ChiMed: A Chinese Medical Corpus for Question Answering

Yuanhe Tian
Department of Linguistics
University of Washington
yhtian@uw.edu

Weicheng Ma
Computer Science Department
New York University
wm724@nyu.edu

Fei Xia
Department of Linguistics
University of Washington
fxia@uw.edu

Yan Song
Tencent AI Lab
clksong@gmail.com

Abstract

Question answering (QA) is a challenging task in natural language processing (NLP), especially when it is applied to specific domains. While models trained in the general domain can be adapted to a new target domain, their performance often degrades significantly due to domain mismatch. Alternatively, one can require a large amount of domain-specific QA data, but such data are rare, especially for the medical domain. In this study, we first collect a large-scale Chinese medical QA corpus called *ChiMed*; second we annotate a small fraction of the corpus to check the quality of the answers; third, we extract two datasets from the corpus and use them for the relevancy prediction task and the adoption prediction task. Several benchmark models are applied to the datasets, producing good results for both tasks.

1 Introduction

In the big data era, it is often challenging to locate the most helpful information in many real-world applications, such as search engine, customer service, personal assistant, etc. A series of NLP tasks, such as text representation, text classification, summarization, keyphrase extraction, and answer ranking, are able to help QA systems in finding relevant information (Siddiqi and Sharan, 2015; Allahyari et al., 2017; Yang et al., 2016; Joulin et al., 2016; Song et al., 2017, 2018).

Currently, most QA corpora are built for the general domain focusing on extracting/generating answers from articles, such as CNN/Daily Mail (Hermann et al., 2015), SQuAD (Rajpurkar et al., 2016), Dureader (He et al., 2017), SearchQA (Dunn et al., 2017), CoQA (Reddy et al., 2018), etc., with few others from community QA forums, such as TrecQA (Wang et al., 2007), WikiQA (Yang et al., 2015), and SemEval-2015 (Nakov et al., 2015).

In the medical domain, most medial QA corpora consist of scientific articles, such as BioASQ (Tsatsaronis et al., 2012), emrQA (Pampari et al., 2018), and CliCR (Šuster and Daelemans, 2018). Although some studies were done for conversational datasets (Wang et al., 2018a,b), corpora designed for community QA are extremely rare. Meanwhile, given that many online medical service forums have emerged (e.g. MedHelp[1]), there are increasing demands from users to search for answers for their medical concerns. One might be tempted to build QA corpora from such forums. However, in doing so, one must address a series of challenges such as how to ensure the quality of the derived corpus despite the noise in the original forum data.

In this paper, we introduce our work on building a Chinese medical QA corpus named *ChiMed* by crawling data from a big Chinese medical forum[2]. In the forum, the questions are asked by web users and all the answers are provided by accredited physicians. In addition to (Q, A) pairs, the corpus contains rich information such as the title of the question, key phrases, age and gender of the user, the name and affiliation of the accredited physicians who answer the question, and so on. As a result, the corpus can be used for many NLP tasks. In this study, we focus on two tasks: relevancy prediction (whether an answer is relevant to a question) and adoption prediction (whether an answer will be adopted).

[1] https://www.medhelp.org

[2] The code for constructing the corpus and the datasets used in this study are available at https://github.com/yuanheTian/ChiMed.

Proceedings of the BioNLP 2019 workshop, pages 250–260
Florence, Italy, August 1, 2019.

# of As per Q	# of Qs	% of Qs
1	5,517	11.8
2	39,098	83.7
≥3	2,116	4.5
Total	46,731	100.0

Table 1: Statistics of *ChiMed* with respect to the number of answers (As) per question (Q).

2 The *ChiMed* Corpus

To benefit NLP research in the medical domain, we create a Chinese medical corpus (*ChiMed*). This section describes how the corpus was constructed, the main content of the corpus, and and its potential usage.

2.1 Data Collection

Ask39[3] is a large Chinese medical forum where web users (to avoid confusion, we will call them *patients*) can post medical questions and receive answers provided by licensed physicians. Each question, together with its answers and other related information (e.g., the names of physicians and similar questions), is displayed on a page (aka a QA page) with a unique URL. Currently, approximately 145 thousand forum-verified physicians have joined the forum to answer questions and there are 17.6 million QA pages. We started with fifty thousand URLs from the URL pool and downloaded the pages using the selenium package[4]. After removing duplicates or pages with no answers, 46,731 pages remain and most of the questions (83.7%) have two answers (See Table 1).

