A Chinese Machine Reading Comprehension Dataset Automatic Generated Based on Knowledge Graph

Hanyu Zhao[1], **Yuan Sha** [1,✉], **Jiahong Leng**[1], **Xiang Pan**[2],
Zhao Xue [1]、 **Quanyue Ma** [1] and **Yangxiao Liang** [1]
[1] Beijing Academy of Artificial Intelligence, Beijing, China
[2] New York University, New York, America
{hyzhao, yuansha, jhleng, xuezhao, maqy, yxliang}@baai.ac.cn
xiangpan@nyu.edu

Abstract

Machine reading comprehension (MRC) is a typical natural language processing (NLP) task and has developed rapidly in the last few years. Various reading comprehension datasets have been built to support MRC studies. However, large-scale and high-quality datasets are rare due to the high complexity and huge workforce cost of making such a dataset. Besides, most reading comprehension datasets are in English, and Chinese datasets are insufficient. In this paper, we propose an automatic method for MRC dataset generation, and build the largest Chinese medical reading comprehension dataset presently named CMedRC. Our dataset contains 17k questions generated by our automatic method and some seed questions. We obtain the corresponding answers from a medical knowledge graph and manually check all of them. Finally, we test BiLSTM and BERT-based pre-trained language models (PLMs) on our dataset and propose a baseline for the following studies. Results show that the automatic MRC dataset generation method is considerable for future model improvements.

1 Introduction

Medical care is closely related to people's lives and helps people keep healthy in several different aspects, including disease prevention, medical examination, disease diagnosis, and treatment. Generally, various medical services are provided by those professionals who have specialized knowledge and rich experience. However, many countries now face severe medical personnel shortages, which means the demand for medical services dramatically exceeds the limit that professionals can supply. The rapidly developed artificial intelligence (AI) technology is a potential solution to the doctor shortage problem. AI medical experts are expected to offer various kinds of service for patients after learning enough knowledge from human experts and thus, reduce doctors' burden to a great extent.

One of the potential uses of AI experts is to provide medical consultation for patients. The technical nature of such kinds of applications is the automated question answering technology. Recently, the automatic question answering system is a significant development direction of the intelligent medical industry. Retrieval-based question answering and knowledge base question answering (KBQA) are two primary forms of present medical QA systems. Retrieval type system selects the candidate in the existing collection of QA pairs with the highest similarity to the user's input question. It returns the corresponding answer as the final response. As for KBQA, the system first extracts a topic entity in the user's question, then searches it in the knowledge base and returns the most suitable neighbor of a topic entity as the final answer. However, the above two methods both have poor performances when answering new questions that have not been appeared in the dataset or knowledge base of the QA system. Thus, they are not adaptive to the medical industry, where new things appear almost all the time. MRC provides a new approach to construct a QA system. It enables the machine to understand the meaning of contexts and answer related questions. The difference is that the MRC QA system answers questions based on its knowledge learned from a large volume of data rather than directly matching one from the dataset or knowledge base.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1066-1075, Hohhot, China, Augest 13 - 15, 2021.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Although MRC systems have many advantages in question answering tasks, they are not developed a lot recently due to the lack of large-scale and high-quality datasets. At present, most MRC datasets are generated and annotated by humans, which is time-consuming and non-objective sometimes. In the medical industry, some data annotation work can only be done by those annotators with specialized knowledge, which also means a high economic cost. In this case, it is helpful to generate and annotate datasets automatically. Here, we propose a knowledge-graph-based automatic method to generate Q&A datasets and construct a medical dataset. We will show the details of the method in section 3.

The main contributions of our work are as follows:

- We propose a reading comprehension dataset automatic generation method based on knowledge graph(KG) and pretrained language models(PLMs).
- By using this method, we release the largest Chinese Medical Reading Comprehension dataset.
- We propose several baseline models based on our medical dataset CMedRC, such as BERT, ERNIE, etc.

