A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets

Changchang Zeng [1,2,3], Shaobo Li [1,4,*], Qin Li [5], Jie Hu [5] and Jianjun Hu [6,*]

1 Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China; zengchangchang16@mails.ucas.ac.cn
2 University of Chinese Academy of Sciences, Beijing 100049, China
3 Department of Computer Science and Engineering, Chengdu Neusoft University, Chengdu 611844, China
4 School of Mechanical Engineering, Guizhou University, Guiyang 550025, China
5 College of Big Data Statistics, GuiZhou University of Finance and Economics, Guiyang 550025, China; qinlee85@126.com (Q.L.); jiehu@mail.gufe.edu.cn (J.H.)
6 Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
* Correspondence: lishaobo@gzu.edu.cn (S.L.); jianjunh@cse.sc.edu (J.H.); Tel.: +1-803-777-7304 (J.H.)

1. **Introduction**
2. **Tasks**
3. **Benchmark Dataset & Open Issues**

1. Introduction

The most common way to test whether a person can fully understand a piece of text is to require she/he answer questions about the text. Just like the human language test, reading comprehension is a natural way to evaluate a computer's language understanding ability.

The goal of a typical MRC task is to require a machine to read a (set of) text passage(s) and then answers questions about the passage(s)

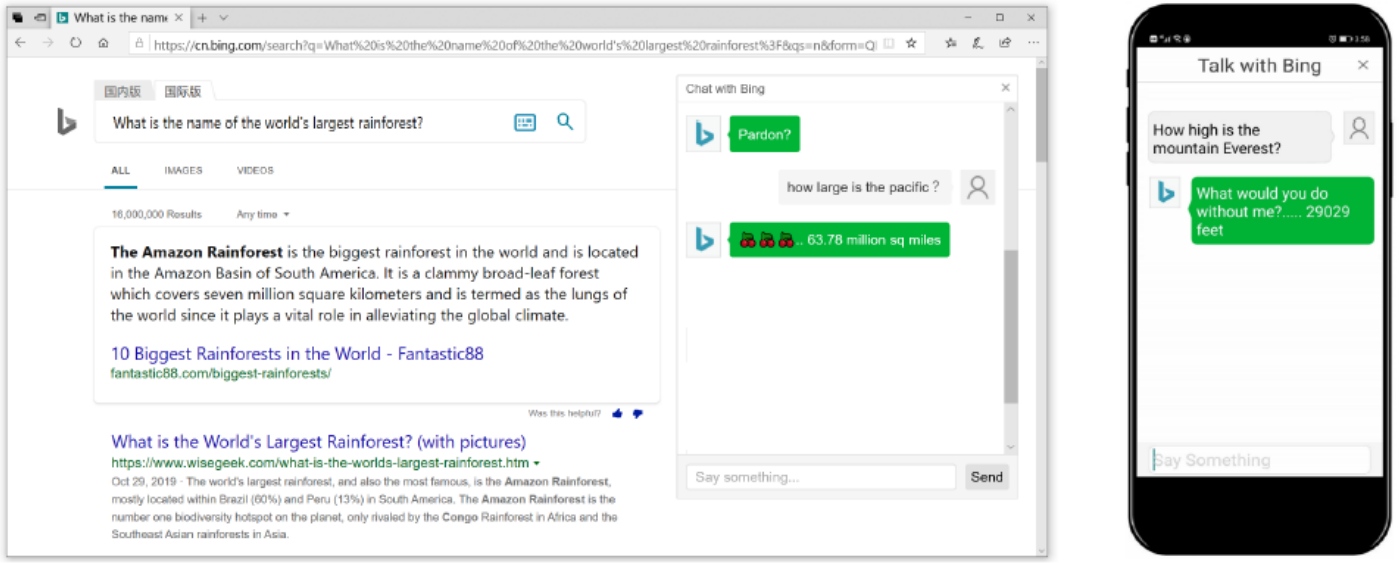


Figure 1. Examples of machine reading comprehension applied to search engine and dialogue system.

