Ways to be a Good Consultant: Answer Ranking in Medical Domain

Yuanhe Tian

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2019

Committee:

Fei Xia

Yan Song

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

Ways to be a Good Consultant: Answer Ranking in Medical Domain

Yuanhe Tian

Chair of the Supervisory Committee:
Fei Xia
Department of Linguistics

In this study, we make a move to answer ranking task of medical community question answering (QA). The task of answer ranking has four different settings based on whether features from questions or other answers are used. We designed multiple approaches under each setting to explore how different features contribute to high answer quality. Experimental results on a Chinese Medical QA Dataset show although question-answer relevance is important, cross-answer features are more crucial to distinguish good answers from bad answers. Therefore, in order to become good consultants, it is first recommended that doctors pay their attention to write high quality answers. Finally, our case study demonstrates that good answers tend to show their concern to patient's feelings and to provide more tips for daily care.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

A-ONLY: Abbreviation of one of the four experimental settings where only features of an answer are used.

A-A: Abbreviation of one of the four experimental settings where features of both answers of a question are used.

MAP: Mean Average Precision, a broadly used evaluation metric for the task of answer ranking.

MRR: Mean Reciprocal Rank, a broadly used evaluation metric for the tesk of answer ranking.

QA: Question Answering.

Q-A: Abbreviation of one of the four experimental settings where features of a question and one of its answers are used.

Q-A-A: Abbreviation of one of the four experimental settings where features of a question and both of its answers are used.

Q-A-A COMB: Abbreviation of a category of approaches that directly concatenate embeddings of question and answers under the setting of Q-A-A.

Q-A-A SIM: Abbreviation of a category of approaches under the setting of Q-A-A. These approaches compare similarities of QA pairs before scoring answers' quality.

Q-A-A SIM + A-A DIFF: Abbreviation of a category of approaches under the setting of Q-A-A. These approaches are combinations of approaches in Q-A-A Sim and approaches under A-A settings.

ACKNOWLEDGMENTS

I would like to express appreciation to University of Washington and the Department of Linguistics, where I obtain the opportunity to study the task of answer ranking in the medical domain. I also thank them for offering technical support that enables me to search for related studies and conduct experiments.

I am grateful to my master thesis advisers Fei Xia and reader Yan Song who offer me patient and detailed guidance during the whole research. I also appreciate our team member Weicheng Ma who collects the Chinese Medical QA Dataset and gives me suggestions. My thanks must go also to other research team members and friends: Chen Li, Sicong Huang, Zhaofeng Wu, and Weifeng Jin who contribute to human annotation and provide their ideas to this research.

DEDICATION

This thesis is dedicated to my mother Yuhong Liu. She encouraged me to keep working when I got stuck in the research and wanted to give up.

Chapter 1

INTRODUCTION

With the explosion of information, it is essential to seek and locate the most helpful information in many real-world applications, such as search engine, customer service, personal assistant, etc. It will be a challenge to distinguish good answers from bad answers in QA forums (e.g. Quora[1]). One intuitive way to address this challenge is to rank all answer candidates and retrieve the top of them. For example, in Quora, answers are ranked based on the number of upvotes/downvotes given by forum users. Therefore, answer ranking is a crucial task where a list of answers are ranked with respect to their quality and relevance to a given question [28, 10, 5, 8]. Past studies mainly focus on general domain [33, 7, 32, 36, 38, 23, 20, 25, 3, 30]. Few studies have been conducted in the medical domain, where language usage, expertise requirements, and even problem settings are different.

As many online medical s ervice platforms have emerged (e.g. MedHelp[2] and 39ask[3]), a huge amount of knowledge on diagnoses and analyses is produced with respect to various questions raised by patients. As a result, we face a critical challenge of how to appropriately rank the answers. The challenge might get even harder if it comes to Chinese, because of the problems of word segmentation, medical name entity recognition, and the lack of existing Chinese medical resources (e.g. knowledge base). We find that approaches may not result in the best performance in the medical domain if they directly model question-answer relevance. As shown in Figure 1.1, because both answers are very similar and good in terms of the relevance with the question, these approaches may fail to distinguish the good answer from the bad answer. However, the good an-

[1]https://www.quora.com

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

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

Question:
经常感觉肚子下面突然特别疼痛，这是怎么回事?
I often feel a sudden pain under my stomach. What is going on?

Good Answer:
可能是消化不良导致，建议患者在平时清淡饮食，多喝水，不吃辛辣的食物。
May be caused by indigestion, it is recommended that the patient should have a light diet, drink plenty of water, do not eat spicy food in daily life.

Bad Answer:
可能是急性肠胃炎，一般选择治疗肠胃炎的药物治疗，要注意好休息。
It might be acute gastroenteritis. In general, you can choose drugs that can treat gastroenteritis. Please rest more.

Figure 1.1: An Example of Medical QA. Both answers are relevant to the question, but the good answer offers more informative tips for daily care than the bad one.

swer selected by the questioner must have some advantages (e.g. offers informative tips for daily care). Therefore, our motivation is to find and analyze factors affecting the way of being a good answer.

In order to achieve this goal, we first introduce a Chinese medical QA dataset collected by one of our team members. Questions in the dataset are raised by patients and their answers are written by government licensed doctors. We next study the quality of answers from two aspects: how an answer is related to the given question; how the quality of an answer is compared to other answers. We design approaches under four experimental settings based on different aspects of modeling the question-answer and answer-answer relations. These approaches include several state-of-the-art text similarity calculation models [9, 13, 6]. Based on the experimental results, we find that answer-answer relation is more important in the medical QA circumstance compared with question-answer relevance. In addition, we conduct a small-scope human annotation. Its preliminary results indicate it is not easy for untrained annotators without medical background to identify good answers based on their intuitive judgment. More importantly, the results also emphasize the necessity of further studying human performance in the medical answer ranking task to obtain a more comprehensive

understanding of the current results.

The following chapters are organized as follows. In Chapter 2, we cover related studies in QA and answer ranking in the general domain, as well as studies in medical QA. In Chapter 3, we introduce a Chinese Medical QA Dataset collected by our team member from the Internet. In Chapter 4, we describe our approaches under four different settings that are used to find out how features from question-answer correspondence and answer-answer difference influence answer's quality. Chapter 5 gives experimental settings. We show experimental results and detailed discussion in Chapter 6. In Chapter 7, we present our conclusions and future work.

Chapter 2

LITERATURE REVIEW

Related studies of answer ranking in the medical domain can be located from two aspects: (1) QA and answer ranking in the general domain, and (2) QA in the medical domain.

2.1 QA and Answer Ranking in the General Domain

The area of QA aims at automatically answering questions posed by humans in natural language. Typical studies in this field focus on retrieving information from articles and documents on the Internet and generating summary-like answers [28, 10, 5, 8]. In these studies, answer ranking is not regarded as a core task, although it always serves as a part of QA systems. However, answer ranking is crucial in the area of community QA. Because questions in community QA (e.g. Quora) are directly answered by other users, it will be a good solution for QA to retrieve the top ranked/re-ranked answers of a question. Therefore, answer ranking or answer selection appears to be significantly important in community QA, and it attracts much attention from researchers.

Most existing answer ranking research address the ranking task by modeling Q-A similarity. They measured the relevance between a question and its answers [11] and achieved good performance in popular general domain answer ranking datasets: TrecQA [33] and WikiQA [35]. Based on the approaches they are using, existing studies can be classified into two categories: non-neural approaches and neural approaches.

Typical non-neural approaches use manually selected features to rank answers. Wang et al. [33] proposed a syntax-driven approach to the task of answer selection. Their approach derived from the idea that questions and their correct answers related to each other via predictable syntactic transformations. They modeled the syntactic transformation by extending a probabilistic Quasi-Synchronous Grammar that was originally developed by Smith and Eisner [26]. Wang et al.

improved the grammar by using extra lexical semantics knowledge from WordNet and by applying conditional maximum likelihood estimation to training. Their experimental results on TrecQA dataset collected by themselves show their approach significantly outperforms strong state-of-the-art baselines.

In addition to the study of Wang et al. [33] that concentrated on syntax tree transformations, research of Heilman and Smith [7], Wang and Manning [32], and Yao et al. [36] modeled Q-A similarity with respect to the syntax tree edit distance.

Heilman and Smith [7] defined nine kinds of edit operations (INSERT-CHILD, INSERT-PARENT, DELETE-LEAF, DELETE-&-MERGE, RELABEL-NODE, RELABEL-EDGE, MOVE-SUBTREE, NEW-ROOT, MOVE-SIBLING) on dependency parse trees. They used a tree kernel as a heuristic in a greedy search approach to extract sequences of tree edit between two sentences. After that, a logistic regression approach was applied to build a sentence pair classifier. The experimental results show that their tree edit distance approach outperforms Wang et al. [33] without using lexical semantics knowledge on TrecQA dataset [33].

Wang and Manning [32] addressed the ranking task in a more probabilistic way. They used nearly 30 edit operations from three different edit types: Surface Edits, Semantic Edits, and Syntactic Edits. Different from the study of Heilman and Smith [7] that did not use lexical semantics knowledge, Wang and Manning took advantages of various linguistics resources, including WordNet and NomBank. Afterward, they used Finite-State Machine whose state transitions stand for edit operations to find edit sequences. In addition, they applied Conditional Random Field to parameterizing their classifier. Compared with results of Heilman and Smith [7], their approach performs better with respect to the metric of Mean Reciprocal Rank (MRR) but it performs worse with respect to the metric of Mean Average Precision (MAP).

