
130859	209926	209927	How do you treat a cat with a cold?	How can you cure a cat of a cold?	1
82425	139763	133638	How much medical evidence is there in support of the claim weed causes cancer?	Does weed give you lung cancer?	1
261370	377490	377491	How can an allergy to sawdust be treated?	How do you treat sawdust allergy?	1
...	...	...	...	...	...

Table 2: Methods comparison

Methods	Average Evaluation Accuracy	Range of change by 30 times experiments
BERT [7]	78.2%	(-1.8%,+1.3%)
MaLSTM [9]	78.4%	(-2.9%,+2.0%)
HBAM	**81.2%**	(-2.4%,+2.2%)

is 10, GoogleNews-vectors-negative300.bin.gz from Word2Vec[14], the activation function is tanh.

6.2.3 Comparison. To evaluate the performance of our system, we compare it with two state-of-the-art sentence pair similarity algorithms, namely BERT and MALSTM [14]. BERT was proposed by Google in 2018 and has refreshed records in 11 NLP tasks, including Q&A (SQuAD v1.1), reasoning (MNLI), and more. MALSTM was proposed by the MIT team in 2016 and has achieved excellent results in calculating the similarity of sentences. It is better than a few well-known sentence similarity comparison algorithms include Dependency Tree-LSTM, ConvNet, and more. Superior performance over these two benchmark means that our system would have achieved a level that is higher or on par with the current state-of-the-art methods.

In Table 2, we display the results of mapping the medical-related words to query 10000 lines medical-related question pairs. We divide the dataset by 6:2:2 for training, validation, and testing in the BERT baseline model for fine-tuning. In other models, we use 9:1 for training and testing. It can be clearly seen that our HBAM has the best performance to check the duplication of two text sentences.

6.3 Experiment 2

We also perform a second experiment on the remaining more than 50,000 medical sentence pairs as well. We separately select thousands of tags from the three kinds of datasets: ehealthforumQAs, questionDoctorQAs, and webmdQAs, respectively. The tags of each dataset are extracted as the keywords to take the intersection with the disease symptom keyword dictionary. Then the intersection results are used to search for matching sentence pairs in the remaining 50,000 medical sentence pairs. Finally, the first 1000 matched sentence pairs are taken out for each dataset, and 10 times evaluation results are obtained, respectively.

Table 3 shows the evaluation results by experiment 10 times. From the three tables, we believe that the HBAM performs better prediction performance in the three test cases.

In Experiment 2, we reuse the models trained from Section 6.2, but test with different datasets. Nevertheless, it has turned out that the accuracy of the three models has not changed in a significant way.

Table 3: Evaluation result for the three medical websites

Medical website name	Method name	Average predict accuracy	Range of change by 10 times experiments
ehealthforumQAs	BERT	78.5%	(-1.8%,+1.1%)
	HBAM	**81.3%**	(-1.2%,+1.1%)
	MaLSTM	78.4%	(-2.9%,+1.5%)
questionDoctorQAs	BERT	78.2%	(-1.4%,+0.9%)
	HBAM	**80.9%**	(-2.1%,+2.5%)
	MaLSTM	78.1%	(-1.7%,+1.9%)
webmdQAs	BERT	78.1%	(-1.6%,+0.9%)
	HBAM	**81.2%**	(-1.2%,+1.3%)
	MaLSTM	78.5%	(-1.5%,+1.9%)

[14] https://code.google.com/archive/p/word2vec/

7 CONCLUSION AND FUTURE WORK

In this paper, we propose a chatbot framework based on a knowledge graph and a text representation and similarity model. The advantage of the knowledge graph lies in that it utilizes structured storage so that it may help easy maintenance and retrieval of domain-specific knowledge. While the advantage of the attention model utilizes deep learning to represent better and comprehend natural language questions. Therefore, we develop a system that combines the advantages of both models by integrating a knowledge graph with a neural-based model. We compare the ability to achieve the text-similarity between some state-of-the-art NLP models and our new HBAM. We consider the scenarios of single-turn question-and-answer dialogue, use the method of the knowledge graph, and combine deep learning methods to present data well. We speculate that one reason that HBAM is better than MaLSTM is by adding the attention layer to help capture the medical keywords from a sentence. Besides, two possible reasons for HBAM's superior performance over BERT is because of BERT is pre-trained based on a general word embedding and the 12-layer transformer which could cause overfitting when we try to capture and understand the medical keywords from a sentence.