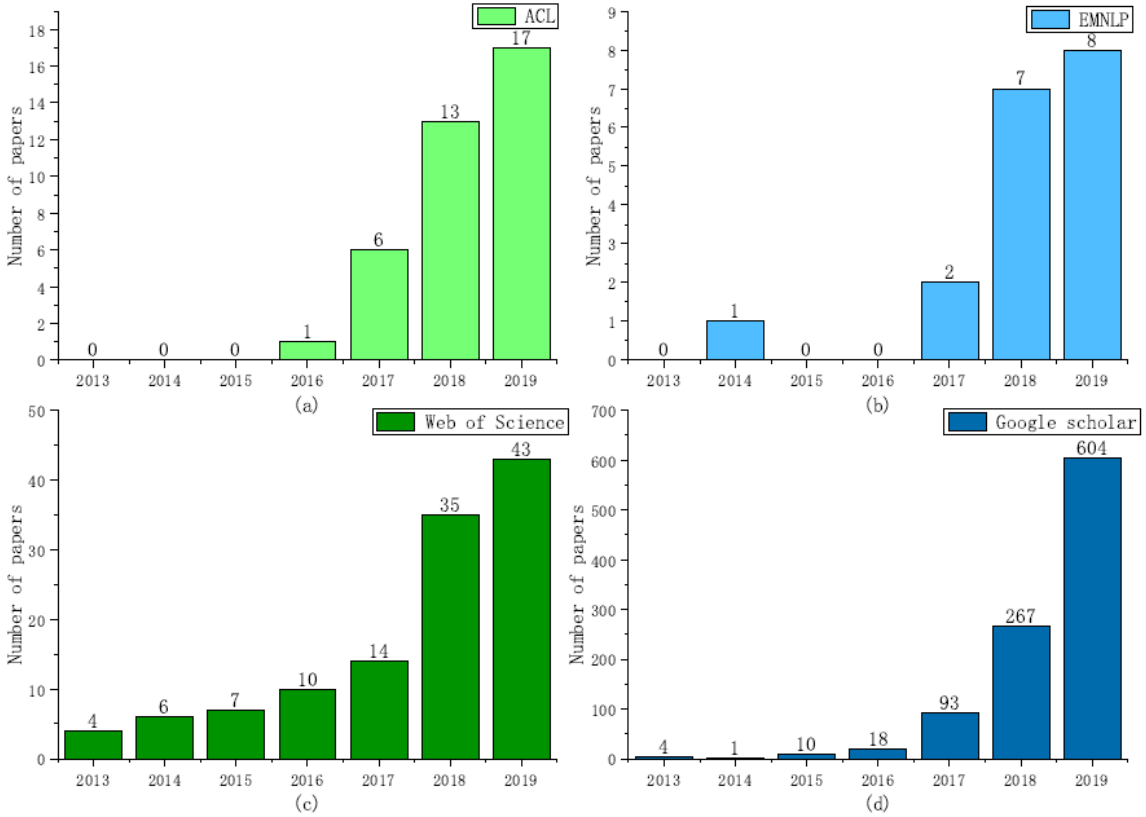


Figure 2. The number of research papers for machine reading comprehension each year: (**a**) The number of research papers on machine reading comprehension (MRC) in ACL from 2013 to 2019. (**b**) The number of research papers on MRC in ENMLP from 2013 to 2019. (**c**) The number of research papers on MRC in Web of Science from 2013 to 2019. (**d**) The number of research papers on MRC in Google scholar from 2013 to 2019.

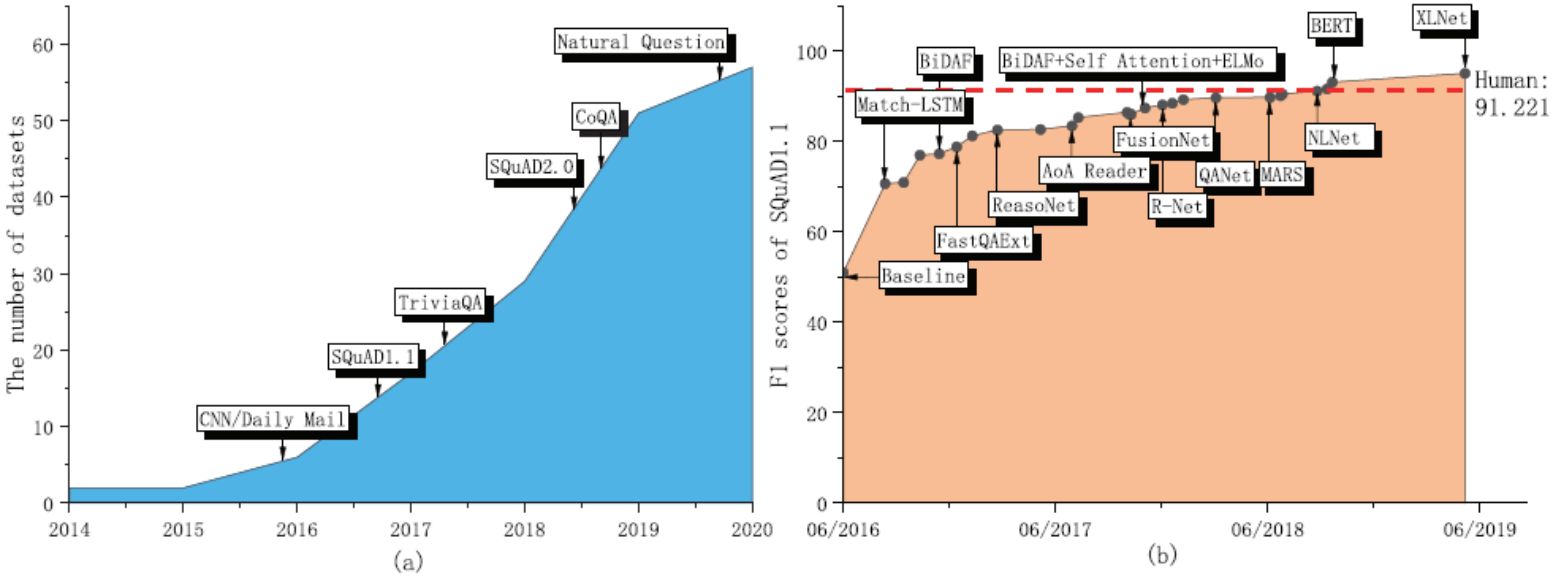
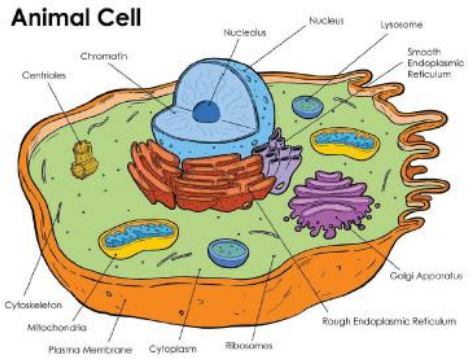