2.2 QA Records

From each QA page, we extract the question, the answers and other related information, and together they form a *QA record*. Table 2 displays the main part of a QA record, which has five fields that are most relevant to this study: (1) "*Department*" indicates which medical department the question is directed to;[5] (2) "*Title*" is a brief description of disease/symptoms (5-20 characters); (3) "*Question*" is a health question with a more detailed description of symptoms (at least 20 characters); (4) "*Keyphrases*" is a list of phrases related to the question and the answer(s); (5) The last field is a list of *answers*, and each answer has an *Adopted* flag indicating whether it has been adopted. Among the five fields, *Title* and *Question* are entered by patients; *Answers* are provided by physicians; *Department* is determined by the forum engine automatically when the question is submitted. As for the *Keyphrases* field and the *Adopted* flag, it is not clear to us whether they are created manually (if so, by whom) or generated automatically.[6] In addition to these fields, a QA record also contains other information such as the name and affiliation of the physicians who answer the question, the patient's gender and age, etc.

Table 3 shows the statistics of *ChiMed* in terms of QA records. On average, each QA record contains one question, 1.96 answers, and 4.48 keyphrases. Overall, 69.1% of the answers in the corpus have an adopted flag.

2.3 Potential Usage of the Corpus

Given the rich content of the QA record, ChiMed can be used in many NLP tasks. For instance, one can use the corpus for **text classification** (to predict the medical department that a Q should be directed to), **text summarization** (to generate a title given a Q), **keyphrase generation** (to generate keyphrases given a Q and/or its As), **answer ranking** (to rank As for the same Q, if adopted As are indeed better than unadopted As), and **question answering** (retrieve/generate As given a Q).

Because the content of the corpus comes from an online forum, before we use the corpus for any NLP task, it is important to check the quality of the corpus with respect to that task. As a case study, for the rest of the paper, we will focus on three closely related tasks, all taking a question and an answer (or a set of answers) as the input: The first one determines whether the answer is relevant to the question; the second determines whether the answer will be adopted for the question (as indicated by the *Adopted* flag in the corpus); the third one ranks all the answers for the question if there are more than one answer. We name them the *relevancy task*, the *adoption prediction task*, and the *answer ranking task*, respectively. The first two are binary classification tasks, while the last one is a ranking task. In the next section, we will manually check a small fraction of the corpus to determine whether its quality is high for those tasks.

[3] http://ask.39.net

[4] https://github.com/SeleniumHQ/selenium

[5] There are 13 departments such as pediatrics, infectious diseases, and internal medicine.

[6] We have made many attempts to no avail to contact the forum about those and other questions.

Department	内科 > 淋巴增生 Internal Medicine > Lymphocytosis
Title	胃部淋巴增生会癌变吗？ Will lymphatic hyperplasia in the stomach cause cancer?
Question	我最近检查出患有胃部淋巴增生的疾病，非常担心，请问它会癌变吗? I recently checked out the disease of lymphoid hyperplasia in the stomach. I am very worried. Will it cause cancer?
Keypharses	慢性浅表性胃炎，幽门螺旋杆菌感染，淋巴增生，胃，消化 Chronic superficial gastritis, Helicobacter pylori infection, lymphatic hyperplasia, stomach, digestion
Answer 1	这一般是幽门螺旋杆菌感染造成的，一般不会造成癌变，所以不必惊慌。建议饮食规律，吃易消化的食物，细嚼慢咽，少量多餐，禁食刺激性食物。 In general, this is caused by Helicobacter pylori infection and does not cause cancer. So do not panic. It is recommended to have a regular diet, eat digest friendly food and chew slowly. Do not eat much in one meal and no spicy food is allowed.
Adopted	True
Answer 2	这是普通的慢性胃粘膜炎症，与幽门螺旋杆菌感染有关。可用阿莫西林治疗。 This is a common chronic gastric mucosal inflammation and has a relationship with Helicobacter pylori infection. You can choose amoxicillin for treatment.
Adopted	False

Table 2: An example of QA record in *ChiMed*. The English translation is not part of the corpus.

# of Questions	46,731
# of Answers	91,416
Avg. # of Answers per Question	1.96
# (%) of Answers Adopted	63,153 (69.1%)
# of Keyphrases	209,261
# of Keyphrases per Q	4.48
# of Unique Keyphrases	10,360

Table 3: Statistics of *ChiMed*.

# of Adopted As	# of Qs	% of Qs
0	30	0.08%
1	25,594	65.46%
2	13,474	34.46%
Total	39,098	100%

Table 4: QA records with exactly two answers.