2 Related Work

2.1 English Machine Reading Comprehension Datasets

There are many reading comprehension datasets for Machine reading comprehension research. Though these datasets concentrate on different question types and refer to different corpora domains, most of them are in English. The following paragraph introduces several famous English datasets. CNN/Daily Mail (Hill et al., 2015) is a large one in the news, which contains approximately 1.4M fill-in questions. NewsQA (Trischler et al., 2016) is another corpus that focuses on the news field, but it is relatively small, and most questions are the span of words. RACE (Lai et al., 2017) collects around 870K multiple-choice questions that appeared in Chinese middle school students' English reading comprehension examination. SQuAD2.0 (Rajpurkar et al., 2016) and TriviaQA (Joshi et al., 2018) are collections of contents and question-answer pairs in Wikipedia. Besides the span of words, some questions in those two datasets are more complex, which need to be inferred by summarizing multiple sentences in corresponding contexts. CoQA (Reddy et al., 2018) gathers its data mainly from children's storybooks. It contains various types of questions, including the span of words, yes/no, and even some unanswerable questions whose answers cannot be obtained according to the given context.

2.2 Chinese Machine Reading Comprehension Datasets

As for the Chinese dataset, HFL-RC (Cui et al., 2016) and C3 (Sun et al., 2020) are the first Chinese cloze-style and free-form multiple-choice MRC datasets respectively. The former is collected from fairy tales and newspapers, while the latter is based on documents in Chinese-as-a-second-language examinations. In the field of law, CJRC (Cui et al.,) is a famous dataset that contains more than 10K documents published by the Supreme People' s Court of China. It contributes a lot to the research of judicial information retrieval and factor extraction. DuReader (He et al., 2017) is a large-scale and open-domain Chinese MRC dataset, which gathers 200K questions from Baidu Search and Baidu Zhidao with manually generated answers.

2.3 Machine Reading Comprehension Models

Considering the characteristics of Chinese MRC tasks, various models have been proposed for solving them. By using attention mechanism (Kadlec et al., 2016), AoA (attention over attention) (Cui et al., 2017), Gated Attention (Dhingra et al., 2017), Hierarchical Attention (Wang et al., 2018) and ConvAtt (Yu et al., 2018), some models have achieved considerable performance increase. From MRC task perspective, there are also various specific models that have been

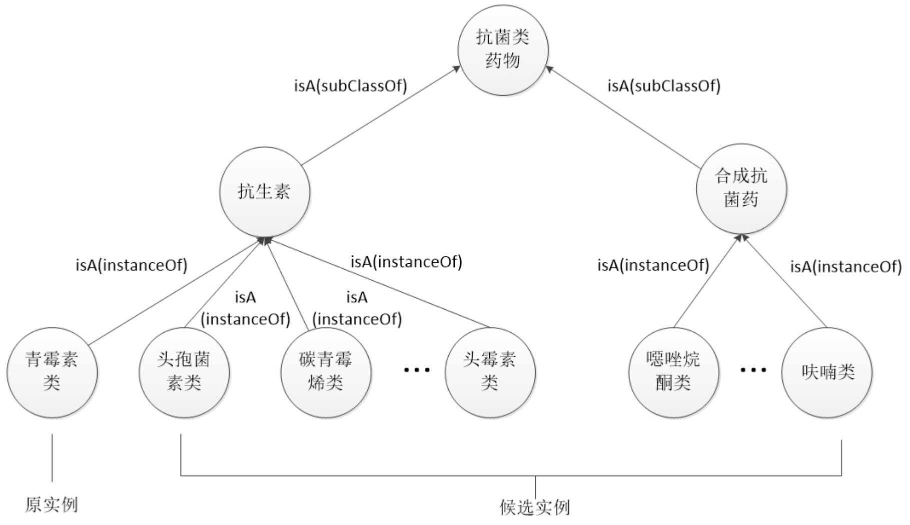

Figure 1: Example of entity replacement using conceptual knowledge graph

proposed rencently (Wang and Jiang, 2016a; Xiong et al., 2016; Liu et al., 2017; Wang et al., 2017; Hu et al., 2018).

