Towards Medical Machine Reading Comprehension with Structural Knowledge and Plain Text

Dongfang Li[1], **Baotian Hu**[1], **Qingcai Chen**[1,2,*], **Weihua Peng**[3,*], **Anqi Wang**[1]
[1]Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Baidu, International Technology (Shenzhen) Co., Ltd.
crazyofapple@gmail.com, {hubaotian, qingcai.chen}@hit.edu.cn
pengweihua@baidu.com, 19s051040@stu.hit.edu.cn

Abstract

Machine reading comprehension (MRC) has achieved significant progress on the open domain in recent years, mainly due to large-scale pre-trained language models. However, it performs much worse in specific domains such as the medical field due to the lack of extensive training data and professional structural knowledge neglect. As an effort, we first collect a large scale medical multi-choice question dataset (more than 21k instances) for the National Licensed Pharmacist Examination in China. It is a challenging medical examination with a passing rate of less than 14.2% in 2018. Then we propose a novel reading comprehension model KMQA, which can fully exploit the structural medical knowledge (i.e., medical knowledge graph) and the reference medical plain text (i.e., text snippets retrieved from reference books). The experimental results indicate that the KMQA outperforms existing competitive models with a large margin and passes the exam with 61.8% accuracy rate on the test set.

1 Introduction

With the advent of large scale datasets such as SQuAD (Rajpurkar et al., 2016, 2018), RACE (Lai et al., 2017), and Natural Questions (Kwiatkowski et al., 2019; Lee et al., 2019) on the open domain, machine reading comprehension (MRC) has become a hot topic in the natural language processing field. In the past few years, the MRC has obtained substantial progress, and many recent models have surpassed the human performance on several datasets. The superiority of these models is mainly attributed to two significant aspects: 1) the powerful representations ability of large pre-trained language models (PLMs), which can cover or remember most of the language variations *implicitly*.

Question: 患者，女，27岁，确诊**慢性乙型肝炎**3年，近日化验结果：HBV-DNA 2×10^5 copies/mL, ALT 122 U/L。拟予以**抗病毒治疗**，首选的药物是哪个? A female patient, aged 27 years old, has been diagnosed with ***chronic hepatitis B*** for 3 years. Recent results show: HBV-DNA 2×10^5 copies/mL, ALT 122 U/L. The initial diagnosis is to take ***antiviral treatment*** for her. Which is the preferred one among the following drugs?
Options: **A**. 阿糖腺苷 Ara adenosine. **B**. 恩替卡韦 Entecavir. ✓ **C**. 泛昔洛韦 Famciclovir. **D**. 利巴韦林 Ribavirin. **E**. 膦甲酸钠 Sodium foscarnet.
Option B retrieved text snippets: 临床用于抗乙型肝炎病毒的药物有拉米夫定, 阿德福韦, 干扰素-α, 利巴韦林, 恩替卡韦等... Drugs used clinically against hepatitis B virus include lamivudine, adefovir, interferon-α, ribavirin, entecavir, ...
Option B knowledge facts: (恩替卡韦, 适应症, 慢性乙型肝炎) (entecavir, indication, chronic hepatitis B) (恩替卡韦, 二级分类, 抗病毒药) (entecavir, second class, antiviral drugs)

Table 1: An example from our multiple-choice QA task in a medical exam (✓: correct answer option).

For example, among the top 10 works on SQuAD 2.0, nine models are based on ALBERT (Lan et al., 2020).[1] 2) the most popular MRC datasets belong to the open domain, which are built from news, fiction, and Wikipedia text, etc. The answers to most questions can be derived from the given plain text directly.

Compared to the open domain MRC, medical MRC is more challenging, while owning the great potential of benefiting clinical decision support. There still lacks the popular benchmark medical MRC dataset. Some recent works are trying to construct medical MRC dataset such as PubMedQA (Jin et al., 2019), emrQA (Pampari et al., 2018) and HEAD-QA (Vilares and Gómez-Rodríguez, 2019), etc. However, either these data sets are noisy (e.g., due to semi-automatically or heuristic rules generated), or the annotated data

* Co-corresponding authors

[1]At the time of submission (June 3, 2020). The leaderboard is at https://rajpurkar.github.io/SQuAD-explorer

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1427–1438,
November 16–20, 2020.

scale is too small (Yoon et al., 2019; Yue et al., 2020). Instead, we constructs a large scale medical MRC dataset by collecting 21.7k multiple-choice problems with human-annotated answers for the National Licensed Pharmacist Examination in China. This entrance exam is a challenging task for humans, which is used to assess human candidates' professional medical knowledge and skills. According to the statistics data, the examinee's pass rate in 2018 is less than 14.2%. [2] The text of the reference books is used as the plain text for the questions. One example is illustrated in Table 1.

Though several pre-trained language models have been introduced for domain-specific MRC, BERT based models are not as consistently dominant as they are in open field MRC tasks (Zhong et al., 2020; Yue et al., 2020). Another challenge is that medical questions are often more difficult; no labeled paragraph contains the answer to a given question. Searching for multiple relevant snippets from possibly large-scale text such as the whole reference books is usually required. In many cases, the answer can not be found explicitly from the relevant snippets, and the medical background knowledge is needed to derive the correct answers from the relevant snippets. Therefore, unlike open domain, just using the powerful pre-trained language model and plain text cannot obtain the high performance for medical MRC. For example, in Table 1, the relevant snippets (the 3rd row) can only induce that *Ribavirin* and *Entecavir* are the possible answers for the given question (the 1st row). If the triples from medical knowledge graph (*entecavir*, *indication*, *chronic hepatitis B*) is used, we can quickly obtain the correct answer as *Entecavir*.