3 Relevancy, Answer Ranking, and Answer Adoption

Given *ChiMed*, it is easy to synthesize a "labeled" dataset for the relevancy task. E.g., given a question, we can treat answers in the same QA record as relevant, and answers in other QA records as irrelevant. The quality of such a synthesized dataset will depend on how often answers in a QA record are truly relevant to the question in the same record. For the adoption prediction task, we can directly use the *Adopted* flag in the QA records.

For the answer ranking task, the answers in a QA record are not ranked. However, if adopted answers are often better than unadopted answers, the former can be considered to rank higher than the latter if both answers come from the same QA record. Table 4 shows among the QA records with exactly two answers, 65.46% of them have exactly one adopted answer and 34.46% have two adopted answers. We can use these 65.46% of QA records as a labeled dataset for the answer ranking task. However, the quality of such a dataset will depend on the correlation between the *Adopted* flag and the high quality of an answer.

To evaluate whether the answers are relevant to the question in the same QA record, and whether adopted answers are better than unadopted ones, we randomly sampled QA records containing exactly two questions, and picked 60 records with exactly one adopted and one unadopted answers (called **Subset-60**) and 40 records with both answers adopted (called **Subset-40**). The union of subset-60 and subset-40 is called **Full-100**, and it contains 100 questions, 200 answers (140 answers are adopted and 60 are not).

3.1 Annotating Relevancy and Answer Ranking

To determine the quality of *ChiMed*, we manually added two types of labels to each QA record in

Possible Relevancy Labels for a (Q, A) pair:

1: The A fully answers the Q
2: The A partially answers the Q
3: The A does not answer the Q
4: Cannot tell whether the A is relevant to Q

Possible Ranking Labels for one Q and two As:

1: The first A is better
2: The second A is better
3: The two As are equally good
4: Neither of As is good (fully answers the Q)
5: Cannot tell which A is better

Properties of Good As:

1: Answer more sub-questions
2: Analyze symptoms or causes of disease
3: Offer advice on treatments or examinations
4: Offer instructions for drug usage
5: Soothe patients' emotions

Properties of Bad As:

1: Answer the Q indirectly
2: The A has grammatical errors
3: Offer irrelevant information

Table 5: Labels and part of annotation Guidelines for relevancy and ranking annotation.

the Full-100 set. The first is *relevancy* label, indicating whether an answer is relevant to a question (i.e., whether the answer field provides a satisfactory answer to the question). There are four possible values as shown in the top part of Table 5.

The second type of labels ranks the two answers for a question. Sometimes, determining which answer is better can be challenging especially when both answers are relevant. Intuitively, people tend to prefer answers that address the question directly, that are easy to understand while supported by evidence, etc. Based on such intuition, we create a set of annotation guidelines, parts of which are shown in the second half of Table 5. Because both types of annotation may require medical expertise, we include a *Cannot tell* label (label "4" for relevancy annotation and label "5" for ranking annotation) for non-expert annotators to annotate different cases.

3.2 Inter-annotator Agreement on Relevancy and Answer Ranking

We hired two annotators without medical background to first annotate the Full-100 set independently and then resolve any disagreement via discussion. The results in terms of percentage and Cohen's Kappa are in Table 6. Inter-annotator agreement on the relevancy label is quite high (90.5% in percentage and 55.6% in kappa), while the agreement on the ranking label is much lower (62.0% in percentage and 43.0% in kappa).

	Relevancy		Ranking	
	%	κ	%	κ
I vs. II	90.5	55.6	62.0	43.0
I vs. Agreed	97.0	83.7	79.0	69.2
II vs. Agreed	93.5	70.4	76.0	64.4

Table 6: Inter-annotator agreement for relevancy and ranking labeling on the Full-100 set in terms of percentage (%) and Cohen's Kappa (κ). I and II refer to the annotations by the two annotators before any discussion, and *Agreed* is the annotation after the annotators have resolved their disagreement.

I \ II	1	2	3	4	Total
1	170	10	0	1	181
2	2	9	1	0	12
3	0	4	2	0	6
4	0	1	0	0	1
Total	172	24	3	1	200

(a) Confusion matrix of two annotators on relevancy labels on the Full-100 set. The agreement is 90.5% (55.6% in Cohen's Kappa) and the four labels are explained in Table 5.

I \ II	1	2	3	4	5	Total
1	25	6	4	1	0	36
2	7	25	5	0	0	37
3	9	4	11	0	0	24
4	0	1	0	1	0	2
5	1	0	0	0	0	1
Total	42	36	20	2	0	100

(b) Confusion matrix of two annotators on ranking labels on the Full-100 set. The agreement is 62.0% (43.0% in Cohen's Kappa) and the five labels are explained in Table 5.