However, most of the proposed methods and their corresponding datasets rely on humans. There are several disadvantages in efficiency and accuracy. Those methods involve resource-intensive processes such as the pre-defining question template and the linkage between question and their paragraph in documents. Those proposed fully automated annotation methods still require humans for evaluation and dataset cleaning. For those professional areas such as medical, the annotation and evaluation require professional knowledge. Those proposed methods are not usable. Furthermore, present methods are mostly straightforward and do not involve multiple-step reasoning and inferences.

3 CMedRC: A Chinese Medical MRC dataset

CMedRC dataset is generated based on a chinese medical knowledge graph[1]. Knowledge graph is a summary and abstraction of existing human knowledge. With the upper and lower relationship in the knowledge graph, new questions can be generated by conducting replacement keywords in seed questions. During the generation process, SimBert is applied to create synonymous sentences of questions to guarantee the corpus' richness. Answers to questions are obtained from a knowledge graph, and corresponding documents that contain answers are crawled from Baidu Baike and Wikipedia. In the following subsections, we will introduce the dataset construction in detail.

3.1 Question and Answer Generation

For a given cypher (seed question), we can extract topic entities and relations from the sentence. Here, a relation means a piece of knowledge related to a topic entity in our medical knowledge graph. Thus, we can replace entities and/or relations in those seed questions to generate new questions automatically.

3.1.1 Entity Replacement

We conduct entity replacing based on a medical conceptual knowledge graph. For each topic entity in our knowledge graph, it has a hypernym and a class. Those topic entities that belong to the same hypernym or other hypernyms under the same class form a candidate set. By selecting a suitable entity from the candidate set for substitution, a new question can be generated. For example, in Figure 1, the seed question is "青霉素类药品的作用是什么？(What is the effect of

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

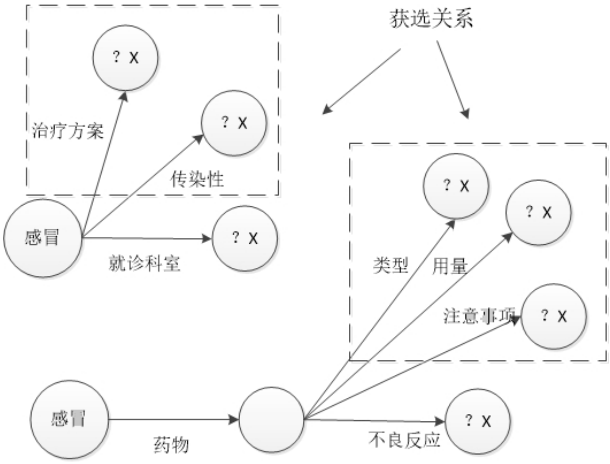

Figure 2: Examples of relation replacement using conceptual knowledge graph

penicillin?)”. The original entity of cypher is “青霉素类 (penicillin)”, which is an instance of its hypernym “抗生素 (antibiotics)”. The original hypernym is a subclass of class “抗菌类药物 (antibacterial drug)”. When conducting replacement, the candidate set includes other entities of original hypernym “抗生素 (antibiotics)” and entities of the hypernym that belongs to the same class “抗菌类药物 (antibacterial drug)”. After entity substitution, one example of the new question is “头孢菌素类药品的作用是什么？(What is the effect of cephalosporin?)”.

3.1.2 Relation Replacement

Similarly, a conceptual knowledge graph is used for relation replacement. After extracting a topic entity and its relation from a cypher, we can choose other relations of that entity in the knowledge graph to replace the original one. Thus, a new question can be generated. If there is a multi-step relation in the seed question, we substitute for the last step relation. Figure 2 shows relation replacement examples of the above two cases. For single-step relation case, the seed question is “感冒的就诊科室是？(Which clinic department should we go when having a cold?)”. We can extract the entity “感冒 (cold)” and its relation “就诊科室 (clinic department)” from cypher. To make a substitution, we randomly choose one relation from the candidate set: the rectangle area enclosed with a dotted line. One example of a generated question is “感冒的治疗方案是？(What is the treatment method for cold?)”. As for the multi-step case, our cypher is “感冒药物双黄连的副作用是？(What are the side effects of cold medicine ShuangHuangLian?)”. Here, we have “感冒 (cold)”, and “双黄连 (ShuangHuangLian)” two entities, their corresponding relations are “药物 (medicine)” and “副作用 (side effect)” respectively. When conducting replacement, only the last step relation “副作用 (side effect)” is substituted by a new relation of entity “双黄连 (ShuangHuangLian)”. Thus, a new question can be “感冒药物双黄连的用量是？(What is the dosage of cold medicine ShuangHuangLian?)”.