Here, we propose a novel medical MRC model KMQA, which exploits the reference medical text and external medical knowledge. Firstly, KMQA models the representations of interaction between question, option, and retrieved snippets from reference books with the co-attention mechanism. Secondly, the novel proposed knowledge acquisition algorithm is performed on the medical knowledge graph to obtain the triples strongly related to questions and options. Finally, the fused representations of knowledge and question are injected into the prediction layer to determine the answer. Besides, KMQA acquires factual knowledge via learning from an intermediate relation classification task and enhances entity representation by constructing a sub-graph using question-to-options paths. Experiments show that our unified framework yields substantial improvements in this task. Further ablation study and case studies demonstrate the effectiveness of the injected knowledge. We also provide an online homepage at http://112.74.48.115:8157.

2 Related Work

Medical Question Answering The medical domain poses a challenge to existing approaches since the questions may be more challenging to answer. BioASQ (Tsatsaronis et al., 2012, 2015) is one of the most significant community efforts made for advancing biomedical question answering (QA) systems. SeaReader (Zhang et al., 2018) is proposed to answer questions in clinical medicine using documents extracted from publications in the medical domain. Yue et al. (2020) conduct a thorough analysis of the emrQA dataset (Pampari et al., 2018) and explore the ability of QA systems to utilize clinical domain knowledge and to generalize to unseen questions. Jin et al. (2019) introduce PubMedQA where questions are derived based on article titles and can be answered with its respective abstracts. Recently, pre-trained models have been introduced to medical domain (Lee et al., 2020; Beltagy et al., 2019; Huang et al., 2019a). They are trained on unannotated biomedical texts such as PubMed abstracts and have been proven useful in biomedical question answering. In this paper, we focus on multiple choice problems in medical exams that are more difficult and diverse, which allows us to directly explore the capabilities of QA models to encode domain knowledge.

Knowledge Enhanced Methods KagNet (Lin et al., 2019) represents external knowledge as a graph, and then uses graph convolution and LSTM for inference. Ma et al. (2019) adopt the BERT-based option comparison network (OCN) for answer prediction, and propose an attention mechanism to perform knowledge integration using relevant triples. Lv et al. (2020) propose a GNN-based inference model on conceptual network relationships and heterogeneous graphs of Wikipedia sentences. BERT-MK (He et al., 2019) integrates fact triples in the KG, while REALM (Guu et al., 2020) augments language model pre-training algorithms with a learned textual knowledge retriever. Unlike previous works, we incorporate external knowledge

[2] http://www.cqlp.org/info/link.aspx?id=3599&page=1

implicitly and explicitly. Built upon pre-trained models, our work combines the strengths of both text and medical knowledge representations.

3 Method

The medical MRC task in this paper is a multiple-choice problem with five answer candidates. It can be formalized as follows: given the question Q and answer candidates $\{O_i\}$, the goal is to select the most plausible correct answer $\hat{O}$ from the candidates. KMQA utilizes textual evidence spans and incorporates **K**nowledge graphs facts for **M**edical multi-choice **Q**uestion **A**nswering. As shown in Figure 1, it consists of several modules: (a) the multi-level co-attention reader that computes context-aware representations for the question, options and retrieved snippets, and enables rich interactions among their representations. (b) the knowledge acquisition which extracts knowledge facts from KG given the question and options. (c) the injection layer that further incorporates knowledge facts into the reader, and (d) a prediction layer that outputs the final answer. And also, we utilize the relational structures of question-to-options paths to further augment the performance of KMQA.

3.1 Multi-level Co-attention Reader

Given an instance, text retrieval system is firstly used to select evidence spans for each question-answer pair. We take the concatenation of question and candidate answer as input, and keep top-N relevant passages. These passages are combined as new evidence spans. Here, we use BM25-based search indexer (Robertson and Zaragoza, 2009) and medical books as text source.

Multi-level co-attention reader is used to represent the evidence spans E, the question Q and the option O. We formulate the input evidence spans as $E \in \mathbb{R}^m$, the question as $Q \in \mathbb{R}^n$ and a candidate answer as $O \in \mathbb{R}^l$, where m, n and l is the max length of the evidence spans, question and candidate answer respectively. Similar to (Devlin et al., 2019), given the input E, Q and O, we apply the WordPiece tokenizer and concatenate all tokens as a new sequence ([CLS],E,[SEP],Q,#,O,[SEP]), where "[CLS]" is a special token used for classification and "[SEP]" is a delimiter. Each token is initialized with a vector by summing the corresponding token, segment and position embedding, and then encoded into a hidden state by the BERT based pre-trained language model.

Generally, the PLMs are pre-trained on the large scale open domain plain text, which lacks the knowledge of the medical domain. There are some recent works show that to further pre-train PLMs on the intermediate tasks can significantly improve the performance of target task (Wang et al., 2019; Clark et al., 2019; Pruksachatkun et al., 2020). Following this observation, we incorporate knowledge from the Chinese Medical Knowledge Graph (CMeKG) (Byambasuren et al., 2019)[3] by intermediate-task training. The CMeKG is a Chinese knowledge graph in medical domain developed by human-in-the-loop approaches based on large-scale medical text data using natural language processing and text mining technology. Currently, it contains 11,076 diseases, 18,471 drugs, 14,794 symptoms, 3,546 structured knowledge descriptions of diagnostic and therapeutic technologies, and 1,566,494 examples of medical concept links, along with attributes describing medical knowledge. The triple in CMeKG consists of four parts: head entity, tail entity and relation along with an attribute description. To acquire factual knowledge, we adopt the relation classification task to further pre-train PLMs on this dataset. This task requires a model to classify the relational labels of a given entity pair based on context. Specifically, we select a subset from CMeKG with 163 distinctive relations and include only the triples in which the relation related to drugs and disease types in the exam. Then, we discard all the relations with fewer than 5,000 entity pairs and retain 40 relations and 1,179,780 facts. After that, we concatenate two entities and insert "[SEP]" between the two as input, and then apply a linear layer to "[CLS]" vector of the last hidden feature of PLM to perform relation classification. Next, we discard the classification layer and initialize the corresponding part of the PLM with other parameters, denoted as $\mathcal{B}$. Finally, we employ $\mathcal{B}$ to get encoding representation $\mathbf{H}_{cls} \in \mathbb{R}^h$, $\mathbf{H}_E \in \mathbb{R}^{m \times h}$, $\mathbf{H}_Q \in \mathbb{R}^{n \times h}$, $\mathbf{H}_O \in \mathbb{R}^{l \times h}$, $\mathbf{H}_{QE} \in \mathbb{R}^{(n+m) \times h}$ respectively, where h is the hidden size.