Table 7: The confusion matrices of two annotators on relevancy labels and ranking labels on the Full-100 set.

Table 7a and Table 7b show the confusion matrices of the two annotators on the relevancy annotation and ranking annotation, respectively. Out of four relevancy labels and five ranking labels, relevancy label "3" and ranking label "4" are rare as most answers in the corpus are relevant; relevancy label "4" and ranking label "5" are also rare, but they do occur as sometimes choosing the relevant/better answer requires medical expertise.

Q	我一直卷发，拉直也没用，这是卷毛性综合症吗？怎么治疗呢？ I have curly hair and straightening is useless. Is this a curl syndrome? How to treat it?
A1	卷发有两个原因，一是先天的自然卷发；另一种是后天的不慎引起，如烫发或染发。先天矫直或化学矫直只能是暂时的。除了洗头和护发产品要调整外，避免使用热吹风机。梳头时要小心。不要用头绳或橡皮筋发夹，防止头发拉伤。 There are two reasons for curly hair: one is congenital natural curly hair; the other is caused by inadvertently acquired, such as perming or dyeing hair. Congenital straightening or chemical straightening can only be temporary. In addition to the shampoo and hair care products need to be adjusted, avoid using a hot hair dryer. Be careful when combing your hair. Do not use a headband or rubber band hairpin to prevent hair strain.
A2	自然卷是一种受遗传因素影响的发型。头发自然卷成一卷。形成的原因是由于人类基因的不同。卷发并不是一件坏事。这种自卷曲的类型是药物无法改变的。如果拉直用的是直板，离子是热的，经过熨烫，一段时间后，它就会回到原来的状态。 Natural rolls are a type of hair that is affected by genetic factors. The hair is naturally rolled into a roll. The reason for the formation is due to differences in human genes. Curly hair is not a bad thing. This type of self-curling is that the drug cannot be changed. If the way of straightening is straight, the ions are hot and after ironing, after a while, it will return to its original state.

Table 8: An example where one annotator thinks the two answers are equally good because they both answer the question informatively. The other annotator thinks A1 is better because it tells the patient how to take care of his/her hair in daily life, although A1 provides less analysis of the causes of the symptom. After discussion, the two annotators reach an agreement that advice on daily care is very important and thus A1 is better than A2.

For ranking annotation, disagreement tends to occur when the two answers are very similar. That is why the majority of disagreed annotations (22 out of 38) occur when one annotator chooses one answer to be better while the other annotator considers the two answers to be equally good (an example is given in Table 8). There are 13 examples where annotators have completely opposite annotation (e.g., one annotates "1" while the other annotates "2"), which further shows the difficulty in identifying which answer is better.

3.3 The *Adopted* flag in *ChiMed*

As is mentioned above, each answer in *ChiMed* has a flag indicating whether or not the answer has been adopted. While we do not know the exact meaning of the flag and whether the flag is set manually (e.g., by the staff at the forum) or automatically (e.g., according to factors such as the physicians' past performance or seniority), we would like to know whether the flag is a good indicator of relevant or better answers.

Among four relevancy labels, we regard answers with label "1" or "2" as relevant answers because they fully or partially answer the question, and answers with label "3" or "4" as irrelevant answers. Table 9 shows that 98.0% of the answers in the Full-100 set are considered to be relevant, according to the *Agreed* relevancy annotation. In other words, approximately 98% of (Q, A) pairs in the corpus are good question-answer pairs. On the other hand, the adopted answers are not more likely to be relevant to the question than the unadopted ones. Therefore, the *Adopted* flag is not a good indicator of an answer's relevancy.

	# of As	# (%) of Relevant As
Adopted	140	137(97.9%)
Unadopted	60	59(98.3%)
Total	200	196(98.0%)

Table 9: The *Adopted* flag vs. relevancy label on the Full-100 set. Here, answers with relevancy label "1" or "2" are regarded as relevant answers.

