Examination-Style Reading Comprehension with Neural augmented Retrieval

Yiqing Zhang[1,2,3], $HaiZhao^{1,2,3,*}$, $ZhuoshengZhang^{1,2,3}$
[1]$DepartmentofComputerScienceandEngineering, ShanghaiJiaoTongUniversity$
[2]$KeyLaboratoryofShanghaiEducationCommissionforIntelligentInteraction$
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]$MoEKeyLabofArtificialIntelligence, AIInstitute, ShanghaiJiaoTongUniversity$
zhangyiqing, zhangzs@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

***Abstract*—In this paper, we focus on an examination-style reading comprehension task which requires a multiple choice question solving but without a pre-given document that is supposed to contain direct evidences for answering the question. Unlike the common machine reading comprehension tasks, the concerned task requires a deep understanding into the detail-rich and semantically complex question. Such a reading comprehension task can be considered as a variant of early deep question-answering. We propose a hybrid solution to solve the problem. First, an attentive neural network to obtain the keywords in question. Then a retrieval based model is used to retrieve relative evidence in knowledge sources with the importance score of each word. The final choice is made by considering both question and evidence. Our experimental results show that our system gives state-of-the-art performance on Chinese benchmarks and shows its effectiveness on English dataset only using unstructured knowledge source.**

***Keywords*-MRC; retrieval; knowledge source;**

I. Introduction

[1] Machine reading comprehension (MRC) which requires computers to answer questions based on acquired knowledge, is regarded as the milestone for deep natural language understanding. A great amount of datasets have been released, including cloze-style CNN/Daily Mail [1], multiply choice type QuaRel [2] and user-query type SQuAD [3], MS MARCO [4]. Most MRC datasets provide one or several corresponding documents which are commonly retrieved by search engines. However, MRC systems may not always find available pre-given document as the hypothesis like all previous cloze-style or user-query types tasks, they have to be both capable of discovering the relationship between question and answer and retrieve critical evidence even without the pre-given documents.

In this work, we thus consider such an examination-style MRC task, which releases the inconvenient requirement about the pre-given document commonly in other types of MRC tasks. Undoubtedly, training machines to take human examinations is more challenging in which answers have to be given on a basis without a standard relevant document that is supposed to contain cues for answering the question.

Why the concerned examination-style MRC task is more challenging than those common MRC types is that evidence knowledge for answering the questions should be found by question-answering (QA) system itself, which makes the solution model much more complicated. The concerned MRC task can be also regarded as a type of deep QA [5], [6] one considering that no standard pre-given document is given for question extraction.

Table I
AN EXAMPLE OF QUESTION TYPE WE FOCUS ON. THERE IS NO CORRESPONDING DOCUMENT PRE-GIVEN.

Question: Some Western scholars think 'Half of all the inventions and discoveries that the modern world depends on may be from China'. The science and technology achievement of Song Dynasty that was introduced to Europe and had a profound impact on the modern world is ?
Candidate answers:
A. Chang's seismograph B. Papermaking skill
C. Typography D. Compass
Correct answer: D

However, in detail, this paper considers solving a subtype of the concerned MRC task, *multiple choice questions*. It slightly differs from common deep QA task that answer candidates have been given in a selectable list [7]. Table I shows a typical multiple choice question in history exam which consists of a question with some background information and four candidate answers. Our system follows the way how human solve this problem. First, we train an attentive neural network as the basic MRC model. During this step, the neural network can learn key words for searching evidence knowledge by attention mechanism, because the important words for choosing the right answer is always the important words for searching the relevant evidence knowledge. Then the importance of the key words and the question are taken into consideration when the retrieval model searches evidence knowledge from knowledge source. Finally, both question and evidence are used to determine the right answer. Experimental results show that, with the neural model indicating word importance, the retrieval model can find more accurate evidence to help judge which candidate is the correct answer. In Chinese dataset the relevant evidence is not tagged, the effectiveness of word importance can only be reflected by the final accuracy of choosing the right answer, so we use an English multiple choice question-answering dataset [8]. In this dataset, paragraph which contains information to answer the certain question is

[1] * Corresponding author. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100) and Key Projects of National Natural Science Foundation of China (No. U1836222 and No. 61733011).