As future work, we foresee the potential of chatbot technologies to play a much more significant role in the medical domain. For example, a software chatbot can be deployed in the real-world to become home healthcare robots or hospital medical inquiry robots. From the application point of view, the paper only considered the single-turn question-and-answer mechanism. An important future direction is to add user profiles into the system and provide a more precise medical assistant to each specific user. Besides, we can combine the data mining method and predict the potential diseases in a region of the population. We plan to recruit some participants to help to evaluate our medical QA system. Also, we hope in the future our chatbot framework can have a chance to be applied in other domains besides healthcare.

8 ACKNOWLEDGEMENT

The authors would like to thank Yang Chen who provides many useful suggestions. This research was supported by summer scholarship funding from the Precision Driven Health research partnership.

REFERENCES

[1] Alfred V Aho and Margaret J Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975.

[2] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, 2017.

[3] Vint Cerf. Parry encounters the doctor. Technical report, 1973.

[4] Ke Chen and Ahmad Salman. Extracting speaker-specific information with a regularized siamese deep network. In *Advances in Neural Information Processing Systems*, pages 298–306, 2011.

[5] Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. Kbqa: learning question answering over qa corpora and knowledge bases. *Proceedings of the VLDB Endowment*, 10(5):565–576, 2017.

[6] Zihang Dai, Lei Li, and Wei Xu. Cfo: Conditional focused neural question answering with large-scale knowledge bases. *arXiv preprint arXiv:1606.01994*, 2016.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Gillian Dobbie and Kevin Ross. Precision driven health: A new zealand research partnership. *International Journal of Integrated Care*, 17(3), 2017.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Peter Jones, Linda Chalmers, Susan Wells, Shanthi Ameratunga, Peter Carswell, Toni Ashton, Elana Curtis, Papaarangi Reid, Joanna Stewart, Alana Harper, et al. Implementing performance improvement in new zealand emergency departments: the six hour time target policy national research project protocol. *BMC health services research*, 12(1):45, 2012.

[11] Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. Neural network-based question answering over knowledge graphs on word and character level. In *Proceedings of the 26th international conference on World Wide Web*, pages 1211–1220. International World Wide Web Conferences Steering Committee, 2017.

[12] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[14] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[15] Lin Ni and Jiamou Liu. A framework for domain-specific natural language information brokerage. *Journal of Systems Science and Systems Engineering*, 27(5):559–585, 2018.

[16] Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. Mandy: Towards a smart primary care chatbot application. In *International Symposium on Knowledge and Systems Sciences*, pages 38–52. Springer, 2017.

[17] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[18] Ming Tan, Cicero Dos Santos, Bing Xiang, and Bowen Zhou. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 464–473, 2016.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[20] Yue Wang, Richong Zhang, Cheng Xu, and Yongyi Mao. The apva-turbo approach to question answering in knowledge base. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1998–2009, 2018.

[21] Joseph Weizenbaum et al. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[22] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[23] Scott Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. 2015.

[24] Wen-tau Yih, Kristina Toutanova, John C Platt, and Christopher Meek. Learning discriminative projections for text similarity measures. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 247–256. Association for Computational Linguistics, 2011.

[25] Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Improved neural relation detection for knowledge base question answering. *arXiv preprint arXiv:1704.06194*, 2017.

[26] Helen Jiahe Zhao and Jiamou Liu. Finding answers from the word of god: Domain adaptation for neural networks in biblical question answering. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.