Yao et al. [36] further improved the results of Heilman and Smith [7] and Wang and Manning [32]. They used nine types of edit operations, and each type contains several possible operations. Compared with the study of Heilman and Smith [7], Yao et al. used dynamic-programming solution [40] to find edit sequences instead of greedy search. Besides, they applied extra 15 new syntactic features and other lexical-semantic relations from WordNet.

Because of the rise of deep neural networks and their good performance, fewer and fewer researchers focus on non-neural approaches. To the best of our knowledge, only one non-neural approach [30] can be found in recent years. The researchers modeled Q-A relevance by taking both intra-pair (Q-A) similarities and cross-pair (Q_1-Q_2 and A_1-A_2) similarities into account. Multiple features including bag-of-n-grams overlap, pre-trained word embeddings, and tree-kernel, were applied to build the ranking model based on logistics regression. The model performance on TrecQA [33] and WikiQA [35] datasets outperformed many complex deep neural network based systems.

In addition to non-neural approaches, neural approaches based on CNN, LSTM and attention mechanism are also applied to the task of answer ranking. The learning approaches can be classified into three categories [11]: (1) **point-wise** approaches (e.g. Yu et al. [38]) that transform the ranking task into a binary classification problem by classifying whether an answer $a_{i,j}$ is a correct answer of question q_i in QA pair $(q_i, a_{i,j})$; (2) **pair-wise** approaches (e.g. Rao et al. [20]) that consider one question q, one correct answer a^+ and one incorrect answer a^- at the same time and use triplet loss function (e.g. Equation 2.2); (3) **list-wise** approaches (e.g. Bian et al. [3]) where a question and all its answer candidates are considered, and where a $softmax$ activation is always applied to the output layer. Compared to traditional non-neural approaches, neural approaches do not rely on manually selected features and in general perform better.

Yu et al. [38] applied a point-wise approach in their system. They used encoders based on bag-of-words and bigram to generate sentence representation vectors and measure Q-A similarity by computing their dot product. Their bigram based model outperformed most non-neural models with respect to MAP on TrecQA dataset [33].

Severyn and Moschitti [23] also used point-wise approach. They applied encoders based on CNN and max pooling to both questions and answers. The similarity of the two sentences is calculated by:

$$sim(\mathbf{x_q}, \mathbf{x_a}) = \mathbf{x_q} M \mathbf{x_a} \tag{2.1}$$

where $\mathbf{x_q}$ and $\mathbf{x_a}$ are embeddings of a question and an answer, respectively, and M is the similarity matrix which will be optimized in training. Identical with Yu et al. [38], Severyn and Moschitti treated the ranking task as a problem of binary classification and applied cross-entropy loss function to training. Taking the advantage of using CNN-based encoders, their approach outperforms previous deep neural approach proposed by Yu et al. [38].

Different from studies of Yu et al. [38] and Severyn and Moschitti [23] that used cross-entropy loss function, Rao et al. [20] used pair-wise approach and triplet loss function:

$$triplet\_loss = max(0, 1 - (sim(q, a^+) - sim(q, a^-))) \tag{2.2}$$

where a^+ and a^- refer to a correct answer and an incorrect answer, respectively. This function tries to optimize parameters by maximizing the margin between positive and negative samples. Their system achieved state-of-the-art performance on TrecQA dataset [33].

In addition to the three above approaches based on Siamese architecture, approaches based on Compare-Aggregate architecture [3, 25] also attracts researchers attention. The two unique parts of Compare-Aggregate architecture are Matching Layer and Aggregation Layer [11].

The model proposed by Bian et al. [3] used a typical Compare-Aggregate architecture and a list-wise approach. Before the Matching Layer, a question and its answer's contextual representations of every position are calculated by an attention mechanism. In the Matching Layer, each contextual representation of the question/answer is compared against all contextual representations of the answer/question by element-wise multiplication. The outputs of two sequences of vectors are then fed into Aggregate Layer where a CNN is used to generate comparison results. Their system achieved state-of-the-art performance on TrecQA [33] and WikiQA [35] datasets.

In general, previous studies focus mainly on modeling question-answer relation. Their successes indicate features of Q-A similarity are important in the general domain.

As for the datasets, TrecQA [33] and WikiQA [35] are two broadly used datasets for general domain answer ranking. There are two versions of TrecQA. Both of them have the same training sets where TRAIN contains manually labeled data and TRAIN-ALL consists of data labeled by

Data	# of Questions	# of QA Pairs	Avg. # of Answers per Question	% Correct
TRAIN	94	4,718	50.2	7.4
TRAIN-ALL	1,229	53,417	43.5	12.0
Raw DEV	82	1148	14.0	19.3
Raw Test	100	1,517	15.2	18.7
Clean DEV	65	1,117	17.2	18.4
Clean Test	68	1,442	21.2	17.2

Table 2.1: Statistics of TrecQA dataset [33]. TRAIN contains manually labeled data and TRAIN-ALL consists of data labeled by automatic methods.

Data	# of Questions	# of QA Pairs	Avg. # of Answers per Question	% Correct
TRAIN	2,118	20,360	9.6	5.1
DEV	296	2,733	9.2	5.1
Test	633	6,165	9.7	4.8
Total	3,047	29,258	9.6	5.0

Table 2.2: Statistics of WikiQA dataset [35].

automatic methods. The development and test sets are different (Table 2.1). The average number of answers per question in the two train sets is around 44. This number gets down to around 15 and 18 in the two versions of development and test dataset, respectively. In addition, the rate of correct answers in TrecQA is less than 20%. The other dataset, WikiQA (Table 2.2), contains around 10 answers for each question on average and the rate of correct answers is around 5% which is lower than the counterpart in TrecQA.

2.2 QA in the Medical Domain

Popular automatic QA research in the medical domain focuses more on retrieving scientific documents or generating summary-like answers from the Internet or knowledge database, instead of answer ranking.

Task b in BioASQ challenge [29] is a typical task in the medical domain QA and many successful medical QA systems can be found (e.g. BioAMA [24]). In this task, systems are required to handle four types of biomedical question: yes/no, factoid, list and summary questions [2]. In phase B of the task, participants need to extract/generate answers or summaries in natural language, given questions and questions related articles and snippets.

From the description of the task, we can learn that it focuses more on the area of document information retrieval instead of answer ranking. Although answer ranking may appear to be one of the steps in automatic QA systems, researchers pay less attention to it because it is not the most important part. For example, BioAMA [24], a system achieved state-of-the-art results on the ideal answer type questions in Task 5b[1] of the BioASQ dataset, only mentioned that it used point-wise ranking classifiers without detailed information on the model architecture.

Therefore, both the task description and researchers' focus demonstrate that in medical QA, it is information retrieval that attracts most researchers' attention, instead of answer ranking. In addition, to the best of our knowledge, few studies can be found in the medical domain answer ranking.

Given this context, our study focuses on ranking doctor written answers in medical community QA. Besides, we address answer ranking task in the medical domain by modeling both question-answer and answer-answer relations, and find the later one plays a more important role.

---

[1]The 5th edition of the Task b.

Chapter 3
THE CHINESE MEDICAL QA DATASET

As is introduced in Chapter 2, existing datasets for answer ranking locate in the general domain. Besides, there is no existing datasets for the medical domain community QA, especially for Chinese. We, therefore, introduce a Chinese medical QA dataset in this chapter. The dataset is collected by one of our teammates through crawling 39ask[1]. It is a Chinese medical consultation forum where patients briefly describe their symptoms, and doctors in the community offer their diagnoses. In the following sections, we describe the dataset, introduce its properties, and provide its potential usage.

3.1 Dataset Details

The first version of our dataset contains 35,000 questions, which cover 15 departments and 2992 disease types. Most questions (65.7%) come from internal medicine, and the rest range from andrology (5.6%), surgical department (5.5%), pediatrics (5.5%), infectious diseases (5.0%), gynecology (5.0%), traditional Chinese medicine department (4.2%), plastic surgery (3.4%), and other seven departments (0.2%). In addition, in the 35,000 questions, 9.0% of them have exactly one answer, 87.6% are followed by two answers and 3.3% have three or more answers (Table 3.1). Among the questions that have exactly two answers, 35.0 % have two good answers, and the rest 65.0 % have exactly one good answer and one bad answer. No questions containing two bad answers can be found. The statistics of our dataset is shown in Table 3.2, where 69.2% of the answers are good answers and 30.8% are bad answers.

Figure 3.1 represents a typical instance in our dataset. It is worth noting that English is not a part of the dataset and it is added for the readers' convenience only. In the question section,

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

# of answers per question	1	2	$\geqslant 3$	Total
# of questions	3,160	30,675	1,165	35,000

Table 3.1: Statistics of the number of answers per question in the full Chinese medical QA dataset.

# of questions	35,000
# of answers	68,717
# of good answers	47,534
# of bad answers	21,183
Avg. # of answers per question	1.96
Avg. length of question (in char)	55.61
Avg. length of answer (in char)	107.14
Avg. length of Q + A (in char)	162.92

Table 3.2: Statistics of the Chinese medical QA dataset. Average lengths of questions and answers are in characters.

the department and the disease type are specified by keyword "*department*". Other information of the question is also presented. It includes "*title*" which is a summary of the patient's question, "*patient_info*" that offers basic patient information of gender, age and the time of onsite, "*question_content*" where the patient describes his/her symptoms, "*time*" of the question posted, and "*labels*" that offers a list of keywords with respect to this question. As for the two answers following the question, the name, specialty and other useful information of the answer provider are shown in key words of "*name*", "*specialty*" and "*other_info*", respectively. The diagnosis ("*answer*"), response time ("*time*") and whether it is selected as a good answer by the patient ("*selected*") are also presented.