Figure 3. The number of MRC datasets created in recent years and the F1 scores of state-of-the-art models on SQuAD 1.1 [19]: (**a**) The cumulative number of MRC datasets from the beginning of 2014 to the end of 2019. (**b**) The progress of state-of-the-art models on SQuAD 1.1 since this dataset was released. The data points are taken from the leaderboard at https://rajpurkar.github.io/SQuAD-explorer/.

2. Tasks

2.1. Multi-Modal MRC vs. Textual MRC

Passage with illustration:
This diagram shows the anatomy of an Animal cell. Animal Cells have an outer boundary known as the plasma membrane. The nucleus and the organelles of the cell are bound by this membrane. The cell organelles have a vast range of functions to perform like hormone and enzyme production to providing energy for the cells. They are of various sizes and have irregular shapes. Most of the cells size range between 1 and 100 micrometers and are visible only with help of microscope.



Question with illustration:
What is the outer surrounding part of the Nucleus?

Choices:
(1) Nuclear Membrane √
(2) Golgi Body
(3) Cell Membrane
(4) Nucleolus

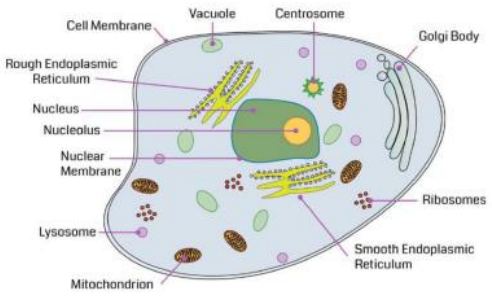
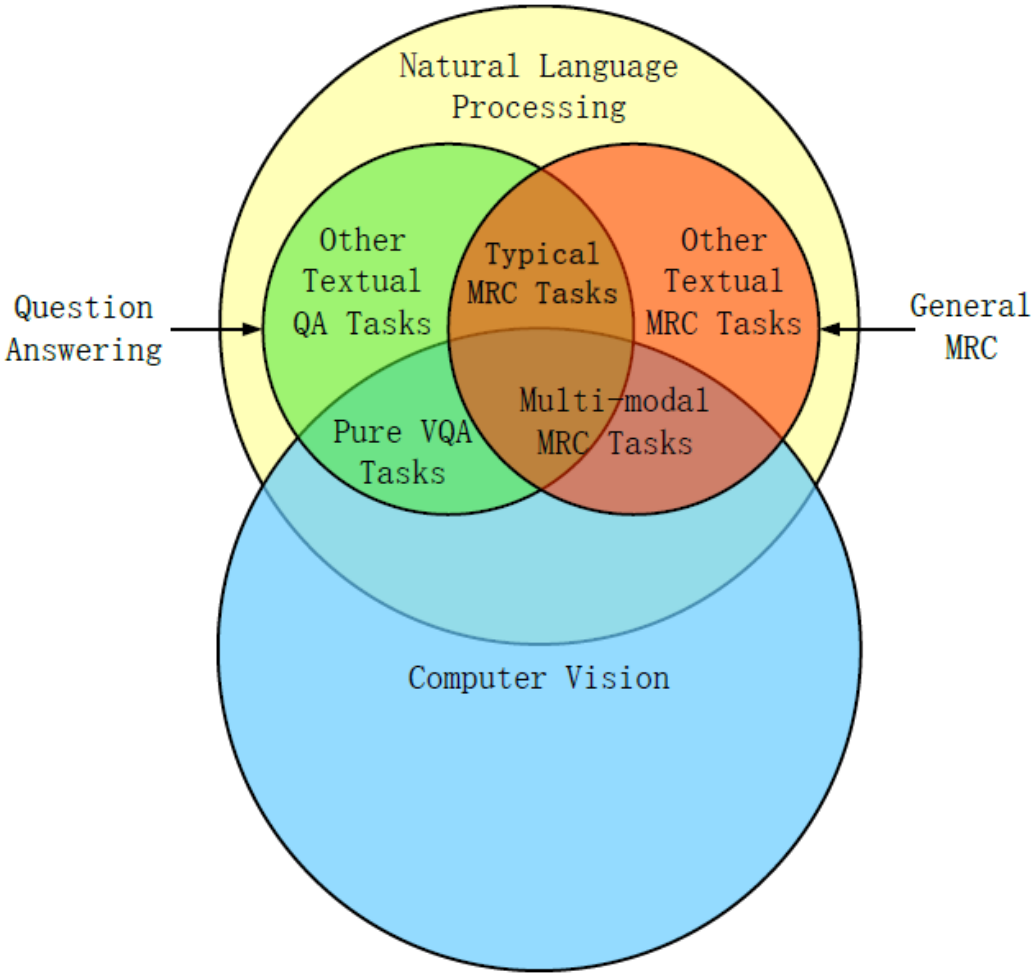