3.2 Synonymous Sentences Generation

To guarantee question corpus diversity, we use SimBERT to produce a synonymous sentence for those new questions after entity and/or relation replacement. SimBERT is a model trained by many synonymous sentence pairs and can predict labels according to the cosine similarity between contextual token embeddings. The structure of SimBERT is shown in Figure 3. In our method, we generate ten synonymous sentences for each new question by SimBERT and randomly select one from those sentences to be the final version of the question. [2] For example, we use a previously generated question “感冒的治疗方案是？(What is the treatment method for cold?)” as the input, the modified new question can be “感冒的治疗方法都有哪些？”(What treatment method can be applied for cold?).

[2] https://github.com/ZhuiyiTechnology/simbert

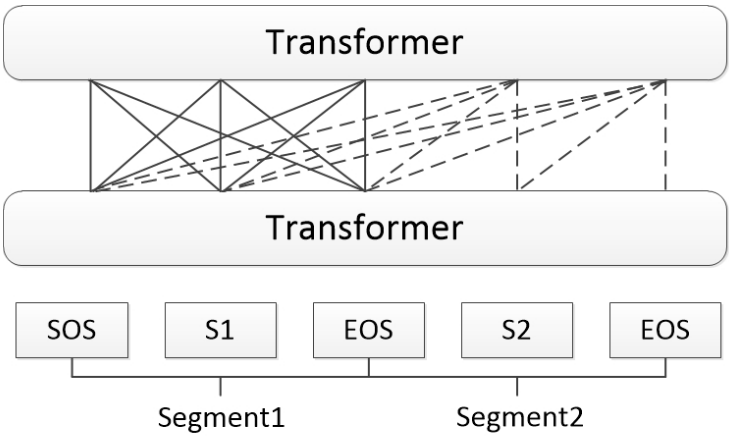

Figure 3: The structure of SimBERT model

3.3 Answer Matching and Document Linkage

We need to match their answers for newly-generated questions and collect those documents containing QA pairs. First, we obtain answers directly from our medical conceptual knowledge graph. We exclude those questions without answers in the dataset. We can crawl the related paragraph from web pages such as Baidu Baike and Wikipedia for those well-matched QA pairs. Suppose the answer to a question cannot match the crawled paragraph exactly, we calculate the Levenshtein Distance between answer and each sliding window of that paragraph. If there is at least one distance larger than 0.7, we record corresponding window with largest distance and submit it for manual check. Otherwise, we delete that question from our dataset.

3.4 Dataset Cleaning

After answer and document matching, we can obtain a set of triples (Q, A, P). To guarantee the quality of the corpus, we conduct data cleaning before adding them to the dataset. Specifically, we check the generated questions and remove those with faulty or misleading wording. Even though our method is more labor-efficient, evaluation is more knowledge-reliable than the human generation.

Table 1: Comparison of CMedRC with existing reading comprehension datasets

DataSets	Language	Domain	Answer Type
RACE	ENG	English Exam	Multi choise
SQuAD	ENG	Wikipe	Span of word, Unanswerable
NewsQA	ENG	CNN	Span of word
CNN	ENG	News	Fill in entity
HFRC	CHN	Fairy/News	Fill in word
CMRC	CHN	Wikipe	Span of word
CJRC	CHN	Law	Span of word, Yes/No, Unanswerable
CMedRC	CHN	Medical	Span of word, Yes/No, Unanswerable

Table 2: Dataset statistics of CMedRC

	Docments	Q&A pairs	Span of word	Yes/No	Unanswerable
train	6000	13000	10002	2255	743
test	2000	4000	3261	693	46

3.5 Data Statistics

Our dataset consists of more than 8k medical documents and 17k QA pairs. It is the largest medical reading comprehension dataset. Table 1 compares our dataset with some other datasets mentioned in section 2, mainly considering three dimensions: language, domain, and answer type. Table 1 shows that CMedRC contains more problematic types of questions, thus can evaluate the ability of algorithms from more perspectives.