To strengthen the information fusion from the question to the evidence spans as well as from the evidence spans to the question, we adopt a multi-level co-attention mechanism, which has been shown effective in previous models (Xiong

[3] http://cmekg.pcl.ac.cn

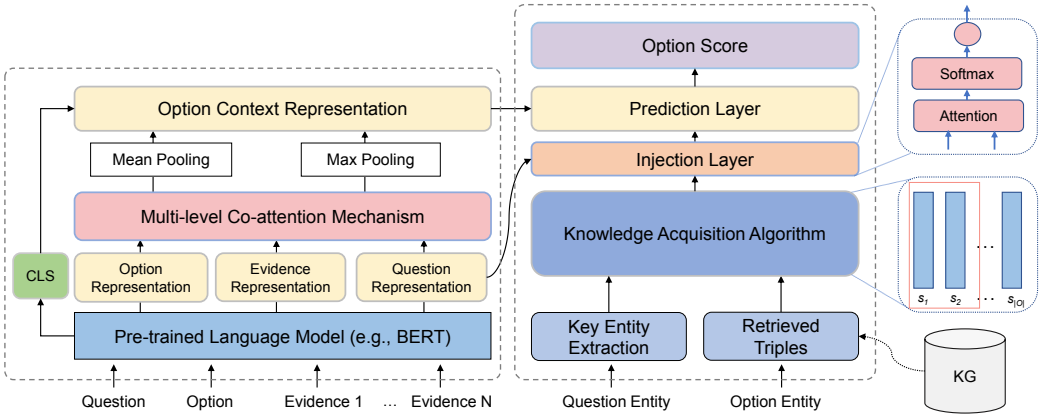

Figure 1: Overall architecture of the proposed KMQA, with multi-level co-attention reader (left) and the knowledge integration part (right) illustrated.

et al., 2017; Seo et al., 2017; Huang et al., 2019b). Taking the candidate answer representation O as input, we compute three types of attention weights to capture its correlation to the question, the evidence, and both the evidence and question, and get question-attentive, evidence-attentive, and question and evidence-attentive representations:

$$\tilde{\mathbf{H}}_O = \mathbf{H}_O \mathbf{W}_t + \boldsymbol{b}_t, \tag{1}$$

$$\mathbf{A}_O^Q = \text{Softmax}(\tilde{\mathbf{H}}_O \mathbf{H}_Q^\top)\mathbf{H}_Q \in \mathbb{R}^{l\times h}, \tag{2}$$

$$\mathbf{A}_O^E = \text{Softmax}(\tilde{\mathbf{H}}_O \mathbf{H}_E^\top)\mathbf{H}_E \in \mathbb{R}^{l\times h}, \tag{3}$$

$$\mathbf{A}_O^{QE} = \text{Softmax}(\tilde{\mathbf{H}}_O \mathbf{H}_{QE}^\top)\mathbf{H}_{QE} \in \mathbb{R}^{l\times h}, \tag{4}$$

where $\mathbf{W}_t$ and $\boldsymbol{b}_t$ are learnable parameters. Next we fuse these representations as follows:

$$\mathbf{T}_O = \text{LSTM}([\mathbf{A}_O^Q; \mathbf{A}_O^E; \mathbf{A}_O^{QE}]) \in \mathbb{R}^{l\times h}, \tag{5}$$

where $[;]$ denotes concatenation operation. Finally, we apply column-wise max and mean pooling on $\mathbf{T}_O$ and concatenate it with $\mathbf{H}_{cls}$. It obtains the new option representation $\tilde{\mathbf{T}}_O \in \mathbb{R}^{3h}$.

3.2 Knowledge Acquisition

In this section, we describe the method to extract knowledge facts from knowledge graph in details. Once the knowledge is determined, we can choose the appropriate integration mechanism for further knowledge injection, such as attention mechanism (Sun et al., 2018; Yang et al., 2019; Ma et al., 2019), pre-training tasks (He et al., 2019) and multi-task training (Xia et al., 2019).

Given a question Q and a candidate answer O, we first identify the entity and its type in the text by entity linking. The identified entity exactly matches the concept in KG. We also perform soft matching of part-of-speech rules and filter out stop words, and obtain key entities for Q according to category description, such as "*western medicine*", "*symptoms*", "*Chinese herbal medicine*" as $\mathcal{E}_Q$. After that, we retrieve all triples $\mathcal{S}_O$ whose head or tail contains the entities of O as knowledge facts for this option. For these knowledge facts, we first convert head-relation-tail tokens into regular words by template function $\mathbf{g}$ in order to generate a pseudo-sentence. For example, "(*chronic hepatitis B*, *Site of disease*, *Liver*)" is converted to "*The site of disease of chronic hepatitis B is liver*". Then we can get re-rank option facts for each question-answer pair with the method shown in Algorithm 1, which uses the word mover's distance (Kusner et al., 2015) as similarity function empirically. The reason we apply it is to be able to find higher-quality knowledge facts that are more relevant to current option and input them into the model. The embedding function $\mathcal{F}$ here is the mean