The next question is whether adopted answers tend to be better answers than unadopted ones. If so, we can use the *Adopted* flag to infer ranking labels as follows: if a QA record in the Full-100 set has exactly one adopted answer, we rank that answer higher than the unadopted one in the same record; if both answers in a QA record are adopted, they are considered to be equally good. Table 10 shows such inferred labels do not correlate well with human annotation. In fact, the correlation between inferred labels and the *Agreed* human annotation is only 0.068, when we use the 97 QA records with ranking label "1", "2", or "3". Therefore, the *Adopted* flag is not a good indicator

	Subset-60	Full-100
Adopted vs. I	43.3%	34.0%
Adopted vs. II	46.7%	32.0%
Adopted vs. Agreed	43.3%	36.0%

(a) Agreements between the ranking labels from annotators (I, II, and Agreed) and the labels induced from the adopted flag (Adopted). The Subset-60 is the subset of the Full-100 set where each question has exactly one adopted answer and one unadopted answer (See Section 3).

Adopted \ Agreed	1	2	3	4	5	Total
1	17	6	9	0	1	33
2	7	9	10	1	0	27
3	14	15	10	1	0	40
Total	38	30	29	2	1	100

(b) Confusion matrix between the agreed human annotation and ranking labels induced from the adopted flag. The meaning of the five labels are explained in Table 5.

Table 10: The adopted flags vs. the ranking labels from annotators on the Full-100 set.

for better answers.

So far we have demonstrated that the *Adopted* flag is not a good indicator for relevant or better answers. So what does the *Adopted* flag really indicate? While we are waiting for responses from the Ask39 forum, there are two possibilities. One is that the flag is intended to mean something totally different from relevant or better answers. The other possibility is that the flag intends to mark relevant or better answers but their criteria for relevant or better answers are very different from ours. Table 11 shows a (Q, A) pair, where the answer is adopted. On the one hand, the answer does not directly answer the question. On the other hand, it does provide some useful information about gallstone, and one can argue that the adopted flag in the original corpus is plausible.

3.4 Two Datasets from *ChiMed*

As shown in Table 9, the majority of answers in *ChiMed* are relevant to the questions in the same QA records. To create a dataset for the relevancy task, we start with the 25,594 QA records which have exactly one adopted and one unadopted answer (see Table 4), Next, we filter out QA records whose questions or answers are too long or too short,[7] because very short questions or answers

[7]We will remove a QA record if it contains a question/answer that is ranked either top 1% or bottom 1% of all questions/answers according to their character-based length.

Q	请问为什么胆结石总是晚上发作? Why does gallstone always occur at night?
A	有些人会出现过度劳累、腹胀、打鼾症状。可能是胆结石的原因，且通常晚上疼痛更严重。可以选择药物治疗。手术复发的可能性很大。建议平时多运动。 Some people have symptoms of fatigue, bloating and snoring. They may be caused by gallstones, and usually the pain is more severe at night. You can choose medication. There is a high probability of recurrence of surgery. It is recommended to exercise more usually.

Table 11: The answer does not directly answer the question, but it has an adopted flag.

	Train	Dev	Test
# of Qs	19,952	2,494	2,494
# of As	39,904	4,988	4,988
Avg. Length of Qs	63.5	63.8	63.3
Avg. Length of As in *ChiMed-QA1*	118.7	118.6	118.0
Avg. Length of As in *ChiMed-QA2*	128.0	127.6	127.1

Table 12: Statistics of the two *ChiMed-QA* Datasets. Average lengths of Qs and As are in characters.

tend to be lack of crucial information, whereas very long ones tend to include much redundant or irrelevant information. The remaining dataset contains 24,940 QA records. We divide it into training/development/testing set with portions of 80%/10%/10% and call the dataset *ChiMed-QA1*. Since each QA record has one adopted and one unadopted answer, we will use the dataset to train an adoption predictor.

For the relevancy task, we need both positive and negative examples. We start with *ChiMed-QA1*, and for each QA record, we keep the adopted answer as a positive instance, and replace the unadopted answer with an adopted answer from another QA record randomly selected from the same training/dev/testing subsets to distinguish relevant vs. irrelevant answers. We call this synthesized dataset *ChiMed-QA2*. We will use those two datasets for the adoption prediction task and the relevancy task (see the next section). We are not able to use the corpus for the answer ranking task as we cannot infer the ranking label from the *Adopted* flag.

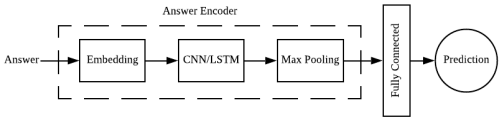

Figure 1: The architecture of CNN- and LSTM-based systems under A-Only setting.

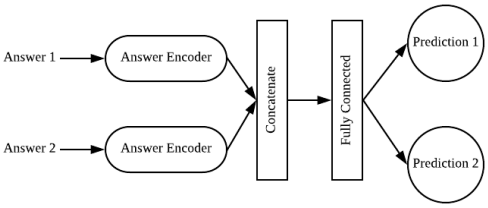

Figure 2: The architecture of our systems under A-A setting. The architecture of answer encoder is identical with the one in Figure 1. Prediction 1 and 2 means the prediction for answer 1 and 2, respectively.