In order to carry out the research better, we divide the dataset into training set and test set, and the data distribution is shown in Table 2. Additionally, Table 3 gives an example of a piece of data.

Table 3: Example of data in CMedRC

Content	鼠疫（plague）是由鼠疫耶尔森菌感染引起的烈性传染病，属国际检疫传染病，也是我国法定传染病中的甲类传染病，在法定传染病中位居第一位。鼠疫为自然疫源性传染病，主要在啮齿类动物间流行，鼠、旱獭等为鼠疫耶尔森菌的自然宿主。鼠蚤为传播媒介。临床表现为发热、毒血症症状、淋巴结肿大、肺炎、出血。本病传染性强，病死率高。鼠疫在世界历史上曾有多次大流行，我国在解放前也曾发生多次流行，目前已大幅减少，但在我国西部、西北部仍有散发病例发生。肺鼠疫可以挂传染科。(Plague is a virulent infectious disease caused by Yersinia pestis. It is an international quarantinable infectious disease and a Class A infectious disease in China. Among all Chinese statutory infectious diseases, it ranks in the first place. Plague is a natural infectious disease, which is mainly prevalent among rodents. Mice and marmots are the natural hosts of Yersinia pestis. Rat fleas are the medium of transmission. The clinical manifestations are fever, toxemia, lymph node enlargement, pneumonia, and hemorrhage. The disease is highly infectious and has a high fatality rate. Plague has had many epidemics in world history, and many epidemics occurred in China before liberation. At present, the epidemic has been dramatically reduced, but sporadic cases still appear in the west and northwest of China. Pneumonic plague can be linked to the department of infection in the hospital.)
QA Pairs	“id”: ”36_0”, “question”: “肺鼠疫有哪些临床上的表现?” (What are the clinical manifestations of pneumonic plague?), “answers”: “发热、毒血症症状、淋巴结肿大、肺炎、出血” (fever, symptoms of toxemia, lymph node enlargement, pneumonia, bleeding), “id”: ”36_1”, “question”: ” 肺鼠疫的病因都有什么?” (What are the causes of pneumonic plague?), “answers”: ” 鼠疫耶尔森菌感染” (the infection of Yersinia pestis), “id”: ”36_2”, “question”: ” 肺鼠疫通过什么途径传播?” (How does pneumonic plague spread?), “answers”: ” 鼠蚤为传播媒介” (Rat fleas are the medium of transmission.), “id”: ”36_3”, “question”: ” 肺鼠疫翻译成英文是?” (What is the Engish translation of ” 肺鼠疫”?), “answers”: ”plague”,

4 Experiment

4.1 Evaluation Metric

In this paper, we adopt two different evaluation metrics to measure the performance of the question-answering task. The metrics are Exaction Position Match and F1-Score.

4.1.1 Extraction Position Match

For the extracted machine reading comprehension(MRC) task, we measure the position match using the exact match(EM). When the extracted beginning position and the end position are the same, the score is one. Otherwise, the score is zero. Such a measure can be regarded as a strict performance indicator.

4.1.2 F1-score

For MRC tasks, if the starting or ending position is close but not the same, the answer is still valuable sometimes. Therefore, we also use F1-score, which can measure the characters' overlap between the prediction and ground truth. Specifically, the formula of F1-score is as follow:

$$Pre = \frac{TP}{LP}$$

$$Rec = \frac{TP}{LG}$$

$$F1 = \frac{2 * Pre * Rec}{Pre + Rec}$$

Here, TP is the overlap length between prediction and ground truth, LP and LG are the length of prediction and ground truth, respectively. Moreover, if there are successive non-Chinese tokens, they are counted as one unit length rather than segmented into characters.