Figure 4. An example of multi-modal MRC task. The illustrations and questions are taken from the TQA [31] dataset.

2.2. Machine Reading Comprehension vs. Question Answering



2.3. Machine Reading Comprehension vs. Other NLP Tasks

1. many useful methods in the field of machine reading comprehension can be introduced into other NLP tasks
   - For example, the stochastic answer network (SAN) [40,41] is first applied to MRC tasks and achieved results competitive to the state of the art on many MRC tasks such as the SQuAD and the MS MARCO. At the same time, the SAN can also be used in natural language processing (NLP) benchmarks, such as Stanford Natural Language Inference (SNLI), MultiGenre Natural Language Inference (MultiNLI), SciTail, and Quora Question Pairs datasets.

2. Secondly, some other NLP research results can also be introduced into the MRC area
3. Thirdly, MRC can be used as a step or component in the pipeline of some complex NLP tasks

2.4. Classification of MRC Tasks

2.4.1. Existing Classification Methods of MRC Tasks

In many research papers [14,26,27], MRC tasks are divided into four categories: cloze style, multiple-choice, span prediction, and free-form answer. Their relationship is shown in Figure 6:

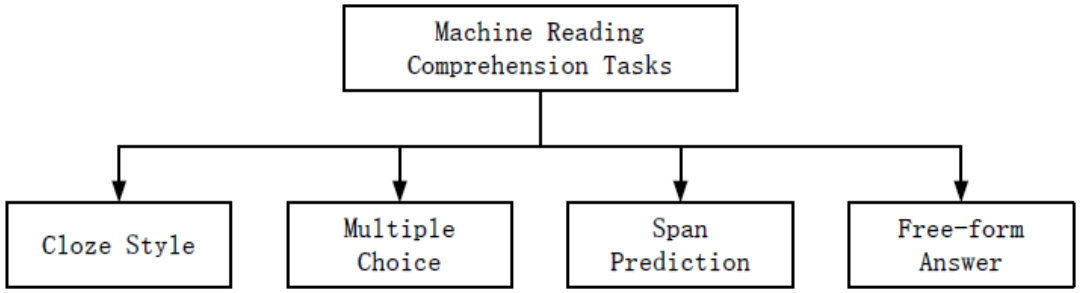


Figure 6. Existing classification method of machine reading comprehension tasks.

2.4.2 Limitations of Existing Classification Method

Passage: Tottenham won 2-0 at Hapoel Tel Aviv in UEFA Cup action on Thursday night in a defensive display which impressed Spurs skipper Robbie Keane... Keane scored the first goal at the Bloomfield Stadium with **Dimitar Berbatov**, who insisted earlier on Thursday he was happy at the London club, heading a second. The 26-year-old Berbatov admitted the reports linking him with a move had affected his performances ... Spurs manager Juande Ramos has won the UEFA Cup in the last two seasons ...
Question: Tottenham manager Juande Ramos has hinted he will allow _____ to leave if the Bulgaria striker makes it clear he is unhappy.
Choices: (A) Robbie Keane (B) **Dimitar Berbatov** √

Figure 7. An example of MRC task. The question-answer pair and passage are taken from the "Who did What" [48].