Algorithm 1 Knowledge Acquisition Algorithm

Require: Question q and entities $\mathcal{E}_Q = \{e\}$, option facts $\mathcal{S}_O = \{(h, r, t)\}$, embedding function $\mathcal{F}$, template function $\mathbf{g}$

Translate triple $s_j = (h_j, r_j, t_j) \in \mathcal{S}_O$ to general text p_j using $\mathbf{g}$
if $\mathcal{E}_Q$ is empty set **then**
  Calculate knowledge-based option scores for each p_j using the word mover's distance $\texttt{wmd}(\mathcal{F}(q), \mathcal{F}(p_j))$
  return `top-K` option facts ranking by score in the ascending order
end if
Initialize similarity vector $\mathbf{o} \in \mathbb{R}^{|\mathcal{S}_O|}$ with infinities.
Calculate the entity-to-triple score $c_{i,j}$ of entity e_i with transformed text p_j: $\texttt{wmd}(\mathcal{F}(e_i), \mathcal{F}(p_j))$
Set the j-th element of similarity vector $o_j = \min_{i\in|\mathcal{E}_Q|}\{c_{i,j}\}$
return `top-K` option facts ranking by $\mathbf{o}$ in the ascending order

pooling of sentence word vectors. The word embedding uses 200-dimension pre-trained embedding for Chinese words and phrases (Song et al., 2018). Although not perfect, the triple text found by Algorithm 1 does provide some useful information that can help the model find the correct answer.

3.3 Knowledge Injection and Answer Prediction

We first concatenate the returned option fact text as F, and then use the $\mathcal{B}$ to generate an embedding of this pseudo-sentence:

$$\mathbf{H}_F = \mathcal{B}(F). \tag{6}$$

Let $\mathbf{H}_F \in \mathbb{R}^{s\times h}$ be the concatenation of the final hidden states, where s is max length, and we then adopt the attention mechanism to model the interaction between $\mathbf{H}_F$ and the PLMs encoding output of question $\mathbf{H}_Q$:

$$\boldsymbol{\mathcal{M}}_{FQ} = (\mathbf{W}_{fq} \circ \mathbf{H}_F)\mathbf{H}_Q^\top, \tag{7}$$

$$\mathbf{A}_Q^F = \text{Softmax}(\boldsymbol{\mathcal{M}}_{FQ})\mathbf{H}_Q, \tag{8}$$

$$\mathbf{A}_F^Q = \text{Softmax}(\boldsymbol{\mathcal{M}}_{FQ})\text{Softmax}(\boldsymbol{\mathcal{M}}_{FQ}^\top)\mathbf{H}_F, \tag{9}$$

$$\mathbf{H}_{\boldsymbol{FQ}} = [\mathbf{H}_F; \mathbf{A}_Q^F; \mathbf{H}_F \circ \mathbf{A}_Q^F; \mathbf{H}_F \circ \mathbf{A}_F^Q], \tag{10}$$

$$\mathbf{T}_F = \text{Tanh}(\mathbf{H}_{\boldsymbol{FQ}}\mathbf{W}_{proj}), \tag{11}$$

where element-wise multiplication is denoted by $\circ$. Specifically, $\mathbf{H}_F$ is linear transformed using $\mathbf{W}_{fq} \in \mathbb{R}^{s\times h}$. Then, the similarty matrix $\boldsymbol{\mathcal{M}}_{FQ} \in \mathbb{R}^{s\times n}$ is computed using standard attention. Then we use $\boldsymbol{\mathcal{M}}_{FQ}$ to compute question-to-knowledge attention $\mathbf{A}_Q^F \in \mathbb{R}^{s\times h}$ and knowledge-to-question attention $\mathbf{A}_F^Q \in \mathbb{R}^{s\times h}$. Finally, the question-aware knowledge textual representation $\mathbf{T}_F \in \mathbb{R}^{s\times h}$ is computed, where $\mathbf{W}_{proj} \in \mathbb{R}^{4h\times h}$. Finally, max pooling and mean pooling are applied on the $\mathbf{T}_F$ to generate final knowledge representation $\tilde{\mathbf{T}}_F \in \mathbb{R}^{2h}$. In the output layer, we combine textual representation $\tilde{\mathbf{T}}_O$ with the knowledge representation $\tilde{\mathbf{T}}_F$. For each candidate answer O_i, we compute the loss as follows:

$$\mathbf{T}_C = [\tilde{\mathbf{T}}_O; \tilde{\mathbf{T}}_F], \tag{12}$$

$$Score(O_i|E,Q,F) = \frac{\exp(\mathbf{W}_{out}^\top \mathbf{T}_C^i)}{\sum_{j=1}^5 \exp(\mathbf{W}_{out}^\top \mathbf{T}_C^j))}, \tag{13}$$

where $\mathbf{W}_{out} \in \mathbb{R}^{1\times 5h}$. We add a simple feed-forward classifier as the output layer which takes the contextualized representation $\mathbf{T}_C$ as input and outputs the answer score $Score(O_i|E,Q,F)$. Finally, the candidate with the highest score is chosen as the answer. The final loss function is obtained as follows:

$$\mathcal{L} = -\frac{1}{C}\sum_i \log(Score(\hat{O}_i|E,Q,F)) + \lambda||\theta||_2, \tag{14}$$

where C is the number of training examples, and $\hat{O}_i$ is the ground truth for the i-th example, θ denotes all trainable parameters.

3.4 Augmenting with Path Information

For concepts in question and options (remove entities that are not diseases, drugs, and symptoms), we combine them in pairs and retrieve all paths between them within 3 hops to form a sub-graph about the option. For example, (*chronic hepatitis B* → *related diseases* → *cirrhosis* → *medical treatment* → *entecavir*) is a path for (*chronic hepatitis B*, *entecavir*).