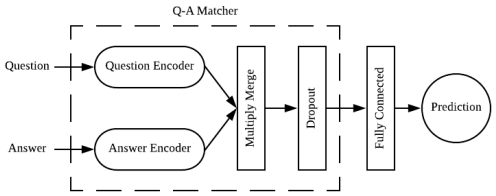

Figure 3: The architecture of our systems under Q-A setting. The architecture of question and answer encoders are identical with the architecture in Figure 1.

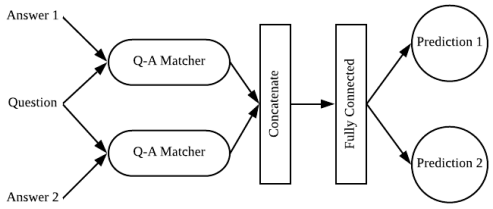

Figure 4: The architecture of our systems under Q-As setting. The architecture of Q-A matcher is shown in Figure 3. We use five Q-A matchers in our experiment: CNN, LSTM, ARC-I, DUET, and DRMM.

Table 12 shows the statistics of the two datasets. The first three rows are the same for the two datasets; the average length of As in *ChiMed-QA2* is slightly longer than that in *ChiMed-QA1* because adopted answers tend to be longer than unadopted ones.

4 Experiments on Two Prediction Tasks

In this section, we use *ChiMed-QA1* and *ChiMed-QA2* (See Table 12) to build NLP systems for the adoption prediction task and the relevancy prediction task, respectively. Both tasks are binary classification tasks with the same type of input; the only difference is the meaning of class labels (relevancy vs. adopted flag). Therefore, we build a set of NLP systems and apply them to both tasks.

4.1 Systems and Settings

We implemented both CNN- and LSTM-based systems, and applied three state-of-the-art sentence matching systems to the two tasks. The three existing systems are: (1) **ARC-I** (Hu et al., 2014) matches questions and answers by directly concatenating their embeddings; (2) **DUET** (Mitra et al., 2017) computes the Q-A similarity by matching exact terms and high-level sentence embeddings (Hadamard production) simultaneously; (3) **DRMM** (Guo et al., 2016) makes its final prediction based on the similarity matrix of each pair of word embeddings in a question and an answer.

We run our CNN- and LSTM-based systems under four different settings: (1) **A-Only** where an answer is the only input (See Figure 1); (2) **A-A** where both answers are input (See Figure 2); (3) **Q-A** where a question and one of its answers are input (See Figure 3); (4) **Q-As** where a question and both of its answers are input (See Figure 4). ARC-I, DUET, and DRMM are run under the settings of Q-A and Q-As, because the systems require a question to be one of the input. The reason we apply the A-Only and A-A settings to the adoption prediction task is that it helps identify whether features from an answer itself will contribute to its adopted flag assignment without knowing its question. To compare the relevancy task and the adoption prediction task, we also apply these two settings to the former task although they are not common settings in previous studies (Lai et al., 2018).

Word segmentation has always been a challenge in Chinese NLP especially when it is applied to a particular domain (Song et al., 2012; Song and Xia, 2012, 2013). Therefore, instead of word embeddings (Song et al., 2018), we use Chinese-character-based embeddings to avoid word segmentation errors. We set the embedding size to 150. We use 155 and 245 as the lengths of questions and answers respectively. Short texts are padded with blank characters. We use 32 filters

Sys ID	Input Setting	NLP System	Relevancy Prediction		Adoption Prediction	
			-CR	+CR	-CR	+CR
1	A-Only	CNN	50.80	51.64	74.10	81.64
2		LSTM	50.66	50.72	74.24	82.00
3	A-A	CNN	49.40	-	84.20	-
4		LSTM	50.28	-	85.00	-
5	Q-A	CNN	74.32	81.84	74.84	81.07
6		LSTM	80.19	87.09	75.28	83.64
7		ARC-I	50.34	50.60	75.20	82.64
8		DUET	81.03	91.74	75.28	82.48
9		DRMM	93.60	98.16	71.49	83.88
10	Q-As	CNN	76.98	-	83.52	-
11		LSTM	88.41	-	84.24	-
12		ARC-I	48.84	-	83.88	-
13		DUET	87.17	-	83.36	-
14		DRMM	98.32	-	83.28	-

Table 13: Results of all systems under different settings with respect to (Q, A) pair prediction accuracy with (+CR) and without (-CR) conflict resolution. We do not present results of +CR in A-A and Q-As settings because they are equivalent to the results of -CR.

with the kernel size 3 for every CNN layer and we set the LSTM hidden size to 32. We apply a pooling size of 2 to all max pooling layers. Besides, the activation function of the output layers under A-Only and Q-A settings is $sigmoid$, that of output layers under A-A and Q-As settings is $softmax$, and that of all other layers is $tanh$.