2.4.2 Limitations of Existing Classification Method

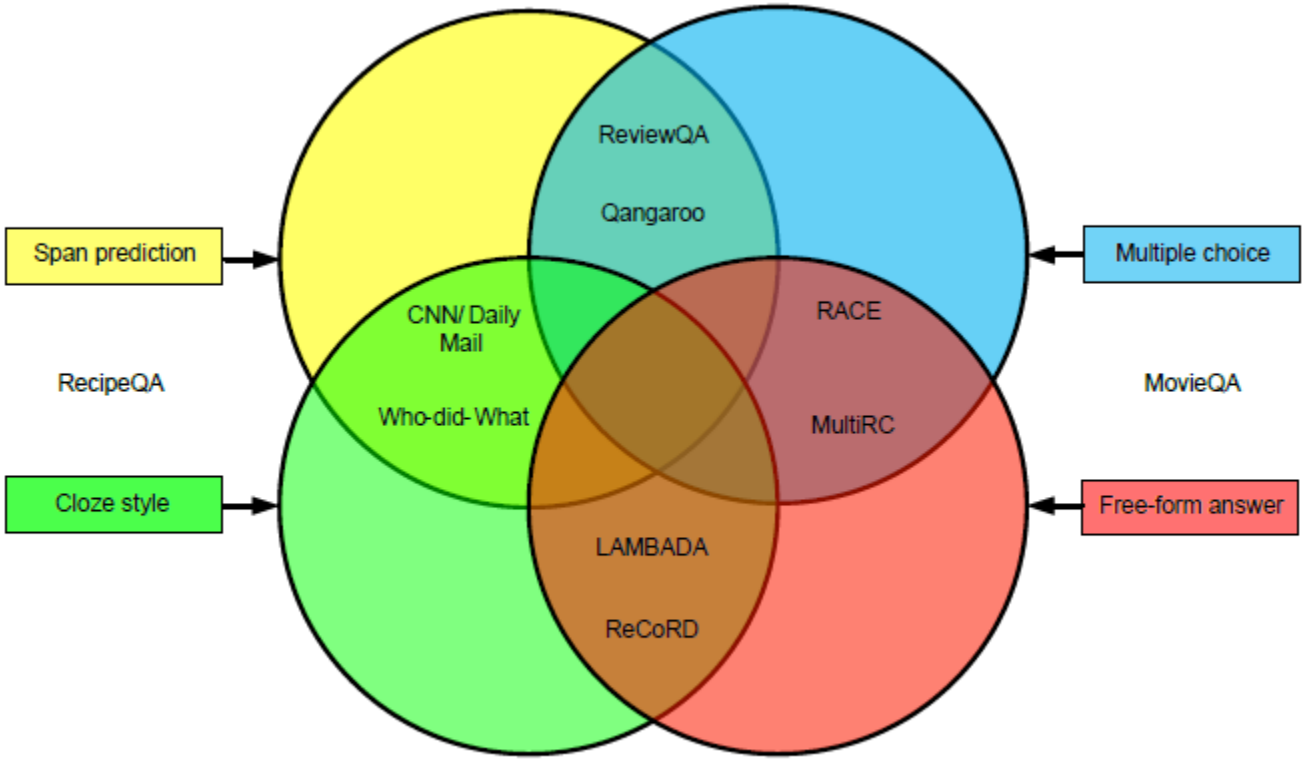


Figure 8. The indistinct classification caused by existing classification method.