Question:

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?
patient_info	男，46 岁，发病时间：不清楚 Male, 46 years old, Onset time: not clear
question_content	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?
time	2018-12-15 14:59
labels	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:

name	刘祥礼 (Liu, Xiangli)
specialty	高血压，冠心病，肺心病，心肌炎心肌病等 Hypertension, coronary heart disease, pulmonary heart disease, myocarditis, cardiomyopathy, etc.
other_info	主任医师（Chief physician）
answer	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.
time	2018-12-19 11:21
selected	True

Answer 2:

name	董春林 (Dong, chunlin)
specialty	全科（General medicine)
other_info	日照市中医医院 (Rizhao City Traditional Chinese Medicine Hospital)
answer	这是一种普通的慢性胃粘膜炎症，与幽门螺旋杆菌感染有关。可以选择阿莫西林治疗。 This is a common chronic gastric mucosal inflammation and has a relationship with Helicobacter pylori infection. You can choose amoxicillin for treatment.
time	2018-12-15 15:12
selected	False

Figure 3.1: An Example in the Chinese Medical QA Dataset. English is not a part of the dataset and it is added for the readers' convenience only.

3.2 Properties of the Dataset

Compared with the general domain community QA datasets (TrecQA [33] and WikiQA [35]), our dataset in the medical domain has two properties: (1) much lower average number of answers per question (around 2 answers per question) and (2) much higher rate of good/correct answers (more than a half). These two properties, in fact, reflect the two different contexts of studying QA in the medical domain.

First, the requirement of being an answer provider is much higher in the medical domain. In the general domain, everyone can be a potential answer provider, because it does not require any professional knowledge to answer a question. However, when it comes to the medical domain, answer providers are required to have government-issued doctor licenses and professional knowledge. Therefore, it is uncommon to see one question followed by dozens of answers in medical QA forums. The quality of answers are on average much better than the answers in the general domain as well.

Next, the intent of patients' questions are much clearer than that of questions in the general domain. A question in general domain can be quite open and can be answered from different aspects (e.g. "*how are aircraft radial engines built?*" in WikiQA [35]). However, patients are asking questions about their medical concerns. When doctors narrow the patients' intent down, the answers will be more specific. Disagreement among doctors will also less likely to happen. As a result, we can find out almost all answers in the medical domain are highly related to their questions and more answers are selected as good answers by patients.

3.3 Potential Usage of the Dataset

There are several potential tasks can be applied to this dataset. Including the task of answer ranking, we list four potential directions of research.

- **Question Classification**: identify the department/disease type a question belongs to;
- **Question Summarization**: generate its title given a question;

- **Keys words generation/extraction**: generate/extract labels of a question;
- **Answer Ranking**: identify which answer is selected by the patient.

Chapter 4

OUR APPROACHES

To fulfill our motivation of finding and analyzing factors affecting the way of being a good answer, we enumerate all possible features' combinations of whether using the question or other answers. We, therefore, design four settings for the answer ranking task: (1) **A-Only** where only features of the answer are used; (2) **Q-A** where features of a question and one of its answers are used; (3) **A-A** where features of both answers[1] of a question are used; (4) **Q-A-A** where features of a question and both of its answers are used. For each setting, we design a set of approaches. These approaches include several previous state-of-the-art text similarity calculation approaches (ARC-I [9], DUET [13], and DRMM [6]).

4.1 A-Only Approaches

To find out how an answer's content influences its quality, we design a set of approaches whose only input is the answer itself. These approaches score an answer's quality without knowing the answer's relevance to its question and other answers. We build answer encoders based on both CNN and bi-directional LSTM to identify answers' quality. Before the encoders, we first use a 150-dimensional word embedding to featurize the input answer. Next, in the CNN model, a 1-dimensional CNN layer with 32 filters and a kernel size of 3 is used to extract phrase level features and the size of hidden units in the bi-directional LSTM is set to 32. Then, a max pooling layer with a pooling size of 2 is conducted to the output of CNN/LSTM to select the most important phrases. After that, we reshape the feature matrix to a vector and pass it through a fully connected layer with 64 hidden units. The architecture of encoders is shown in Figure 4.1.

[1]Because 87.6% of the questions in Chinese Medical QA Dataset have exactly two answers, we only use those questions in our experiment.

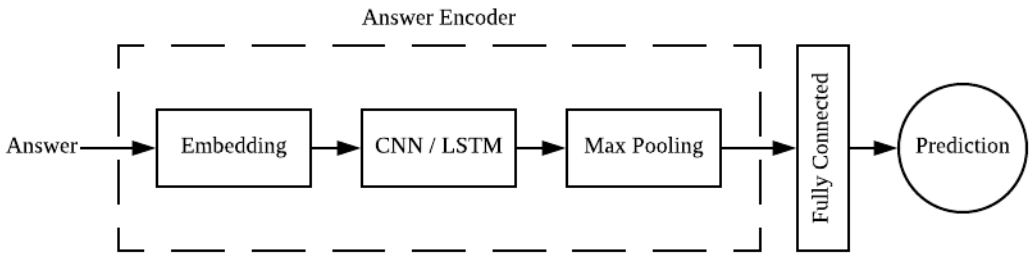

Figure 4.1: Architecture of our approaches under A-Only setting.

4.2 Q-A Approaches

To find out how question-answer relevance influences the answer's quality, we design a set of approaches whose inputs are a question and one of its answer. These approaches focus on modeling the similarity of the question and the answer without knowing other answers.

We apply five approaches to answer ranking task under the setting of Q-A.

1. **CNN** is an approach measures Q-A similarity by making element-wise multiplication of question and answer embeddings generated by CNN-based encoders.

2. **LSTM** uses the same Q-A similarity calculation approach as CNN model (element-wise multiplication) and its architecture is also identical with CNN except for the encoder based on bi-directional LSTM. The architecture of our approaches is shown in Figure 4.2.

3. **ARC-I** [9] captures Q-A similarity by directly concatenating question and answer embedding vectors generated by encoders based on CNN and max pooling. The architecture of ARC-I is identical with CNN-based Q-A Matcher shown in Figure 4.2, except for the merging strategy that is concatenation in ARC-I.

4. **DUET** [13] calculates Q-A similarity by summing up the scores from Local Model (LM) and Distributed Model (DM). LM estimates Q-A relevance based on patterns of exact matches of

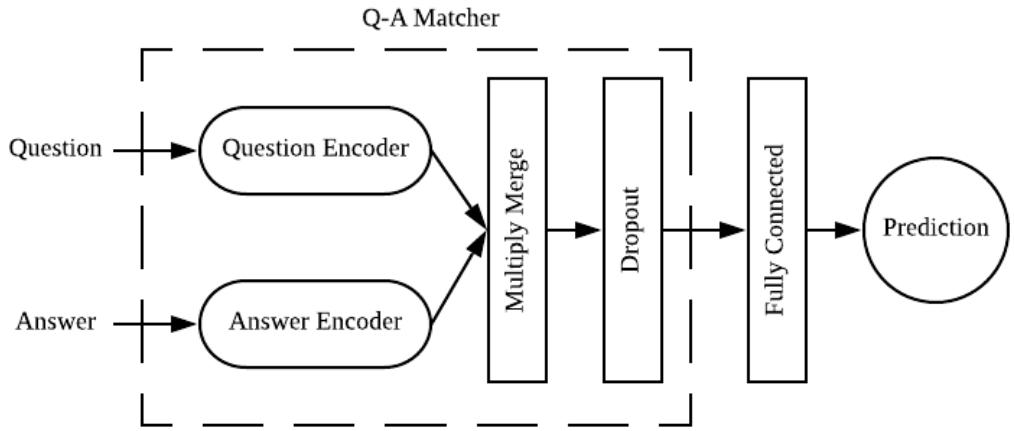

Figure 4.2: Architecture of our approaches under Q-A setting. The architecture of question and answer encoders are identical with the architecture in Figure 4.1.

question terms in the answers. And DM first generates an embedded question feature vector and answer feature matrix by CNN-based encoders, then use Hadamard product to merge the two embeddings. Figure 4.3 shows the architecture of DUET.

5. **DRMM** [6] measures Q-A similarity by conducting dot product of word embeddings between a question and one of its answer. Then, for each word in the question, top k (in our experiment, k equals to 10) similarities are selected. After that, two fully connected layers are conducted to generate high-level similarity feature vectors for every word in the question, and an attention mechanism is applied to the question to help final relevance score calculation. Figure 4.4 shows the architecture of DRMM,

4.3 A-A Approaches

To find out how answer-answer relation influences the answers' quality, we design a set of approaches whose inputs are both answers of a question. These approaches aim at modeling the difference between answers' content that allows an answer to be a good answer, without using