Then, we apply L layer graph convolutional networks (Kipf and Welling, 2017) to update the representation of the nodes, which is similar to (Lin et al., 2019; Yang et al., 2019). Here, we set L equals 2. The vector $\boldsymbol{h}_i^{(0)} \in \mathbb{R}^h$ for concept c_i in the sub-graph g is initialized by the average embedding vector of tokens similar to §3.2. Then, we update them at $(l+1)$-th layer using the following equation:

$$\boldsymbol{h}_i^{(l+1)} = \sigma(\mathbf{W}_{gcn}\boldsymbol{h}_i^{(l)} + \sum_{j\in\mathcal{N}_i}\frac{1}{|\mathcal{N}_i|}\mathbf{W}_{gcn}\boldsymbol{h}_j^{(l)}), \tag{15}$$

where $\mathcal{N}_i$ is the neighboring nodes, σ is ReLU activation function, $\mathbf{W}_{gcn}$ is the weight vector. After that, we update i-th tokens representation $\boldsymbol{t}_i \in \mathbf{T}_O$ with the corresponding entity vector via a sigmoid gate to the new token representation $\boldsymbol{t}_i'$:

$$g_i = \text{Sigmoid}\left(\mathbf{W}_s\left[\boldsymbol{t}_i; \boldsymbol{h}_i^L\right]\right), \tag{16}$$

$$\boldsymbol{t}_i' = g_i \circ \boldsymbol{t}_i + (1-g_i) \circ \boldsymbol{h}_i^L. \tag{17}$$

4 Dataset

We use the National Licensed Pharmacist Examination in China [4] as the source of questions. The exam is a comprehensive evaluation of the professional skills of candidates. Medical practitioners have to pass the examination to obtain the qualification for licensed pharmacist in China. Passing the

[4] http://english.nmpa.gov.cn/2019-07/19/c_389177.htm

exam requires getting a minimum of 60% of the total score. The pharmacy comprehensive knowledge and skills part of the exam consists of 600 multiple-choice problems over four categories. To test the generalizability of MRC models, we use the examples of this part in the previous five years (2015-2019) as the test set, and exclude questions of multiple-answer type. In addition to that, we also collected over 24,000 problems from the Internet and exercise books. After removing duplicates and incomplete questions (e.g. no answer), we randomly divide it into training, development sets according to a certain ratio, and remove the problems similar to the test set according to the condition that the edit distance is less than 0.1. The detailed statistics of the final problem set, named as NLPEC, are shown in Table 2.

	Train	Dev	Test
# Questions	18,703	2,500	550
Avg. words of questions	16.72	17.15	42.82
Avg. words of candidate options	3.48	3.38	3.62
Avg. words of retrieval evidences	84.17	81.75	86.09
Avg. sentences of each evidence	3.82	3.79	4.02
Candidate options per problem		5	

Table 2: Statistics of our NLPEC dataset.

We use the official exam guide book of the National Licensed Pharmacist Examination as text source (NMPA, 2018). It has 20 chapters, including pharmaceutical practice and medication, self-medication for common diseases, and medication for organ system diseases. The book covers most of the necessary contents of the examination. In order to ensure the quality of retrieval, we first convert it into structured electronic versions through OCR tools, and then manually proofread and divide all the texts into paragraphs. Meanwhile, we also extract passages from other literature and add it to the text source, including the pharmacological effects and clinical evaluation of various drugs, explanations of drug monitoring and descriptions of essential medicines.

5 Experiment

5.1 Experiment Settings

We use the Google-released BERT-base model as the PLM (Devlin et al., 2019). We also compare the performance of KMQA, which uses the pre-trained RoBERTa large model (Liu et al., 2019). The pre-trained weights that we adopt are the version of whole word masking in Chinese text (Cui et al., 2019). Our model is also orthogonal to the choice of the pre-trained language model. We use AdamW optimizer (Loshchilov and Hutter, 2019) with a batch size of 32 for model training. The initial learning rate, the maximum sequence length, the learning rate warmup proportion, the gradient accumulation steps, the training epoch, the hidden size h, λ, the number of evidence spans N, and the hyperparameter K are set to 3×10^{-5}, 512, 0.1, 8, 10, 768, 1×10^{-6}, 1, and 3 respectively. The learning parameters are selected based on the best performance on the development set. Our model takes approximately 22 hours to train with 4 NVIDIA Tesla V100. In order to reduce memory usage, in our implementation, we concatenate the knowledge text and the retrieved evidence spans, and then obtain separate encoding representations. For other models, the dimension of word embeddings is 200, the hidden size is 256, and the optimizer is Adam optimizer (Kingma and Ba, 2015). We also pre-trained word embeddings on a large-scale Chinese medical text.

Model	Accuracy (%) DEV	TEST
IR baseline	36.4	34.1
Random guess	21.3	22.8
Co-Matching (Wang et al., 2018)	56.1	45.8
BiDAF (Seo et al., 2017)	52.7	43.6
SeaReader (Zhang et al., 2018)	58.2	48.4
Multi-Matching (Tang et al., 2019)	58.4	48.7
BERT-base (Devlin et al., 2019)	64.2	52.2
ERNIE (Sun et al., 2019)	64.7	53.4
RoBERTa-wwm-ext-large (Cui et al., 2019)	70.8	57.9
KMQA (BERT-base)	67.9	57.1
KMQA (RoBERTa-wwm-ext-large)	**71.1**	**61.8**

Table 3: Performance comparison on the test set. Additional details about baselines can be found in the Appendix.