2.4.3 A New Classification Method

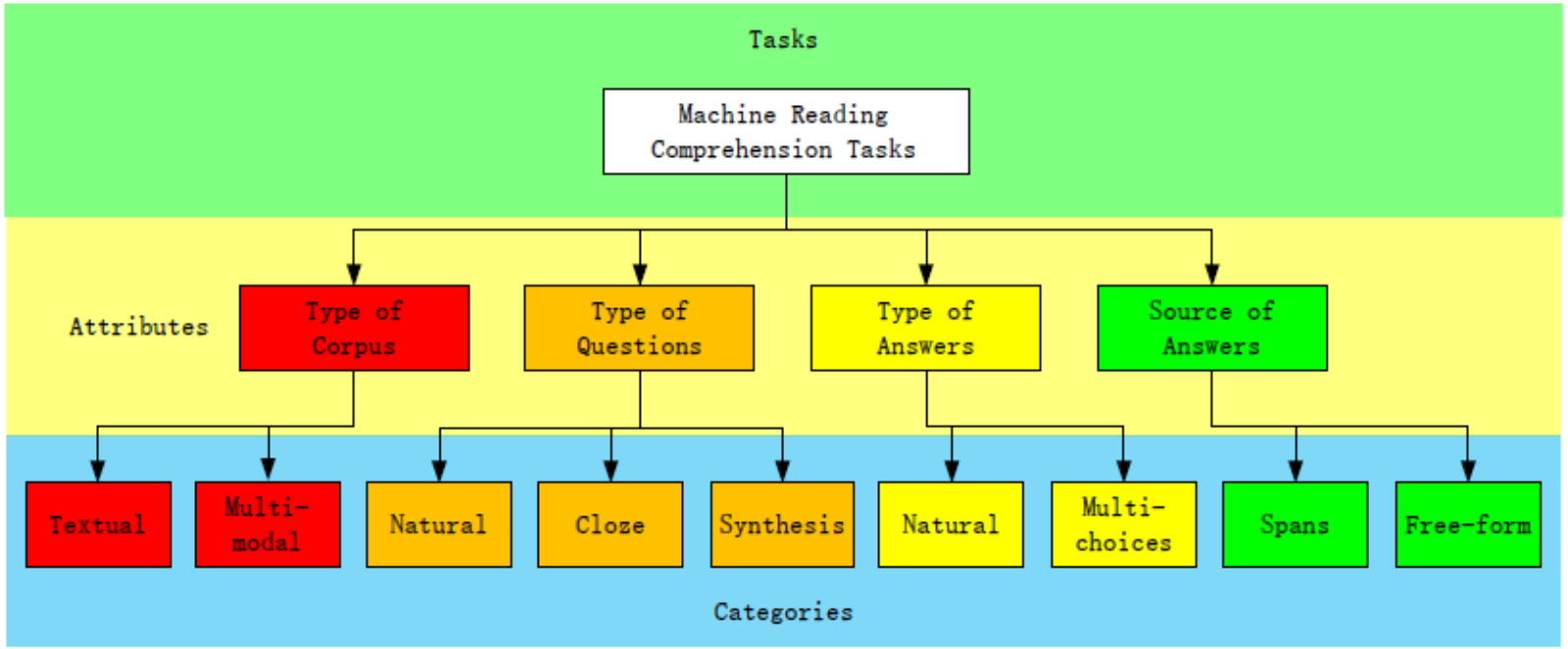
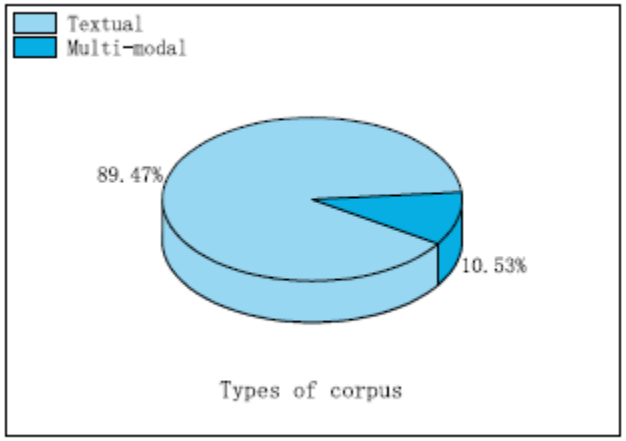
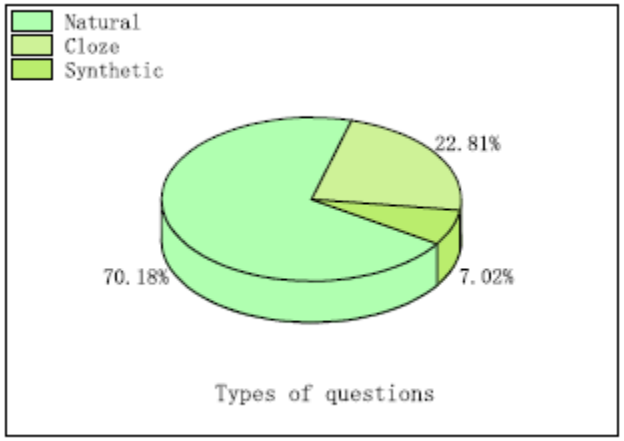


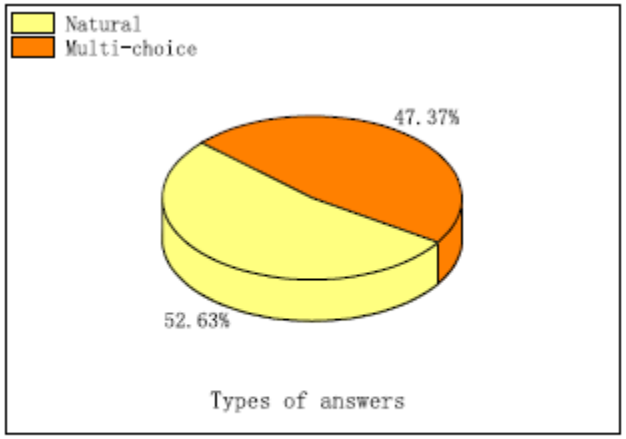
Figure 9. A new classification method of machine reading comprehension tasks.


Textual
Multi-modal
89.47%
10.53%
Types of corpus

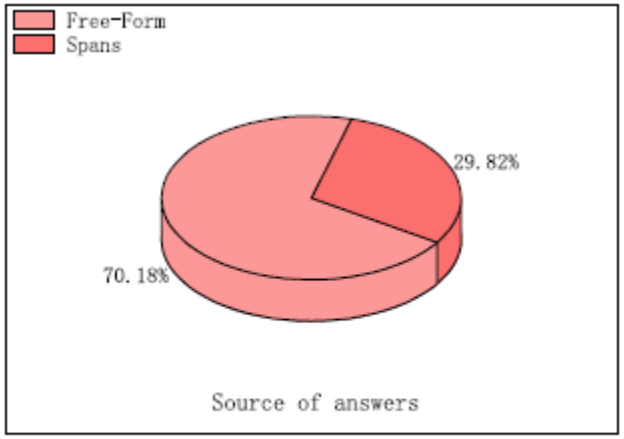
(a)


Natural
Cloze
Synthetic
22.81%
70.18%
7.02%
Types of questions

(b)


Natural
Multi-choice
47.37%
52.63%
Types of answers

(c)


Free-Form
Spans
29.82%
70.18%
Source of answers

(d)

3. Benchmark Dataset & Open Issues

1. MRC with Unanswerable Questions

The existing MRC datasets often lack training sets for unanswerable questions, which weaken the robustness of the MRC systems. As a result, when the MRC models answer unanswerable questions, the models always try to give a most likely answer, rather than refuse to answer these unanswered questions. In this way, no matter how the model answers, the answers must be wrong.