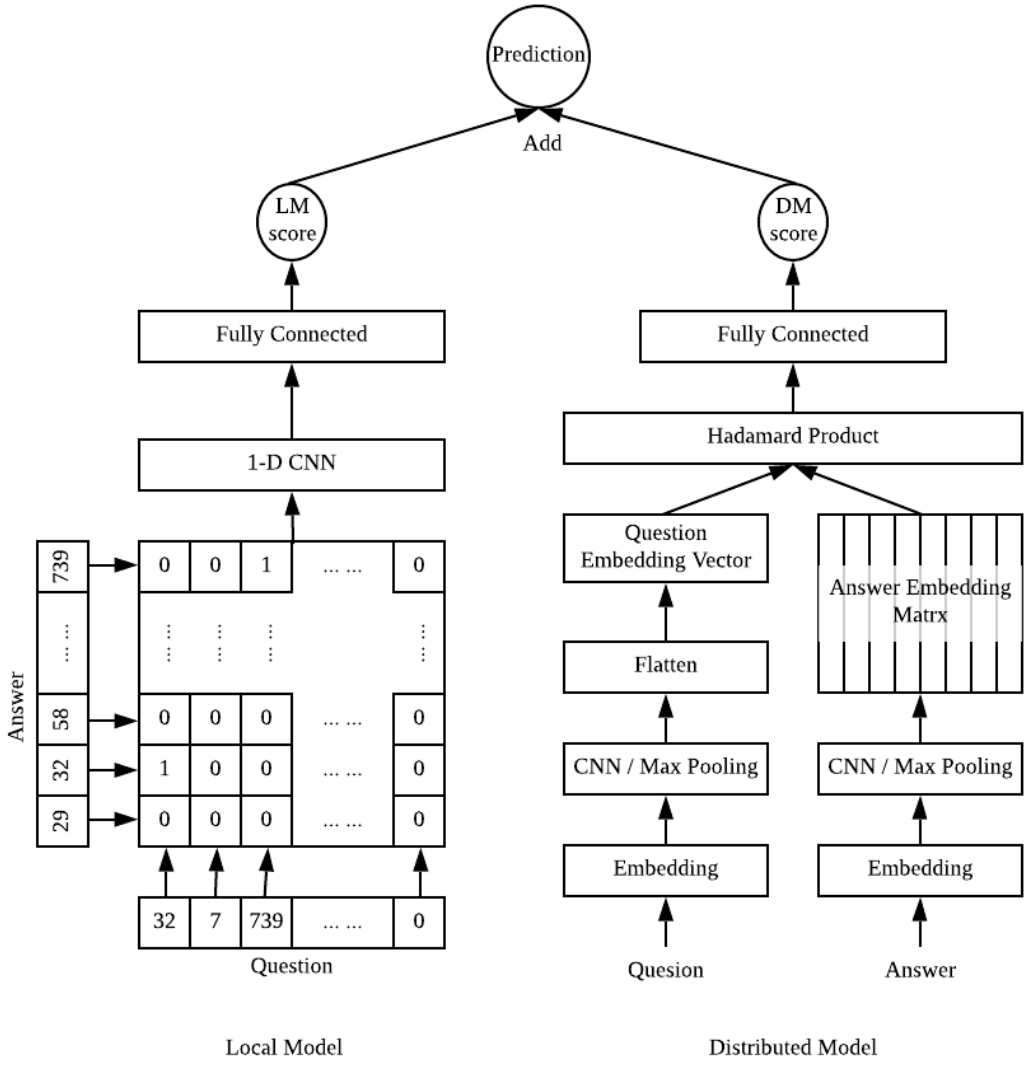

Figure 4.3: Architecture of DUET. In Local Model, the number at each position of questions and answers represents the index of the term in the vocabulary. The value at the position of i, j in the merging matrix will equal to 1 if the term at position i in the answer is identical with the term at position j in the question. The value will equal to 0 elsewhere.

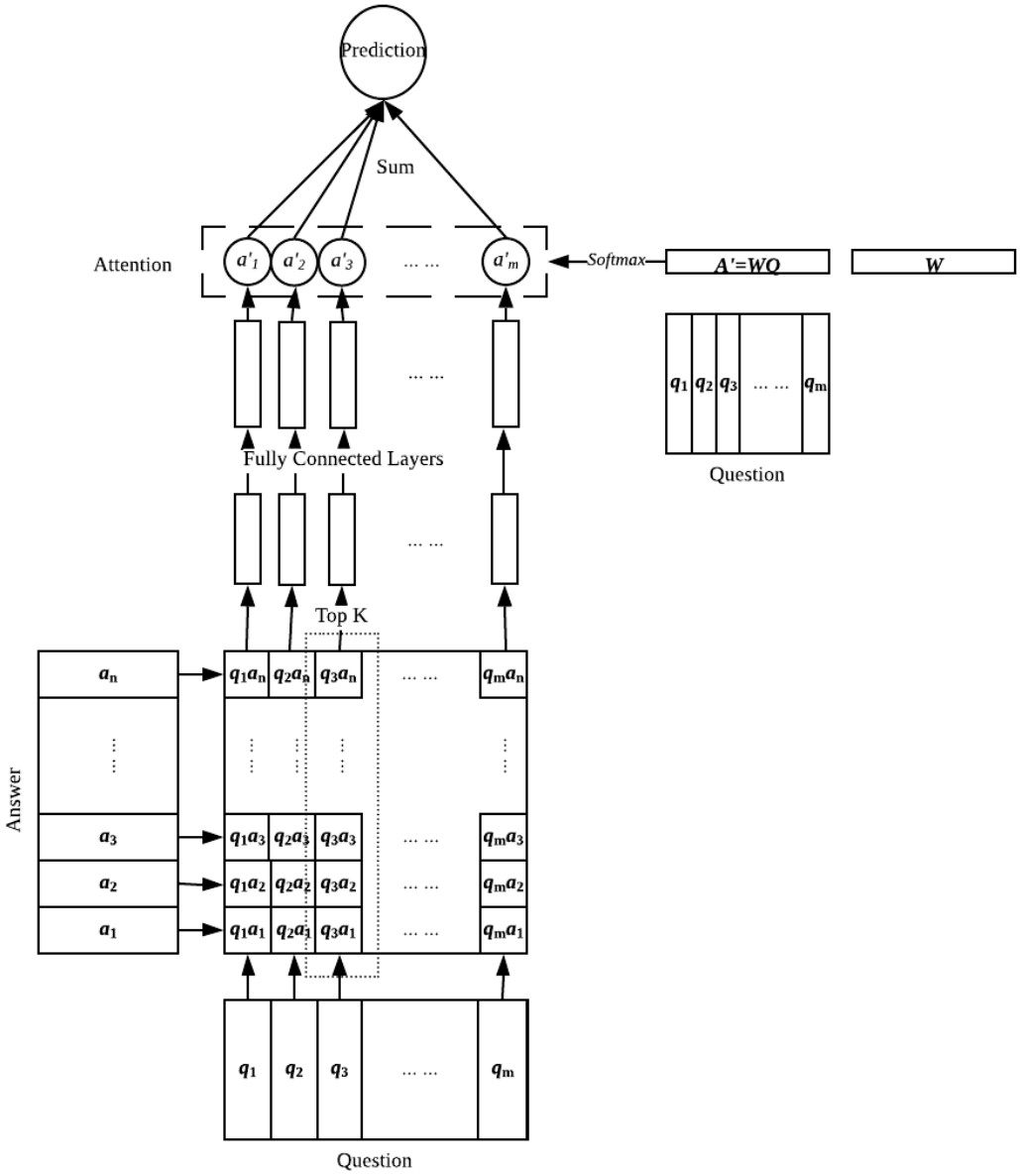

Figure 4.4: Architecture of DRMM. In the similarity matrix, $a_i q_j$ ($i \leqslant n, j \leqslant m$) represents the dot product of the word embedding at position i in an answer and the word embedding at position j in its question. a'_k ($k \leqslant m$) refers to the weight of attention mechanism at position k.

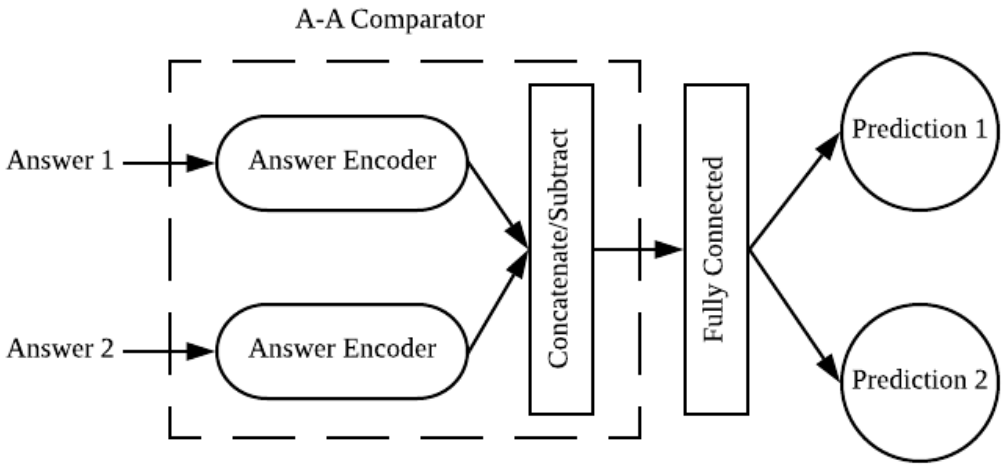

Figure 4.5: Architecture of our approaches under A-A setting. The architecture of answer encoder is identical with the one in Figure 4.1.

features of their question. We use two different ways to model A-A relation. The first is to directly concatenate the two answer embeddings generated by CNN/LSTM encoders. It simply combines features of both answers to medel their content difference. The second is to use the subtraction. It offers a more intuitive way to model answers' difference. After that, the vector presenting A-A difference is passed through a fully connected layer with 64 hidden units and the final scores are calculated without normalization. The architecture of our approaches under A-A setting is shown in Figure 4.5, where the architecture of answer encoder is identical with the one in Figure 4.1.

4.4 Q-A-A Approaches

To find out how they influence an answer's quality if features of a question and both of its answers are taken into account, we design a set of approaches whose inputs are the question and both of its answers. We apply three categories of approaches to this setting: Q-A-A Combination approaches, Q-A-A Similarity approaches and Q-A-A Similarity + A-A Difference approaches.