5.2 Main Results

The comparison between our method and previous works on the multi-choice question answering task over our dataset is shown in Table 3. IR baseline refers to the selection of answers using the ranking of the score of the retrieval system, and random guess refers to the selection of answers according to a random distribution. The third to fifth lines show the results of the previous state-of-the-art models. These models all employ the co-matching model and perform better than those two baselines. They use attention mechanisms to capture the correlation between retrieved evidence, questions, and candidate answers, and tend

to choose the answer that is closest to the semantics of the evidence. Pre-trained language models with fine-tuning achieve more than 18% improvement over baselines. By fusion of knowledge source and text over BERT-base, the performance is further improved, which demonstrates our assumption that incorporating knowledge from the structure source can further enhance the option contextual understanding of BERT-base. Furthermore, our single model of KMQA-RoBERTa large, which employs RoBERTa large model pre-trained with whole word mask achieves better performance on both development set and test set and also outperforms RoBERTa large. This result also slightly surpasses the human passing score. These results demonstrate the effectiveness of our method.

Types	Number	Accuracy
Statement Best Choice	200	64.0
Best Compatible Choice	257	58.4
Case Summary Best Choice	90	66.7
Conceptual Knowledge	279	61.3
Situational Analysis	42	64.3
Logical Reasoning	226	62.0
Positive Questions	433	61.9
Negative Questions	114	61.4

Table 4: Performance of our model on different question category.

In the exam, the questions are divided into three types, namely, type A (statement best choice), type B (best compatible choice), and type C (case summary best choice). The evaluation results are listed in Table 4. We observe that the best compatible choice type accounts for the highest proportion of the questions, and the model performance is lower than the other two. According to the different methods required for answering questions, we further divide them into three types: conceptual knowledge, situational analysis, and logical reasoning. For the problem of conceptual knowledge, they account for a lot and are usually related to specific concept knowledge. It means that we also need to improve our retrieval module. According to the needs of the problem to be deduced in a positive or negative direction, we divide the problem into two categories: positive questions and negative questions. We find that their performance is similar, but the positive part accounts for a more significant proportion.

5.3 Ablation Study

To study the effect of each KMQA component, we also conduct ablation experiments. The results are shown in Table 5. From the experimental results, if there is no external information but only questions and options, the model is only 2.5% higher than the retrieval baseline. After adding the information retrieved by the text retrieval model and knowledge graph, the model is improved by 26.3% and 6.4% respectively, which shows the effectiveness of external information. Further, we find that pre-training on relation classification can also improve the performance of our downstream QA tasks. When the path information from the question to the option is further added, the model has 0.8% improved accuracy. If we only use retrieved snippets from reference books with the co-attention mechanism, the model has more performance drops. We also change the hyper-parameter K, and results show that the setting $K = 3$ performs best. Due to the max length of BERT model, a larger K will not bring more improvements.

Model	Accuracy (DEV)
Ours (BERT-base)	67.9
w/o relation classification	66.4
w/o extracted facts	65.2
w/o path information	67.1
w/o text source	45.3
w/o knowledge source	64.6
only option	38.9
K = 1 (RoBERTa)	70.2
K = 2 (RoBERTa)	70.6
K = 3 (RoBERTa)	71.1

Table 5: Ablation study in development set.

5.4 Case Study

As shown in Table 6, we choose an example to visualize joint reasoning using KG and retrieval text. In Example 1 of Table 6, we find that limited by the process of retrieval, some of the descriptions of the indications of the option are not completely relevant to the question stem, and the paragraphs contain descriptions of the chemical composition of this drug, which is noisy for answering the question. In contrast, our model is able to answer this question using both KG and textual evidence, alleviating the noise problem to some extent. Since many of the questions in our dataset are about diseases and drugs that require descriptions of their underlying meanings, using the medical KG may be the most convenient for our research.

Type	Examples
Positive Example	**Question**: 患者，男，38岁，因腹部受寒致胃痉挛性疼痛，应选用的药物是? The patient, male, 38 years old, suffers from **stomach spasmodic pain** caused by **abdominal cold**. Which of the following drugs should be chosen? **Options:** ✓ (A). 山莨菪碱 Anisodamine. × (B). 布洛芬 Ibuprofen. × (C). 麦角胺咖啡因 Ergotamine caffeinee. × (D). 卡马西平 Carbamazepine. × (E). 吗啡 Morphine. **Evidence spans:** 对腹痛较重者或反复呕吐性腹泻者腹痛剧烈时可服山莨菪碱片，一次5mg，一日3次或痛时服用... 山莨菪碱与莨菪碱在结构上的区别是，结构中醇部分为6-(S)-羟基莨菪醇（亦称山莨菪醇），与托品醇相比，在6位多了一个β-取向的羟基，这使得山莨菪碱分子的极性增强，难以透过血-脑屏障，中枢作用很弱... Anisodamine tablets can be taken for severe **abdominal pain** or recurrent vomiting diarrhea when abdominal pain is severe, 5 mg once, 3 times a day or when pain occurs... The structural difference between anisodamine and scopolamine is that the alcohol part in the structure is 6-(S)-hydroxy scopolamine (also known as anisodamine), which has a β-oriented hydroxyl group at the 6-position compared with tropinol, which makes the polarity of the anisodamine molecule enhanced, it is difficult to penetrate the blood-brain barrier, and the central role is weak... **Knowledge facts: 1.** (山莨菪碱, 适应症, 疼痛) The **indication** for anisodamine is **pain**. **2.** (山莨菪碱, 适应症, 胃肠绞痛) The **indication** for anisodamine is **spasm**. **3.** (山莨菪碱, 适应症, 痉挛) The indication for anisodamine is gastrointestinal colic. **A sample path:** 胃痉挛 → 相关疾病 → 胃病 → 临床症状及体征 → 急性单纯性胃炎 → 治疗方案 → 山莨菪碱 gastric spasm → related diseases → gastropathy → clinical symptoms and signs → acute simple gastritis → treatment plan → anisodamine
Negative Example 1 (Noisy Evidence)	**Question**: 从事驾车、高空作业的患者不宜服用的药物是? Which drugs should not be taken by patients engaged in driving and high altitude work? **Golden answer**: 氯苯那敏 Chlorpheniramine **Predicted distractor**: 伪麻黄碱 Pseudoephedrine **Evidence spans**: 组胺H2受体阻断剂雷尼替丁、西咪替丁、法莫替丁能引起幻觉、定向力障碍。因此，对驾车司机、高空作业者、精密仪器操作者慎用，或提示在服用后休息6h再从事工作。 Histamine H2 receptor blockers ranitidine, cimetidine and famotidine can cause hallucination and disorientation. Therefore, drivers, high-altitude operators, precision instrument operators should be cautious to use, or prompt to rest for 6 hours before working. **Knowledge facts**: (氯苯那敏, 注意事项, 驾驶员、机械操作人员在工作进行时不宜使用)。 The precaution of chlorpheniramine is that it should not be used by drivers and mechanical operators during work. **Evidence spans of wrong answer**: ...氨酚伪麻美芬片II/氨麻苯美片、美扑伪麻片中还含有H1受体拮抗剂成分，可能引起头晕、嗜睡，故服药期间不宜驾车或高空作业、操纵机器... ..., paracetamol pseudoephedrine tablets II/amphetamine tablets, and melphalan pseudoephedrine tablets also contain H1 receptor antagonist components, which may cause dizziness and sleepiness. ***So, it is inappropriate to drive or operate machines at high altitude during medication administration***...
Negative Example 2 (Weak Reasoning)	**Question**: 下列中药、化学药联合应用，不存在重复用药的是? The following Chinese medicine and chemical medicine are used together. Which option does not exist for repeated medicine? **Golden answer**: 曲克芦丁片＋维生素C片 Troxerutin Tablets + Vitamin C Tablets **Predicted distractor**: 珍菊降压片＋氢氯噻嗪片 Zhenju Antihypertensive Tablets + Hydrochlorothiazide Tablets **Evidence spans**: （2）充分询问进食情况及用药史，避免重复用药引发维生素D中毒... (2) Fully inquire about food intake and medication history to avoid vitamin D poisoning caused by repeated medication... **Knowledge facts of wrong answer**: (珍菊降压片, 注意事项, 对氢氯噻嗪、可乐定、磺胺类药物过敏者忌用) ***The precautions of Zhenju Antihypertensive Tablets are to avoid the use of hydrochlorothiazide***, clonidine and sulfonamides in allergic patients...