SQuAD 2.0

2. Multi-Hop Reading Comprehension

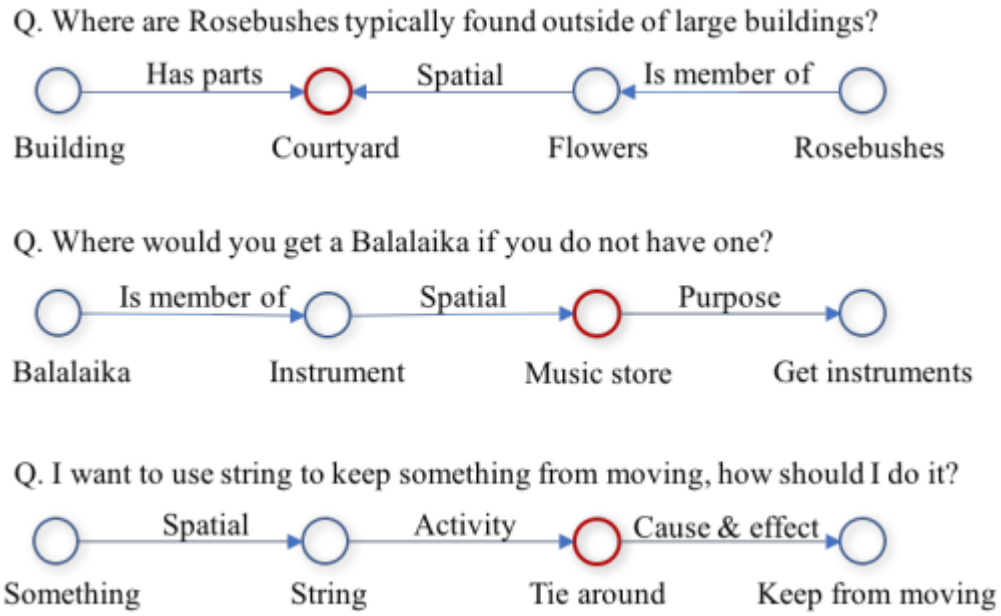
In most MRC dataset, the answer to a question usually can be found in a single paragraph or a document. However, in real human reading comprehension, when reading a novel, we are very likely to extract answers from multiple paragraphs. Compared with single passage MRC, the multi-hop machine reading comprehension is more challenging and requires multi-hop searching and reasoning over confusing passages or documents.