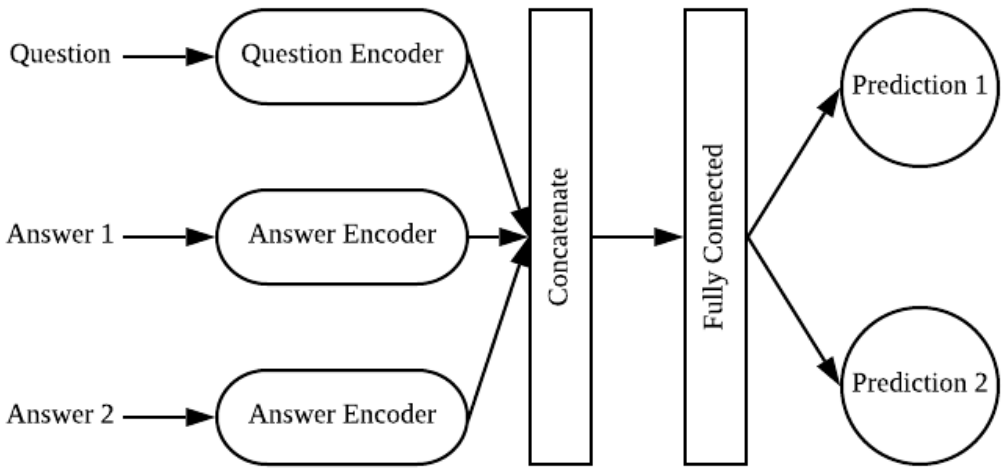

Figure 4.6: Architecture of our Q-A-A Combination approach under Q-A-A setting. The architecture of encoders is identical with the one in Figure 4.1

.

4.4.1 Q-A-A Combination Approaches

Q-A-A Combination (Q-A-A Comb) approaches merge features of a question and both of its answers by directly concatenating their embeddings. They provide a simple way of considering all features from the inputs, without specifically modeling their relations. In this kind of approaches, both CNN- and LSTM-based encoders are used and the architecture is shown in Figure 4.6.

4.4.2 Q-A-A Similarity Approaches

Q-A-A Similarity (Q-A-A Sim) approaches compare the similarities of a question between both of its answers before scoring every answer. They examine whether the comparison between Q-A similarities contributes to answer ranking.

We apply all five approaches introduced in Section 4.2 to the setting of Q-A-A. These approaches are regarded as Q-A Matchers whose inputs are a question and one of its answer, and whose output is a similarity feature vector (CNN, LSTM, and ARC-I) or score (DUET, DRMM).

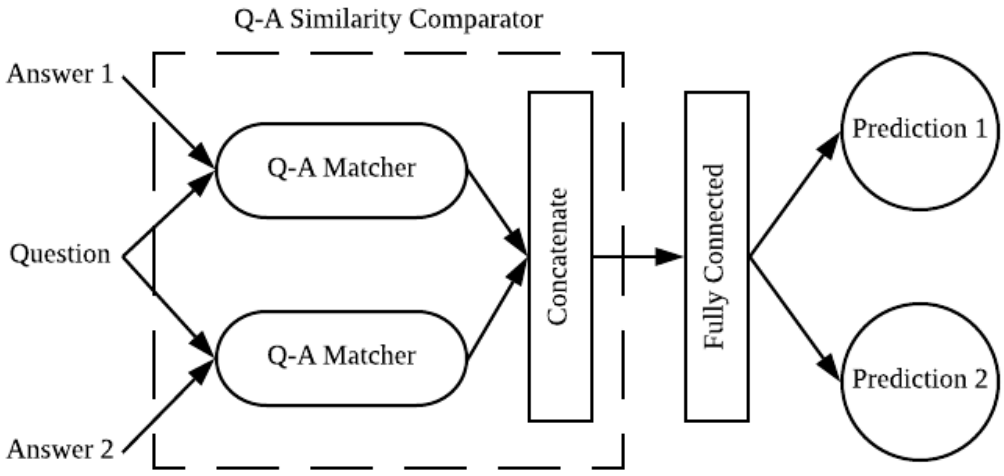

Figure 4.7: Architecture of our Q-A-A Similarity approach under Q-A-A setting. Q-A matcher can stand for any question-answer similarity calculator. We use five Q-A mathcers in our experiment: CNN, LSTM, ARC-I, DUET, and DRMM.

Q-A similarities generated by a Q-A Matcher are directly concatenated and passed through a fully connected layer with 64 hidden units to generate the vector of matching result. The vector is finally fed into the output layer to calculate final scores without normalization (Figure 4.7).

4.4.3 Q-A-A Similarity + A-A Difference Approaches

Q-A-A Similarity + A-A Difference (Q-A-A Sim + A-A Diff) approaches combine Q-A-A Similarity approaches (Section 4.4.2) and A-A approaches[2] (Section 4.3) to rank answers. These approaches concatenate the outputs of a Q-A Similarity Comparator (Figure 4.7) and an A-A Comparator (Figure 4.5) right before scoring the quality of answers (Figure 4.8). They allow us to see whether A-A difference contributes to the ranking task compared with Q-A-A Similarity ap-

[2]We use "A-A Diff" instead of "A-A" to refer to our approaches under A-A setting. So "A-A" refers to one of our experimental setting and "A-A Diff" refers to a kind of approaches applied to A-A setting.

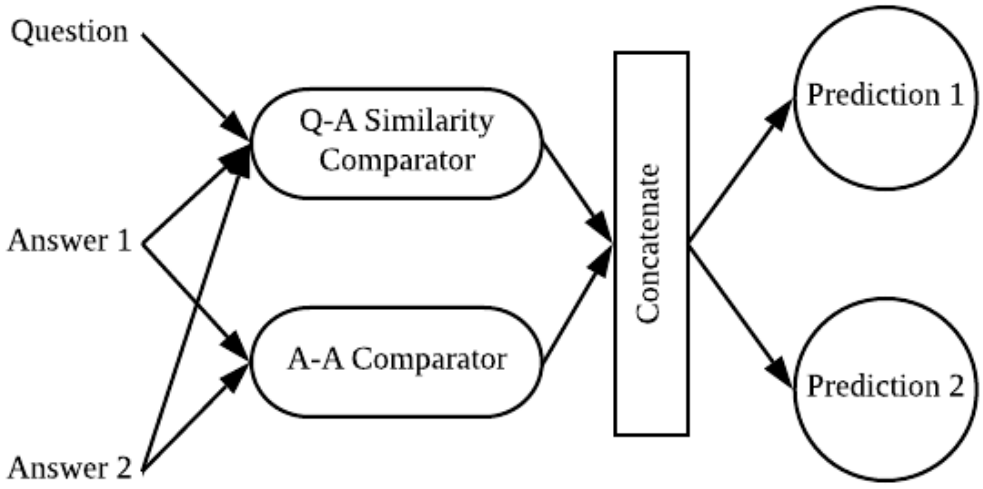

Figure 4.8: Architecture of our Q-As Similarity + A-A Difference approaches under Q-A-A setting. The architecture of Q-A Similarity Comparator is introduced in Figure 4.7 and the architecture of A-A Comparator is shown in Figure 4.5

.

proaches. We apply all five Q-A Matchers introduced in Section 4.2 to the combination. In addition, we only use CNN-based A-A Comparator in the experiment.

Chapter 5

EXPERIMENTS

To evaluate our approaches under four different settings, we first sample a subset of the Chinese Medical QA Dataset for our experiment. Then, we introduce the loss function of training and three different kinds of evaluation metrics applied to the answer ranking task. Finally, we conduct a small-scope human annotation to obtain an intuitive feeling of task difficulty.

5.1 Dataset

The dataset used in our experiments is a subset of the Chinese Medical QA Dataset which is introduced in Chapter 3. Because 87.6% of the questions in the full dataset are followed by two answers, and 65.0% of the two-answer questions have exactly one good and one bad answers, we select all two-answer questions that contain exactly one good and one bad answers as the dataset for the experiment. In addition, we clean the dataset by removing all questions and answers whose character-based lengths are too long or too short[1].

Finally, we sample a dataset with 19,924 questions and 39,848 answers for our experiment (Table 5.2 and one typical instance for experience is shown in Table 5.1. We select 75% (14,943 questions and 29,886 answers) of the dataset as the training set and the rest 25% (4,981 questions and 9,962 answers) for testing. In addition, we apply character-based embeddings, because the word vocabulary in Chinese is too large and word segmentation approaches do not perform well in the medical domain.

[1]We rank the questions and answers by their character-based length and remove those at the top 1% and those at the bottom 1%.

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?
Good Answer	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.
Bad Answer	这是一种普通的慢性胃粘膜炎症，与幽门螺旋杆菌感染有关。可以选择阿莫西林治疗。 This is a common chronic gastric mucosal inflammation and has a relationship with Helicobacter pylori infection. You can choose amoxicillin for treatment.

Table 5.1: An example of QA for our experiment. Compared with the bad answer, the good answer not only soothes the patient but also provides a more informative suggestion by offering tips for daily care.

5.2 Loss Function and Evaluation Metrics

Based on the nature of the binary answer classification of all our approaches under four different settings, we apply binary cross-entropy loss function to training.

In addition, we apply three evaluation metrics to our experiment:

1. **MRR** (Mean Reciprocal Rank) and **MAP** (Mean Average Precision) are two broadly used metrics in the task of answer ranking [11]. In our experiment, we assume all approaches only

# of questions	19,924
# of answers	39,848
# of answers per question	2
Avg. length of question (in char)	62.99
Avg. length of answer (in char)	118.93
Avg. length of Q + A (in char)	181.92

Table 5.2: Statistics of the experiment data from Chinese medical QA dataset.

retrieve the answer with a higher score. Because only one good answer for each question exists in our dataset, the two metrics are identical [11].