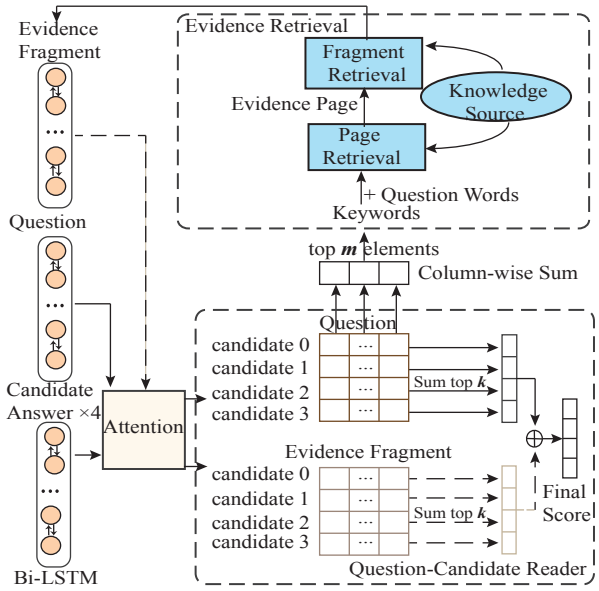

Figure 1. An overview of our system. The data flow denoted by dotted arrow appears after Evidence Retrieval module generates output.

indeed given, we tag the given paragraph as the best paragraph and all the paragraphs are collected to be the external knowledge source. After that we can directly know the effectiveness of the neural model indicating word importance to the retrieval model. The code of our system is available here[2].

Our contribution in this paper is three-folder:

- We formally focus on examination-style MRC task that the evidence document is not given and needed to be found by QA system itself.
- We use unstructured knowledge source which makes our system can be easily improved by using high-quality knowledge source in specific areas.
- We design an MRC system that combines the retrieval based and neural based method in a novel way for multiple choice question-answering.

II. Related Work

In recent years, an ambitious AI task that develops QA system to pass examinations in different levels and subjects has been paid a lot of attention with the development of deep learning [9]. The Project Aristo [10] devoted to having the computer pass Elementary School Science and Math exams. Another more ambitious project is the Todai Robot Project [11] which aims to let the computer take the university admission examination in Japan and reach the acceptance score of University of Tokyo.

During the development of MRC, different types of datasets have been released. The SQuAD [3] poses a question-answering mode over the pre-given document, in which the question is related to the document and the answer is a span in the document. MS-MARCO [4] provides several documents for each question and the question should be answered based on these documents. RACE [8] is from real-word examination with the same QA mode as SQuAD except for questions in multiple choice form. Anyway, it is generally accepted that providing a pre-given document as the above MRC task surely alleviates the difficulty of the task.

For a more challenging task without such a pre-given document, external knowledge source has to be taken into account. Knowledge base (KB) can be processed easily by computer, however, it has the inherent limitation of being hard to build. Researchers thus consider KB extension or enhancement. Berant [12] scaled up Freebase by learning from question-answer pairs. However, even the largest KB still suffers from the insufficiency, so unstructured resources like Wikipedia have been widely used in QA systems [13].

Recently, a few works considered using more diverse sources other than Wikipedia. For example, Sachan [14] used instructional materials to tackle science question. Zhang [15] studied answering questions in clinical medicine using knowledge in a large-scale document collection. Guo [16] solved multiple choice questions in history exam by using textbooks resources and relevant information collected from Baidu Encyclopedia. Zhang [17] recently adopted strong machine learning model like oneshot learning to overcome the data deficiency difficulty in examination-style MRC.