HotpotQA

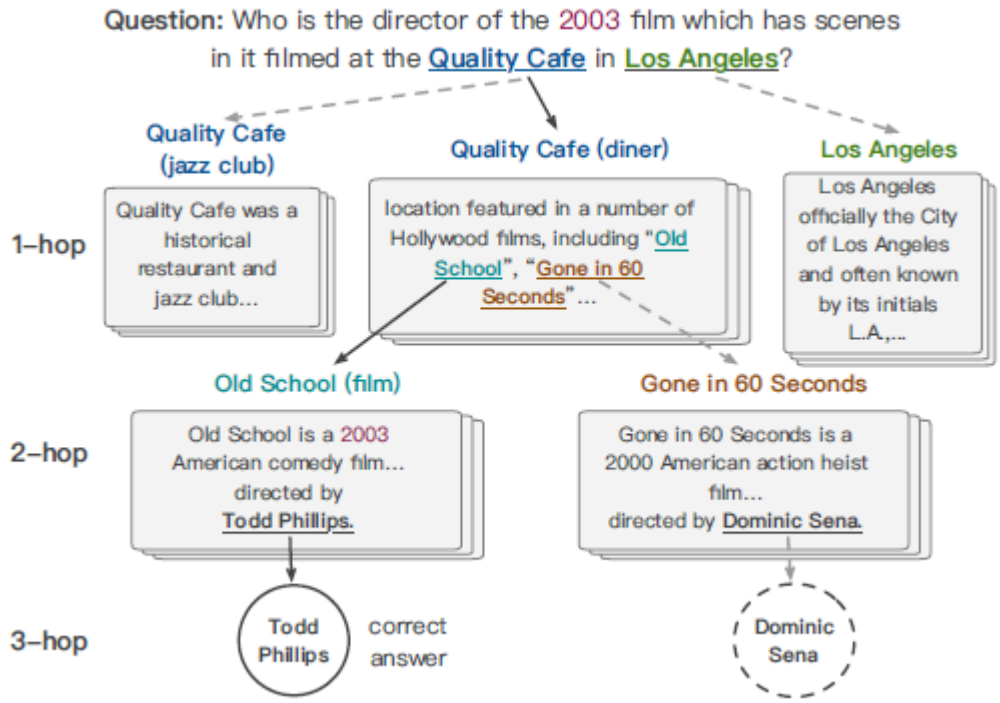
3. Multi-Modal Reading Comprehension

4. Reading Comprehension Require Commonsense or World Knowledge

CommonSenseQA



5. Complex Reasoning MRC



6. Conversational Reading Comprehension

It is a natural way for human beings to exchange information through a series of conversations. In the typical MRC tasks, different question and answer pairs are usually independent of each other. However, in real human language communication, we often achieve an efficient understanding of complex information through a series of interrelated conversations. Similarly, in human communication scenarios, we often ask questions on our own initiative, to obtain key information that helps us understand the situation. In the process of conversation, we need to have a deep understanding of the previous conversations in order to answer each other's questions correctly or ask meaningful new questions. Therefore, in this process, historical conversation information also becomes a part of the context.

In recent years, conversational machine reading comprehension (CMRC) has become a new research hotspot in the NLP community, and there emerged many related datasets, such as CoQA [47], QuAC [65], DREAM [81] and ShARC [38].

7. Domain-Specific Datasets

such as science examinations, movies, clinical reports

8. MRC with Paraphrased Paragraph

Paragraph paraphrasing refers to rewriting or rephrasing a paragraph using different words, while still conveying the same messages as before. The MRC dataset with paraphrased paragraph has at least two versions of context which expresses the same meanings while there is little word overlap between the different versions of context. The task of paraphrased MRC requires the computer to answer questions about contexts. To answer these questions correctly, the computer needs to understand the true meaning of different versions of context. So far, we only find that the DuoRC [68] and Who-did-What [48] are datasets of this type.

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

EMNLP2020

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).

Collecting 21.7k multiple-choice problems with human-annotated answers for the National Licensed Pharmacist Examination in China.

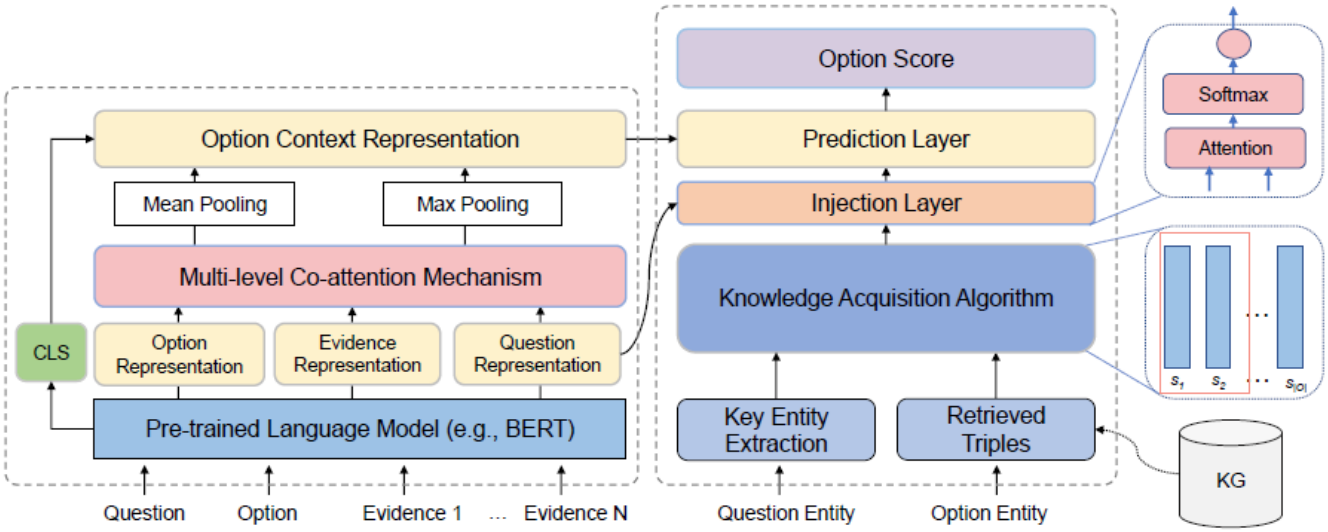


Figure 1: Overall architecture of the proposed KMQA, with multi-level co-attention reader (left) and the knowledge integration part (right) illustrated.

(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.

Multi-level Co-attention Reader

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.

Finetune

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 B.

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:

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

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

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

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

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

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

Knowledge Acquisition

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 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

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 g

Translate triple $s_j = (h_j, r_j, t_j) \in \mathcal{S}_O$ to general text p_j using 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 o in the ascending order

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:

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

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

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

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

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

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

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 $\tilde{\mathbf{T}}_F \in \mathbb{R}^{2h}$.

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

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

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* $\rightarrow$ *related diseases* $\rightarrow$ *cirrhosis* $\rightarrow$ *medical treatment* $\rightarrow$ *entecavir*) is a path for (*chronic hepatitis B*, *entecavir*).

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

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

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

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.

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.

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...