2. **QA Set Accuracy** is a reference metric that measures if systems correctly label two answers simultaneously by scoring more than 0.5 to the good answer while less than 0.5 to the bad answer. Obviously, this metric is more strict than MRR and MAP.

3. **QA Pair Accuracy** is another reference metric that evaluates models' performance on every single QA pair. An answer with a score of more than 0.5 is regarded as a good answer and is regarded as a bad answer if its score is lower than 0.5.

In our experiment, the first metric measures whether a system give a good answer a higher score regardless of whether the two answers are both classified as good or bad. The other two metrics, in addition to whether good answers have a higher score, evaluate whether the system correctly classifies good and bad answers by a threshold of 0.5.

5.3 Human Annotation

We conduct a small-scope human annotation by asking five annotators without medical background to identify good answers of 20 randomly selected questions under four different settings (80 in

total): (1) they can only see one answer; (2) they can see one QA pair; (3) they can see both two answers; (4) they can see one question and both two answers.

The annotators are untrained, are required to give an intuitive judgment without deep analysis, and are told that there is only one good answer in the answer list when they are given more than one answer. Samples in the human annotation questionnaire can be found in Appendix A.

Chapter 6

RESULTS AND DISCUSSION

We first analyze the results with respect to how features from different aspects contribute to answer's quality. Next, we discuss the results of human annotation. Finally, we conduct a case study to find out which factors affect answer's quality.

6.1 Model Results and Analysis

Table 6.1 and Table 6.2 present the experimental results of all approaches under different settings and metrics (Section 5.2). Approaches in Table 6.1 are under settings of: (1) A-Only where only features of the answer are used, (2) Q-A where features of a question and one of its answers are used, and (3) A-A where features of both answers of a question are used; approaches in Table 6.2 are under the setting of (4) Q-A-A where features of a question and both of its answers are used. In addition, because MRR and MAP are identical given only one correct answer in gold standard [11], we only present MAP metric in these tables. Based on the experimental results, we present our observations of how features of question-answer similarity and answer-answer difference influence answer's quality.

6.1.1 Q-A vs. A-Only

Cross-setting comparison between approaches under two settings of Q-A and A-Only in Table 6.1 tells us features of answer content are important to answer's quality. Only using features form answer content, approaches (A-[1,2]) can achieve comparable results to approaches under Q-A setting (QA-[1-5]), where features of Q-A similarity are taken into account. It indicates features of the answer itself may contribute to a good quality.

Setting	Expt. id	Approach	MAP	QA Set Acc	QA pair Acc
A-Only	A-1	CNN	81.23	51.56	73.50
	A-2	LSTM	81.81	51.27	73.90
Q-A	QA-1	CNN	81.29	52.36	74.32
	QA-2	LSTM	81.53	53.10	74.01
	QA-3	ARC-I	80.35	53.56	74.13
	QA-4	DUET	81.35	51.86	74.40
	QA-5	DRMM	82.07	49.83	71.21
A-A	AA-c1	CNN (conc)	83.15	82.69	83.12
	AA-c2	LSTM (conc)	83.12	82.83	83.03
	AA-s1	CNN (sub)	**83.22**	82.75	**83.27**
	AA-s2	LSTM (sub)	83.18	**83.08**	83.16

Table 6.1: Experimental results of all approaches under settings of A-Only, Q-A and A-A. "conc" and "sub" refer to "concatenate" and "subtract", the two ways of modeling A-A difference. MAP evaluates whether the only answer retrieved by systems with a higher score is a good answer; QA Set Acc requires systems to correctly label two answers simultaneously by scoring more than 0.5 to the good one while less than 0.5 to the bad one; QA pair Acc evaluates every single QA pair separately. Because MRR and MAP are identical given only one correct answer in gold standard [11], we only present MAP metric in this table. A model's results under different metrics may using parameters from different training epoch.

6.1.2 Q-A: Different Q-A Matchers

When comparing the performance of different approaches under Q-A setting (QA-[1-5]), we find that the more features from answer content are reserved, the higher performance approaches tend to achieve in QA Set Accuracy (the second metric column). When ranked by this metric, the results of CNN-based models from the best to the worst are ARC-I (QA-3, 53.56%) $>$ CNN (QA-1, 52.36%) $>$ DUET (QA-4, 51.86%). The same order appears when we rank those models from the most to the least with respect to how many features of answer content are reserved. ARC-I computes Q-A similarity by directly concatenate vectors of the question embedding and the answer

Set of Approaches	Expt. id	Approach	MAP	QA Set Acc	QA pair Acc
Q-A-A Comb	QAC-1	CNN	82.41	81.99	82.44
	QAC-2	LSTM	83.16	82.94	83.17
Q-A-A Sim	QAS-1	CNN	82.37	82.11	82.34
	QAS-2	LSTM	81.63	81.33	81.63
	QAS-3	ARC-I	82.01	81.75	82.03
	QAS-4	DUET	81.13	81.11	81.12
	QAS-5	DRMM	81.57	81.45	81.58
Q-A-A Sim + A-A Diff	QAS+AA-c1	CNN+conc	82.77	82.31	82.72
	QAS+AA-c2	LSTM+conc	83.02	82.35	82.94
	QAS+AA-c3	ARC-I+conc	82.71	82.23	82.73
	QAS+AA-c4	DUET+conc	82.81	82.41	82.78
	QAS+AA-c5	DRMM+conc	82.41	81.95	82.38
	QAS+AA-s1	CNN+sub	82.43	81.81	82.39
	QAS+AA-s2	LSTM+sub	81.99	81.65	82.04
	QAS+AA-s3	ARC-I+sub	82.96	82.59	82.97
	QAS+AA-s4	DUET+sub	**83.42**	**82.98**	**83.41**
	QAS+AA-s5	DRMM+sub	82.55	82.11	82.60

Table 6.2: Experimental results of all approaches under the setting of Q-A-A where all features of a question and its two answers are used. MAP evaluates whether the only answer retrieved by systems with a higher score is a good answer; QA Set Acc requires systems to correctly label two answers simultaneously by scoring more than 0.5 to the good one while less than 0.5 to the bad one; QA pair Acc evaluates every single QA pair separately. "Sim", "Diff" and "Comb" are abbreviation of "Similarity", "Difference" and "Combination". In approaches of Q-A-A Sim + A-A Diff, we use CNN-based encoder to calculate A-A difference and "conc" and "sub" mean the ways to calculate A-A difference are "concatenate" and "subtract", respectively. Because MRR and MAP are identical given only one correct answer in gold standard [11], we only present MAP metric in this table. A model's results under different metrics may use parameters from different training epoch.

embedding. It, therefore, reserves most features from the answer. CNN multiplies the two embeddings so that positions that both the question and the answer emphasize will have a higher value, and positions that they do not emphasize will have a lower value. It, therefore, reserves relatively fewer features from the answer. Local Model of DUET is responsible for catching exactly word level Q-A correspondence and its Distributed Model merges the question and the answer before the answer embedding matures (Figure 4.3). DUET, therefore, reserves the least features from the answer content. In addition, this analysis can be strengthened by the performance of DRMM (QA-5). Because it only takes word-embedding-based Q-A term match into account (Figure 4.4), it rarely reserves features from answer content and has the lowest QA set accuracy (49.83%) among approaches under Q-A setting (QA-[1-5]).

6.1.3 A-A vs. A-Only

Compared with approaches under A-Only setting (A-[1,2]), all metrics are improved when features of another answer are taken into account (AA-[c1, c2, s1, s2]). It demonstrates the difference between answers helps identify answers' quality without referring to their question. Besides, no obvious difference appears between different ways (concatenation or subtraction) of comparing A-A difference (AA-[c1, c2, s1, s2]).

6.1.4 Q-A-A Comb vs. A-A

In this subsection, we compare approaches of Q-A-A Combination (QAC-[1,2]) and A-A approaches that use concatenation to model A-A difference (AA-c[1,2]). As is introduced in Section 4.4.1 and Section 4.3, the only difference between the two kinds of approaches is whether to concatenate question embeddings. When features of questions are taken into account, no significant improvement is acquired (e.g. AA-c2 vs. QAC-2, MAP: 83.12 $\rightarrow$ 83.16) and the performance even gets worse (e.g. AA-c1 vs. QAC-1, MAP: 83.15 $\rightarrow$ 82.41). This phenomenon, therefore, shows the unstable contribution of features from questions.

6.1.5 Q-A-A Sim vs. Q-A

Compared with Q-A approaches (QA-[1-5]) in Table 6.1, Q-A-A Sim approaches (QAS-[1-5]) in Table 6.2 achieve around 30% and 6% absolute improvement in metrics of QA Set Accuracy and QA pair Accuracy, and can reach a comparable result with respect to MAP. The improvements in the two metrics reveal that by comparing similarities from different answers, systems can not only identify which answer is better (MAP) but also classify the answer's quality by a threshold of 0.5 (QA Set Accuracy and QA Pair Accuracy).

However, with respect to MAP, the contribution of Q-A similarity comparison is not that clear: the performance of CNN (QA-1 vs. QAS-1), LSTM (QA-2 vs. QAS-2) and ARC-I (QA-3 vs. QAS-3) are slightly improved while the performance of DUET (QA-4 vs. QAS-4) and DRMM (QA-5 vs. QAS-5) become slightly worse.

Therefore, the contribution of Q-A similarity comparison is controversial. On the one hand, it will help much if we want to build systems that can classify good and bad answers by a threshold of 0.5; on the other hand, its contribution seems to be unstable if we want to build systems that just retrieve the answer with the highest score.