In addition, noting that the two answers for the same question have opposite labels in both tasks, we evaluate all systems in terms of (Q, A) pair predication accuracy with and without **conflict resolution** (CR), with which the model resolves conflicts when either two relevant/adopted answers or two irrelevant/unadopted answers are predicted. Because the activation function of the output layers under A-A and Q-As settings is $softmax$ and because there are always two answers for each question, systems under these two settings never generate conflict predictions. We do not apply MAP (Mean Average Precision) (Lai et al., 2018) to the tasks because the number of candidate answers of each question in the datasets is limited to 2.

4.2 Experimental Results

Table 13 shows the experimental results of running the five predictors on the testing set under four different settings. There are a few observations.

First, for the relevancy task, by designing only half of the (Q, A) pairs in *ChiMed-QA2* come from the same QA records. When Q is not given as part of the input (System 1-4), it is impossible for the predictors to determine whether an answer is relevant; therefore, the system performances are no better than random guesses. In contrast, for the adoption prediction task, by designing all the (Q, A) pairs in *ChiMed-QA1* come from the same QA records, and according to Table 9 we also know that about 98% of the answers, regardless of whether they are adopted or not, are relevant. Therefore, the absence of Qs in System 1-4 does not affect system performance a lot.

Second, when both Q and A are present (System 5-9), the accuracy of relevancy prediction is higher than that of adoption prediction, because the former is an easier task (at least for humans). The only exception is ARC-I (System 7), whose results on relevancy is close to random guess (50.34% and 50.60%) while the result on adoption is comparable with other systems. This is due to the way that ARC-I matches questions and answers. Because embeddings of a question and an answer are directly concatenated in ARC-I, Q-A similarity are not fully captured, leading to low performance on relevancy. On the contrary, the adoption prediction does not rely much on the Q-A similarity (as explained above).

Third, for the relevancy task, systems that capture more features of Q-A similarity tend to have a better result. For example, under the Q-A setting, DUET (System 8) outperforms CNN, LSTM and ARC-I (System 5-7) because DUET has an additional model of exact phrase matching between questions and answers. DRMM (System 9) performs better than DUET (System 8) because DRMM uses word embedding instead of exact phrase when matching pairs of phrases between a question and an answer. In contrast, the performances of the five systems on the adoption task are very similar.

In addition, except for the relevancy task evaluated with CR, the contrast between System 10-14 vs. System 5-9 indicates comparing two As always helps predictors in both tasks because intuitively knowing both answers would help us to decide which one is relevant/adopted. On the contrary, the comparison between the same two groups of systems with CR in the relevancy task indicates comparing two As may hurt the relevancy predictors (System 5, 7, 8) because the relevancy is really between Q and A, which might be affected by the existence of other As.

Finally, all the systems under A-Only and Q-A settings (Systems 1-2 and 5-9) benefit from CR. It

is also worth noting that running the models under Q-A setting and to evaluate them without CR in previous studies (Lai et al., 2018) is much more common. Under this setting, the highest performance achieved is 93.60% (System 9). The score is not as high as our expectation and there still exist room for improvement.

4.3 Error Analysis for Relevancy Prediction

We go though errors of system 9 in the relevancy prediction task without CR and find three main types of errors. Note that we artificially build *ChiMed-QA2* for the relevancy prediction task by keeping the adopted answer a of a question q and replacing the unadopted answer of q with an adopted answer a' from another question q'. And we therefore regard a as a relevant answer of q and a' as an irrelevant answer of q (See Section 3.4).

The first type of error is that the answer a is actually irrelevant to the question q. In other words, the gold standard is wrong; system 9 does make a correct prediction. This is not surprising as there are around 2% irrelevant answers in the dataset according to our annotation (See Table 9).

Second, the system fails to capture the relationship between a disease and a corresponding treatment. E.g., a patient describes his/her symptoms and asks for treatment. The doctor offers a drug directly without analyzing the symptoms and causes of disease. In that case, the overlap between the question and the answer is relatively low. The system therefore cannot predict the answer to be relevant without the help of a knowledge base.