Table 6: Case study and error examples of the proposed KMQA.

In addition, we randomly select 50 errors made by our approach from the test set, and categorize them into 4 groups:

Information Missing: In 44% of the errors, the retrieved evidence and extracted knowledge cannot provide useful information to distinguish different answer candidates, which is the major error type in our model. Taking the case "*What does the abbreviation - p.c. - stand for in prescription?*" as an example, to correctly predict the answer, we need to know that "*p.c.*" is the abbreviation that means "*after meals*" (from the Latin "*post cibum*").

Noisy Evidence: In 32% of the errors, the model is misled by noisy knowledge of other wrong answers. The reason may be that the context is too long and overlaps with the problem description. For example, in Example 2 of Table 6, both the right answer and wrong prediction could be potentially selected by retrieval evidence. However, we can intuitively get the answer through mutual verification of essential information in KG and retrieved texts.

Weak Reasoning Ability: 14% of the errors are due to the weak reasoning ability of the model, such as the understanding of symbolic units in options. For example, in Example 3 of Table 6, the model needs to first understand the joint meaning of options using common sense, and then eliminate the wrong answer with counterfactual reasoning through knowledge and text.

Numerical Analysis: 10% of the errors are from mathematical calculation and analysis questions. The model cannot handle the question like "*To prepare 1000ml 70% ethanol with 95% ethanol and distilled water, what is the volume of 95% ethanol needed?*" properly since it cannot be directly entailed by the given paragraph. Instead, it requires mathematical calculation and reasoning ability of the model.

6 Conclusion

In this work, we explore how to solve multi-choice reading comprehension tasks in the medical field based on the examination problems of licensed pharmacists, and propose a novel model KMQA. It explicitly combines knowledge and pre-trained models into a unified framework. Moreover, KMQA implicitly takes advantage of factual information via learning from an intermediate task and also transfers structural knowledge to enhance entity

representation. On the test set from the real world, the KMQA is the single model that outperforms the human pass line. In the future, we will explore how to apply our model to more domains, and enhance the interpretability of the reasoning path when the model answers questions.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions. This work is supported by Natural Science Foundation of China (61872113, U1813215, 61876052), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20180306172232154), and the fund of the joint project with Beijing Baidu Netcom Science Technology Co., Ltd.

References

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Odma Byambasuren, Yunfei Yang, Zhifang Sui, Damai Dai, Baobao Chang, Sujian Li, and Hongying Zan. 2019. Preliminary study on the construction of chinese medical knowledge graph. *Journal of Chinese Information Processing*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *ACL*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL-HLT*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Yu Fang, Shimin Yang, Siting Zhou, Minghuan Jiang, and Jun Liu. 2013. Community pharmacy practice in china: past, present and future. *International Journal of Clinical Pharmacy*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *ICML*.

Bin He, Di Zhou, Jinghui Xiao, Qun Liu, Nicholas Jing Yuan, Tong Xu, et al. 2019. Integrating graph contextualized knowledge into pre-trained language models. *arXiv preprint arXiv:1912.00147*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019a. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019b. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In *EMNLP-IJCNLP*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP-IJCNLP*.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *ACL*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *EMNLP*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP-IJCNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *AAAI*.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. 2019. Towards generalizable neuro-symbolic systems for commonsense question answering. In *EMNLP*.

Certification Center For Licensed Pharmacist of National Medical Products Administration in China NMPA. 2018. *National Licensed Pharmacist Exam Book 2019 Western Medicine Textbook Licensed Pharmacist Exam Guide Pharmacy Comprehensive Knowledge and Skills (Seventh Edition)*. China Medical Science and Technology Press.