4.2 Baselines

For MRC tasks, there is usually a sequence model for contextual encoding and a task-specific layer for index prediction. We adopt traditional BiLSTM as our baseline model. For BERT-based PLMs, we test the normal BERT-based model and BERT-WWM (Whole Word Masking), which considers the Chines language characteristic. We also test the ERNIE (Sun et al., 2019) as the SOTA model for our evaluation. Finally, we provide estimated human performance as the human performance for reference.

4.2.1 Sequential Model

Traditional MRC model uses BiLSTM as the input sequence encoding layer and the final linear layer for the answer index prediction (Wang and Jiang, 2016b). BiLSTM considering two directions of information flow, and it is lightweight and classic to be the baseline model.

4.2.2 BERT

As the representation of pre-training language models(PLMs), BERT is pre-trained on a large text corpus (like Wikipedia). Thus it can capture the general language features and be used as a general contextual embedding layer for downstream tasks. In MRC, BERT is used as the contextual encoding layer to get the sentence embedding. There are various versions of BERT, and we chose the word piece MLM version and whole word masking version (WWM) to be the Bert-based models in our experiment.

4.2.3 ERNIE

ERNIE (Enhanced Language Representation with Informative Entities) can learn abundant information from broad knowledge resources, including knowledge graphs. Thus it is more potent in knowledge-oriented downstream tasks, such as MRC. We use it as a SOTA model in the experiment.

4.3 Evaluation

Experimental results on the test set are shown in Table 4. From this table, it is obvious that the pre-trained language models(PLMs) such as BERT is about 20 percentage points better than BiLSTM in EM and F1, which indicates that PLMs on large corpora can help it learn general language representation. Fine-tuning based on PLMs can improve downstream tasks' effects and avoid training models of downstream tasks from scratch. Besides, we also experiment with two improved BERT models—BERT_WWM and ERNIE. The table shows that BERT_WWM can be one percentage point higher than BERT_Base in EM by using Whole Word Masking. And ERNIE can be one percentage point higher in F1 by introducing knowledge into BERT. Finally, we also evaluate the results of Human Performance, which is approximately seven percentage points higher than PLMs. It implies that the automatically generated data can markedly improve models in future research.

Table 4: Experiment results

Model	EM	F1
Human	0.861	0.912
BiLSTM	0.532	0.623
BERT_Base	0.737	0.832
BERT_WWM	0.746	0.837
ERNIE	0.740	0.841

5 Conclusion and Further Research

5.1 Conclusion

We present an automatic generation method for machine reading comprehension dataset. The method merely relies on an existing conceptual knowledge graph and a small number of seed questions to automatically construct a Chinese reading comprehension dataset. According to this approach, we construct the largest Chinese medical reading comprehension dataset presently, which contains more than 17k QA pairs and 8k documents. In comparison with those human-annotated datasets, our dataset's construction cost is reduced to a great extent. Next, we propose baseline models based on our dataset for BiLSTM and BERT-based PLMs. After comparing with human answers, the results show that those models have room for further improvement. Although our method is practiced in the medical area, such a method can be easily adopted and generalized into other areas.

5.2 Further Research

In our method, we only consider the one-step replacement. We can use similar methods for generated multi-hop reasoning reading comprehension datasets, which is a more challenging task for evaluating the MRC model and machine intelligence.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61806111, NSFC for Distinguished Young Scholar under Grant No. 61825602 and National Key R&D Program of China under Grant No. 2020AAA010520002.

References

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. page 7.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786. The COLING 2016 Organizing Committee.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602. Association for Computational Linguistics.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846. Association for Computational Linguistics.

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

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4099–4106. International Joint Conferences on Artificial Intelligence Organization, 7.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2018. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794. Association for Computational Linguistics.

Ting Liu, Yiming Cui, Qingyu Yin, Wei-Nan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 102–111. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

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

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:1904.09223 [cs]*, April. arXiv: 1904.09223.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Shuohang Wang and Jing Jiang. 2016a. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Shuohang Wang and Jing Jiang. 2016b. Machine Comprehension Using Match-LSTM and Answer Pointer. *arXiv:1608.07905 [cs]*, November. arXiv: 1608.07905.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198. Association for Computational Linguistics.

Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714. Association for Computational Linguistics.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.