Finally, it is quite common that a patient describes his/her symptoms at the beginning of the question q and asks something else at the end (e.g. whether drug X will help with his/her illness). In this case, if q' (the original question of the irrelevant answer a') describes similar symptoms, the system may fail to capture what exactly q wants to ask and therefore mistakes a' for a relevant answer. Table 14 gives an error in this type where q and q' describe similar diseases but they are in fact expecting totally different answers.

Given the three types of errors, we find out the latter two are relatively challenging. This therefore requires further exploration on the way of modeling (Q, A) pairs in the relevancy prediction task. In addition, because current irrelevant answers are randomly sampled from the entire dataset, the current dataset does not include many challenging examples. This makes relevancy prediction task appear easier than what it could be. For future work, we plan to balance the easy and hard instances in the dataset by adding more challenging examples to *ChiMed-QA2*.

q	我上周感冒咳嗽，现在感冒好了，但咳嗽更加厉害了。蜂蜜可以治疗咳嗽吗? I had a cold and cough last week. Now, the cold has gone, but the cough is even worse. Can honey treat cough?
q'	我是支气管扩张患者，最近感冒病情加重。支气管扩张病人感冒怎么治疗? I am a patient with bronchiectasis. I have recently become worse with a cold. How to treat a cold for a bronchiectasis patient?
a'	正常的情况下，支气管病人如果感冒，就应该立即到医院就医，并在医生的指导下用药物治疗。如果耽误治疗的话病情会加重，而且会出现一些并发症。 Normally, if a bronchial patient has a cold, he should go to the hospital immediately and take medication under the guidance of a doctor. If the treatment is delayed, the condition will worsen and complications will occur.

Table 14: An example where system 9 mistakes irrelevant answer a' for a relevant answer. Both questions q and q' are talking about cold and cough, but they are totally different because q is asking whether honey is helpful for cough while q' is looking for treatment.

5 Conclusion and Future Work

In this paper, we present *ChiMed*, a Chinese medical QA corpus collected from an online medical forum. Our annotation on a small fraction of the corpus shows that the corpus is of high quality as approximately 98% of the answers successfully address the questions raised by the forum users. To demonstrate the usage of the corpus, we extract two datasets and use them for two prediction tasks. A few benchmark systems yield good performance on both tasks.

For the future work, we are collecting data to expand the corpus and plan to add more challenging samples to the datasets. In addition, we plan to use *ChiMed* for other NLP tasks such as automatic answer generation, keyphrase generation, summarization, and question classification. We also plan to explore various methods of adding more annotations (e.g., answer ranking) to the corpus.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64. ACM.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Tuan Manh Lai, Trung Bui, and Sheng Li. 2018. A review on deep learning techniques applied to answer selection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2132–2144.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1291–1299. International World Wide Web Conferences Steering Committee.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 269–281.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

Sifatullah Siddiqi and Aditi Sharan. 2015. Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications*, 109(2).

Yan Song, Prescott Klassen, Fei Xia, and Chunyu Kit. 2012. Entropy-based training data selection for domain adaptation. In *Proceedings of COLING 2012: Posters*, pages 1191–1200, Mumbai, India.

Yan Song, Chia-Jung Lee, and Fei Xia. 2017. Learning word representations with regularization from prior knowledge. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 143–152, Vancouver, Canada.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*, pages 175–180, New Orleans, Louisiana.

Yan Song and Fei Xia. 2012. Using a goodness measurement for domain adaptation: A case study on Chinese word segmentation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3853–3860, Istanbul, Turkey.

Yan Song and Fei Xia. 2013. A common case of jekyll and hyde: The synergistic effect of using divided source training data for feature augmentation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 623–631, Nagoya, Japan.

Simon Šuster and Walter Daelemans. 2018. Clicr: a dataset of clinical case reports for machine reading comprehension. *arXiv preprint arXiv:1803.09720*.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Nan Wang, Yan Song, and Fei Xia. 2018a. Coding structures and actions with the COSTA scheme in medical conversations. In *Proceedings of the BioNLP 2018 workshop*, pages 76–86, Melbourne, Australia.

Nan Wang, Yan Song, and Fei Xia. 2018b. Constructing a Chinese medical conversation corpus annotated with conversational structures and actions. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan.

Liu Yang, Qingyao Ai, Jiafeng Guo, and W Bruce Croft. 2016. anmm: Ranking short answer texts with attention-based neural matching model. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 287–296. ACM.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.