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

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? In *ACL*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL*.

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

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *NAACL-HLT*.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W. Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *EMNLP*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Min Tang, Jiaran Cai, and Hankz Hankui Zhuo. 2019. Multi-matching network for multiple choice reading comprehension. In *AAAI*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Éric Gaussier, Patrick Gallinari, Thierry Artières, Michael R. Alvers, Matthias Zschunke, and Axel-Cyrille Ngonga Ngomo. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI*.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *ACL*.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *ACL*.

Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. A co-matching model for multi-choice reading comprehension. In *ACL*.

Jiangnan Xia, Chen Wu, and Ming Yan. 2019. Incorporating relation knowledge into commonsense reading comprehension with multi-task learning. In *CIKM*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *ICLR*.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with

rich knowledge for machine reading comprehension. In *ACL*.

Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2019. Pre-trained language model for biomedical question answering. In *Machine Learning and Knowledge Discovery in Databases - International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part II*.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrqa dataset. In *ACL*.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *AAAI*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *AAAI*.

A Compared Methods

BiDAF (Seo et al., 2017) is a representative network for machine comprehension. It is a multi-stage hierarchical process that represents context at different levels of granularity and uses a bi-directional attention flow mechanism to achieve a query-aware context representation without early summarization.

Co-matching (Wang et al., 2018) uses the attention mechanism to match options with the context that composed of paragraphs and the question, and output the attention value to score the options. It is used to solve the single paragraph reading comprehension task of a single answer question.

Multi-Matching (Tang et al., 2019) applies the Evidence-Answer Matching and Question-Passage-Answer Matching module to gather matching information and integrate them to get the scores of options.

SeaReader (Zhang et al., 2018) is proposed to answer questions in clinical medicine using knowledge extracted from publications in the medical domain. The model extracts information with question-centric attention, document-centric attention, and cross-document attention, and then uses a gated layer for denoising.

BERT (Devlin et al., 2019) achieves remarkable state-of-the-art performance across a wide range of related tasks, such as textual entailment, natural language inference, question answering. It first trains a language model on an unsupervised large-scale corpus, and then the pre-trained model is fine-tuned to adapt to downstream tasks.

RoBERTa (Liu et al., 2019) is based on BERT's language masking strategy and modifies key hyper-parameters in BERT, including changing the target of BERT's next sentence prediction, and training with a larger bacth size and learning rate. It has achieved improved results than BERT on different data sets.

ERNIE (Sun et al., 2019) is designed to learn language representation enhanced by knowledge masking strategies, which includes entity-level masking and phrase-level masking. It achieves state-of-the-art results on five Chinese natural language processing tasks.

	TRAIN	DEV	TEST
# Knowledge facts	1,129,780	50,000	50,000

Model	Accuracy (TEST)
RoBERTa-wwm-ext-large (Cui et al., 2019)	89.4
RoBERTa-wwm-ext-large (w/o fine-tuning)	50.8
BERT-base (Devlin et al., 2019)	88.8
BERT-base (w/o fine-tuning)	50.6
DPCNN (Johnson and Zhang, 2017)	82.6
TextCNN (Kim, 2014)	67.8
ESIM (Chen et al., 2017)	77.8

Table 7: Data statistics of relation classification task and accuracy results.

B Relation Classification

We also show the dataset that used to pre-train on the relation classification task and the performance of the pre-trained models in this task. We compare several common text classification and matching models, including TextCNN (Kim, 2014), ESIM (Chen et al., 2017), DPCNN (Johnson and Zhang, 2017). For text classification, the input of the model is the concatenation of two entity words. For ESIM, the input layer is softmax multi-classification. Through learning with the relation classification task, pre-trained models achieve improved performance on the divided test set.

C Introduction to Exam

The detailed statistics of exams in recent years are listed in Table 8. The professional qualifications for licensed pharmacists are subject to a national unified outline, unified proposition, and unified organized examination system (Fang et al., 2013). The qualification exam for licensed pharmacists is held on every October. The examination takes

Years	# Applicants (k)	# Participants (k)	Exam ratio (%)	# Passing (k)	Pass ratio (%)
2018	687.5	566.6	82.41	79.9	14.10
2017	675.2	523.2	77.50	153.0	29.19
2016	884.7	728.6	82.38	151.0	20.74
2015	1121.4	937.7	83.62	235.0	25.16
2014	840.2	702.4	83.61	137.1	19.52
2013	402.3	329.8	81.99	51.8	15.72
2012	188.1	146.8	78.09	26.0	17.68
2011	145.9	109.7	75.16	14.4	13.13
2010	132.7	100.6	75.76	11.2	11.12

Table 8: Statistics of this exam in recent years.

two years as a cycle, and those who take the examination of all subjects must pass the examination of all subjects within two consecutive examination years. The professional qualification examination for licensed pharmacists is divided into two professional categories: pharmacy and traditional Chinese pharmacy. The pharmacy exam subjects are (1) pharmacy professional knowledge (first part) (2) pharmacy professional knowledge (second part) (3) pharmacy management and regulations, and (4) pharmacy comprehensive knowledge and skills. The subjects for the examination of traditional Chinese medicine are (1) professional knowledge of traditional Chinese medicine (first part) (2) professional knowledge of traditional Chinese medicine (second part) (3) pharmaceutical management and regulations, and (4) comprehensive knowledge and skills of traditional Chinese medicine.

D Source of Questions

The source website and books of collected questions are (1) www.51yaoshi.com (2) Sprint Paper for the State Licensed Pharmacist Examination-China Medical Science and Technology Press (3) State Licensed Pharmacist Examination Golden Exam Paper - Liaoning University Press (4) Practicing Pharmacist Quiz App (5) The Pharmacist 10,000 Questions App (6) Practicing Pharmacist Medical Library App