6.1.6 Q-A-A Sim + A-A Diff vs. Q-A-A Sim

Comparing results of Q-As Sim approaches with results of Q-As Sim + A-A Diff approaches, we find features of A-A difference appear to have a stable and positive influence on the task of medical answer ranking. All Q-A-A Sim approaches (QAS-[1-5]) are improved when features of A-A difference are taken into account (QAS+AA-c[1-5], QAS+AA-s[1-5]), which is presented Table 6.2. It may indicate that A-A difference contains features not only crucial to medical answer ranking but also hard to be captured by just modeling Q-A similarity.

6.1.7 Q-A-A Sim + A-A Diff vs. A-A

One interesting phenomenon appears in this comparison is that almost all results in Q-As Sim + A-A Diff (QAS+AA-c[1-5], QAS+AA-s[1-3,5]) are lower than the lowest result under A-A setting

(AA-c2, MAP: 83.12) with respect to MAP. The only exception is DUET (QAS+AA-s4, MAP: 83.42) that achieves the best results among all results in Table 6.1 and Table 6.2. This phenomenon shows different combinations of approaches modeling Q-A similarity and A-A difference may lead to different results. For most combinations in our experiment, features of Q-A similarity could hurt approaches only modeling A-A difference. However, an appropriate combination (QAS+AA-s4) may strengthen the performance of both the approach for Q-A similarity (DUET) and the approach for A-A difference (CNN-conc).

To conclude the above discussions, we find that Section 6.1.1, Section 6.1.2, Section 6.1.3 and Section 6.1.6 demonstrate the important influence of answer content and A-A difference to the task of medical answer ranking; Section 6.1.4 and Section 6.1.5 present a not obvious contribution of question and Q-A similarity to answers' quality. These discussions demonstrate features of A-A difference may play more important roles than features of Q-A similarity in the medical domain answer ranking. In addition, discussion in Section 6.1.7 further demonstrate the performance of A-A comparators may get better if they are combined with appropriate Q-A matchers (e.g. CNN (sub) + DUET in QAS+AA-s4).

As a result, in order to become a good consultant, the doctor is first recommended to write answers with a high quality. In addition, he/she can try to make the answer be more related to the question.

6.2 Human Annotation Results and Discussion

Table 6.3 shows the results and annotator agreement of our small-scope human annotation. We find the average accuracy of human annotation (around 60%) under all settings are significantly lower than model performance (around 80%). Besides, when we compare results under the settings of Q+A, As and Q+As, human performance and agreement do not improve even though more information is provided.

The big decreases of all three agreements when two answers are given (A Only and Q+A vs. As and Q+As) may indicate the difficulty for untrained annotators to distinguish the good answer from the bad answer based on their intuitive judgment. However, we also need to note the probability

Annotator / Settings	I	II	III	IV	V	**Avg. Acc**	**$\geqslant$ 4 Agr**	**5 Agr**	**Avg. pairwise Agr**
A Only	45	45	55	55	50	50	0.90	0.25	0.69
Q + A	75	55	55	55	60	60	0.85	0.50	0.77
As	65	55	65	70	55	62	0.55	0.35	0.60
Q + As	55	50	55	65	70	59	0.50	0.25	0.60

Table 6.3: Accuracy and agreement of five human annotators under different settings. "Acc" and "Agr" are abbreviations of accuracy and agreement, respectively. "$\geqslant$ 4 Agr" refers to the agreement of at least four annotators give the same annotation; "5 Agr" refers to the agreement of all five annotators give the same annotation; "Avg. pairwise Agr" refers to the average agreement of every two annotators.

"A Only", "Q+A", "As", and "Q+As" refer to the settings that annotators are given only one answer, one question and its answer, both two answers, and one question and its two answers, respectively.

that the results fail to present the real human performance because of the noise coming from the small sample size. Therefore, to obtain an estimated upper bound for automatic systems in the medical domain answer ranking, as well as to better explain these unexpected phenomena, further studies of human annotation are required (e.g. train the annotators before they start).

6.3 Case Study

Discussion in Section 6.1.4 and Section 6.1.5 indicates an unstable contribution of features of Q-A similarity to the task of medical answer ranking. We, therefore, conduct case studies and find Q-A similarity contributes less because it is less likely for doctors to provide unrelated solutions. Almost all answers are highly related to their questions and people can even infer the question when its answer is given. Therefore, the contribution of Q-A relation in the medical domain is weakened. What is more, answers that repeat patients' symptoms in the question are not what the questioners want. In this case, a high relevance of Q-A actually hurts the answer's quality. Table 6.4 shows an example where the bad answer has a high word overlap (highlighted in blue) with the

Question	遇到阴雨天气时，就会胸闷、头晕、混身无力。以前有脑供血不足、心脏神经官能症、血压高。怎么治疗，应该注意什么? When it is rainy, I will feel chest tightness, dizziness, and weakness. In the past, I suffered from Cerebral insufficiency, cardiac neurosis, and high blood pressure. How to treat and what should I pay attention to?
Good Answer	根据你的描述，这种情况结合症状主要还是跟你这些基础病是有关系的，目前来说就是平时吃药控制为好。可以吃降压药，定期检测血压情况，平时还可以吃点补品。 According to your description, this combination of symptoms is mainly related to your underlying diseases. At present, it is usually better to take medication control. You can take antihypertensive drugs, check blood pressure regularly, and eat some supplements.
Bad Answer	阴雨天气气压低，氧气不足，会造成人体缺氧的症状，表现为，头晕乏力胸闷。建议注意休息，控制血压，可以口服改善脑供血的药物，也可以每天低流量吸氧一个小时。 When it is rainy, the air pressure is low, and the oxygen is insufficient, which will cause symptoms of hypoxia in the human body. It is manifested as dizziness, fatigue and chest tightness. It is recommended to rest, control blood pressure, take drugs can improve brain blood supply, or you can breathe oxygen for one hour every day.

Table 6.4: An example in Chinese medical QA dataset. Because the overlaps between the bad answer and the question concentrate on the symptoms in which patients are not interested, the answer is not selected as a good answer by the patient.

question. Because those overlaps concentrate on the symptoms in which patients are not interested, the answer is not selected as a good answer by the patient.

We further find that the importance of A-A difference comes from two main characteristics of good answers. Good answers tend to show their concern to patient's feelings and to provide more

tips for daily care instead of simply asking them to go to a hospital. Table 5.1 represents a good example. The good answer captures the worry of the patient and soothes him/her by telling him/her the disease is not as much severe as he/she thinks. In addition, the good answer offers several tips that can be easily conducted in daily life (have a regular diet, chew slowly, do not eat spicy food, etc.). On the contrary, the bad answer says nothing about what the patient needs to pay attention to in daily life to relieve the symptoms. The importance of offering suggestions for daily care may result from the nature of online medical consulting. Because patients are aware of that online medical consulting forums are not substitutes for hospitals, where clear diagnoses can be made with the help of careful medical examinations, suggestions for daily care or immediately doable actions will be one of the most important information that patients expect. Therefore, answers providing medical daily care exactly meet the patients' expectation, so that they result in good quality.

Additional examples and their analyses can be found in Appendix B.

Chapter 7

CONCLUSIONS AND FUTURE WORK

In this study, we apply answer ranking techniques to the medical domain and explore how features from question-answer relevance and answer-answer difference contribute to the ranking task. Experimental results demonstrate a higher value of features of answer-answer difference compared with features of question-answer relevance. Therefore, in order to be good consultants, it is first recommended that doctors pay their attention to write high quality answers. Our case study demonstrates that it will be helpful to an answer's quality if the answer shows concerns to patients feelings and provides more tips for daily care. In addition, preliminary small-scope human annotation results show the potential difficulty for untrained people to identify the good answer based on their intuitive judgment. The results also indicate the necessity of conducting further studies on human annotation.

Future studies can be conducted in the following three directions:

1. In addition to directly concatenating and subtracting, we will apply more complex approaches to model features of answer-answer difference. For these approaches, we will also try to find appropriate approaches of modeling Q-A similarity.

2. We will further study the performance of human annotation to find out the real human performance towards the task of medical domain answer ranking and find a convincing explanation of current results. We will enlarge the size of human annotation questionnaire and train annotators before they start.

3. We will study the factors contribute to good answers via a more statistical way. In the current study, the useful factors (e.g. good answers tend to provide more tips for daily care.) are supported by our case study, which is based on a few samples. To provide a more

comprehensive analysis of these factors, it is better to conduct a more statistical method (e.g. apply attention mechanism to current method and use the weights to indicate the importance of different terms in the answer).

BIBLIOGRAPHY

[1] Asma Ben Abacha and Pierre Zweigenbaum. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information processing & management*, 51(5):570–594, 2015.

[2] Georgios Balikas, Ioannis Partalas, Aris Kosmopoulos, Sergios Petridis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Nicolas Baskiotis, Eric Gaussier, Thierry Artieres, et al. Evaluation framework specifications. *Project deliverable D*, 4, 2013.

[3] Weijie Bian, Si Li, Zhao Yang, Guang Chen, and Zhiqing Lin. A compare-aggregate model with dynamic-clip attention for answer selection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1987–1990. ACM, 2017.

[4] Georgios-Ioannis Brokos, Polyvios Liosis, Ryan McDonald, Dimitris Pappas, and Ion Androutsopoulos. Aueb at bioasq 6: Document and snippet retrieval. *arXiv preprint arXiv:1809.06366*, 2018.

[5] Nagehan Pala Er and Ilyas Cicekli. A factoid question answering system using answer pattern matching. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 854–858, 2013.

[6] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 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, 2016.

[7] Michael Heilman and Noah A Smith. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Human Language Technologies: The 2010 Annual*

Conference of the North American Chapter of the Association for Computational Linguistics, pages 1011–1019. Association for Computational Linguistics, 2010.

[8] Phu Mon Htut, Samuel R Bowman, and Kyunghyun Cho. Training a ranking function for open-domain question answering. *arXiv preprint arXiv:1804.04264*, 2018.

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

[10] Jeongwoo Ko, Teruko Mitamura, and Eric Nyberg. Language-independent probabilistic answer ranking for question answering. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 784–791, 2007.

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

[12] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[13] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 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, 2017.

[14] Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 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, 2015.

[15] Henry Nassif, Mitra Mohtarami, and James Glass. Learning semantic relatedness in community question answering using neural models. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 137–147, 2016.

[16] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, Georgios Paliouras, and Ioannis Kakadiaris. Results of the fifth edition of the bioasq challenge. *BioNLP 2017*, pages 48–57, 2017.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[18] Shengwen Peng, Ronghui You, Hongning Wang, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. Deepmesh: deep semantic representation for improving large-scale mesh indexing. *Bioinformatics*, 32(12):i70–i79, 2016.

[19] John Prager et al. Open-domain question–answering. *Foundations and Trends® in Information Retrieval*, 1(2):91–231, 2007.

[20] Jinfeng Rao, Hua He, and Jimmy Lin. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1913–1916. ACM, 2016.

[21] Fabio Rinaldi, James Dowdall, Gerold Schneider, and Andreas Persidis. Answering questions in the genomics domain. In *Proceedings of the Conference on Question Answering in Restricted Domains*, 2004.

[22] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[23] Aliaksei Severyn and Alessandro Moschitti. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 373–382. ACM, 2015.

[24] Vasu Sharma, Nitish Kulkarni, Srividya Pranavi, Gabriel Bayomi, Eric Nyberg, and Teruko Mitamura. Bioama: Towards an end to end biomedical question answering system. In *Proceedings of the BioNLP 2018 workshop*, pages 109–117, 2018.

[25] Gehui Shen, Yunlun Yang, and Zhi-Hong Deng. Inter-weighted alignment network for sentence pair modeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1179–1189, 2017.

[26] David A Smith and Jason Eisner. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 23–30. Association for Computational Linguistics, 2006.

[27] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers on large online qa collections. *Proceedings of ACL-08: HLT*, pages 719–727, 2008.

[28] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 41–47. ACM, 2003.

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

[30] Kateryna Tymoshenko and Alessandro Moschitti. Cross-pair text representations for answer sentence selection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2162–2173, 2018.

[31] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. A deep architecture for semantic matching with multiple positional sentence representations. In *AAAI*, volume 16, pages 2835–2841, 2016.

[32] Mengqiu Wang and Christopher D Manning. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1164–1172. Association for Computational Linguistics, 2010.

[33] Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 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)*, 2007.

[34] Wei Wu, SUN Xu, and WANG Houfeng. Question condensing networks for answer selection in community question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1746–1755, 2018.

[35] Yi Yang, Wen-tau Yih, and Christopher Meek. 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, 2015.

[36] Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. Answer extraction as sequence tagging with tree edit distance. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 858–867, 2013.

[37] Wenpeng Yin, Dan Roth, and Hinrich Schütze. End-task oriented textual entailment via deep explorations of inter-sentence interactions. In *Proceedings of the 56th Annual Meeting of*

the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 540–545, 2018.

[38] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*, 2014.

[39] Kaizhong Zhang and Dennis Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262, 1989.

[40] Sheng Zhang, Jiajun Cheng, Hui Wang, Xin Zhang, Pei Li, and Zhaoyun Ding. Furongwang at semeval-2017 task 3: Deep neural networks for selecting relevant answers in community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 320–325, 2017.

Appendix A

SAMPLES IN HUMAN ANNOTATION QUESTIONNAIRE

This chapter offers four samples in the human annotation questionnaire. Table A.1, A.2, A.3, and A.4 present samples under settings of A-Only. Q+A, As, and Q+As, respectively.

Answer	这是一种普通的慢性胃粘膜炎症，与幽门螺旋杆菌感染有关。可以选择阿莫西林治疗。 This is a common chronic gastric mucosal inflammation and has a relationship with Helicobacter pylori infection. You can choose amoxicillins for treatment.
Annotation	Is this a good answer (y/n)?

Table A.1: An example in human annotation questionnaire (A-Only setting).

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?
Answer	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.
Annotation	Is this a good answer (y/n)?

Table A.2: An example in human annotation questionnaire (Q+A setting).

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.
Answer 2	这是一种普通的慢性胃粘膜炎症，与幽门螺旋杆菌感染有关。可以选择阿莫西林治疗。 This is a common chronic gastric mucosal inflammation and has a relationship with Helicobacter pylori infection. You can choose amoxicillin for treatment.
Annotation	Which answer is better (1/2)?

Table A.3: An example in human annotation questionnaire (As setting).

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?
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.
Answer 2	这是一种普通的慢性胃粘膜炎症，与幽门螺旋杆菌感染有关。可以选择阿莫西林治疗。 This is a common chronic gastric mucosal inflammation and has a relationship with Helicobacter pylori infection. You can choose amoxicillin for treatment.
Annotation	Which answer is better (1/2)?

Table A.4: An example in human annotation questionnaire (Q+As setting).

Appendix B

EXTRA EXAMPLES IN OUR CHINESE MEDICAL QA DATASET

In this chapter, we offer three extra examples that present the characteristics of good answers. We also give a brief analysis for each example. Table B.1 shows an example whose good answer show its concern to the patient's feelings; Table B.2 and B.3 offer two examples whose good answers give suggestions with informative tips of daily care.

Question	我家小孩打吊瓶，拔针之后发现手肿了，请问该怎么办? My child received infusion therapy. After the winged infusion set is removed, the hand is swollen. What should I do?
Good Answer	一般来说这种情况过一段时间就会恢复的，不要太过于焦虑。平时尽量多吃一些补充维生素c的食物。然后可以多吃一些苹果西瓜之类的，日常生活当中尽量做好杀菌消毒的工作。 Generally this symptom will recover after a while, do not be too anxious. Usually, eat as much food that contains vitamins as possible. Then you can eat apple and watermelon etc. Try to do the sterilization work in daily life.
Bad Answer	挂瓶手肿了是有一定原因的，很有可能是药液注入皮下了，尤其是这种小孩子。症状一般可自行消失，可用毛巾热敷。慢慢会缓解的。 There is a reason for the swollen after infusion therapy. It is very likely that the liquid is injected into the skin, especially for such children. Symptoms generally disappear on their own. You can be applied with a hot towel. It will be slowly relieved.

Table B.1: An example in Chinese medical QA dataset. The good answer soothes the patient by saying “do not be too anxious”, which shows its concern about the patient.

Question	本人今年63岁，最近一段时间不知为什么总是觉得胸口疼痛，想问下是怎么回事? I am 63 years old. I don't know why I always feel chest pain in the recent period. What is going on?
Good Answer	根据患者的问题来分析，可能是由于冠心病所导致，建议患者尽早到医院进行检查治疗，平常不要盲目用药，以免一些不良反应的发生，平时不要剧烈运动，饮食上要以清淡为主，不要吃辛辣食物。 Analyzing based on the patient's problem, it might due to coronary heart disease. It is recommended that patients go to the hospital as soon as possible to conduct examination and treatment. In daily life, do not blindly use drugs, so as to avoid some adverse reactions. In daily life, do not exercise vigorously, the main diet should be light, don't eat spicy food.
Bad Answer	据患者的问题来分析，胸口痛首先需注意明确有无心脏疾病，可到医院进一步行心电，心脏彩超，动态心脏彩超等检查明确，同时注意有无胃部疾病可能，必要时可进一步体检胃镜检查明确。 Analyzing based on the patient's problem, chest pain first needs to pay attention to and clarify whether there is heart disease. Can make it clear by going to the hospital for further ECG, cardiac color Doppler ultrasound, dynamic heart color Doppler and other checks. At the same time, paying attention to whether it is stomach disease. If necessary, make it clear by further Physical examination and gastroscopy.

Table B.2: An example in Chinese medical QA dataset. Although both answers recommend the patient to go to a hospital to have a clear diagnosis, the good answer, give informative tips for daily care (highlighted in blue), which meets the expectation of the patients.

Question	最近感觉头痛，听说可以使用阿司匹林维生素c泡腾片，请问这个药物有用吗? Recently I felt a headache. I heard that I can use Aspirin and Vitamin C Effervescent Tablets. So is this drug useful?
Good Answer	这种情况可以使用这种药物。它对缓解头痛有一定的帮助。平常要注意个人卫生，不要吃辛辣的食物，注意休息，多喝水，多吃些蔬菜水果。 In this case, you can use this drug. It helps to relieve headaches. Usually pay attention to personal hygiene, do not eat spicy food, rest more, drink plenty of water, eat more fruits and vegetables.
Bad Answer	阿司匹林维生素C泡腾片可以用于普通感冒或流行性感冒引起的发热，也用于缓解轻至中度疼痛如头痛、关节痛、牙痛、肌肉痛、神经痛、痛经等，效果不错的。 Aspirin and Vitamin C Effervescent Tablets can be used for fewer caused by the cold or influenza, and also used to relieve mild to moderate pain such as headache, joint pain, toothache, muscle pain, neuralgia, dysmenorrhea, etc., and the effect is good.

Table B.3: An example in Chinese medical QA dataset. The good answer not only directly responds to the patient's question but also offers tips for daily care.

VITA

Yuanhe Tian is a graduate student in the program of Master of Science in Computational Linguistics at the University of Washington.

He welcomes your comments to yhtian@uw.edu.