Different from all the mentioned work that either seeks specific support data sources or strengthens the model itself, this work adopts a hybrid mechanism by effectively integrating retrieval model and neural models instead for the concerned examination-style MRC task. We regard the DrQA system [18] is the most similar one to this work for both using a retrieval module inside. However, this work significantly differs from [18] from the following factors. First, this work is about strict examination-style MRC task, while DrQA actually still works for a SQuAD-like span MRC task but simply assuming that the pre-given document is missing. Second, this work uses neural module to feed retrieval module for the query and the final solution to the question is determined jointly by both retrieval and neural modules, while DrQA uses a simple retrieval module to return a Wikipedia page as the pre-given document, and then span MRC system turns on for all the rest processing.

III. System

Each multiple choice question can be regarded as a triple (Q, Cs, A) where Q is the question, Cs is a set of four candidate answers and A is the correct answer. Figure 1 is an overview of our system.

A. Question-Candidate Reader

Our Question-Candidate Reader follows the backbone of the AOA model [19] which has proved its good performance. We keep the attention mechanism of it and change its embedding layer and answer prediction layer in order to make the model adapt to our concerned task. The following gives more model details.

[2]https://github.com/Yiqingss/Gaokao-history-QA

Contextual Embedding For English dataset, we use GloVe [20] word embedding. For Chinese datsets, we use word2vec [21] to pretrain word embeddings on Chinese Wikipedia. Then we use 2 Bi-LSTM [22] to get contextual representations $h(x)$ of the question and candidate answers.

Attention The attention score follows the AOA model [19] which is given by

$$\begin{aligned} M(i,j) &= h_{question}(i)^T \cdot h_{candidate}(j) \\ \alpha = softmax_{column}(M), &\beta = softmax_{row}(M) \\ \beta_{avg} &= \frac{1}{|Q|}\sum_{t=1}^{|Q|}\beta(t:) \\ S &= \alpha \cdot \beta_{avg} \end{aligned} \tag{1}$$

where $M(i,j) \in \mathbb{R}^{|Q|*|C|}$ is the matching matrix of each question-candidate pair, as each question has four candidate answers, we get four matrices for each question. The value of i-th row and j-th column of M is filled by the dot product of i-th word in question and j-th word in candidate answer. Then a column-wise and a row-wise softmax function is applied to the matching matrix M to get its question-level attention ($\alpha \in \mathbb{R}^{|Q|*|C|}$) and candidate-level attention ($\beta \in \mathbb{R}^{|Q|*|C|}$) respectively. β_{avg} is the result of column-wise average of β, $\beta(t:)$ is t-th row vector of β. $S \in \mathbb{R}^{|Q|}$ is the attention-over-attention score for each question-candidate pair.

Answer Prediction As each question has four candidate answers, each question will get four S. We denoted the attention score S calculated by Eq. (1) with the i-th candidate answer as $S_i, i \in \{0,1,2,3\}$. Then we compute a final score for each question-candidate pair by summing the value of elements in S_i with high score.

$$P(C_i|Q) = \sum top(k; S_i), i \in \{0,1,2,3\} \tag{2}$$

where $top(k; S_i)$ means the k elements with the highest scores in S_i, $P(C_i|Q)$ is the sum of the value of these elements. The predicted answer is candidate answer with the highest $P(C_i|Q)$ score. K is set to 4 through empirical tuning. The words corresponding to the high-score elements make up the set of importance words which will be used in Evidence Retriever.

Training Giving a training corpus of multiple choice question: $U = \{(Q^{(i)}, Cs^{(i)}, A^{(i)})\}_{i=1}^{|U|}$, the training objective of our model is to minimize the cross entropy of the training data.

$$A_j^{(i)} = \begin{cases} 1, & Cs_j^{(i)} = A^{(i)} \\ 0, & Cs_j^{(i)} \neq A^{(i)} \end{cases}$$

$$\begin{aligned} \mathcal{L} = -\sum_{i=1}^{|U|}\sum_{j=0}^{3} &[A_j^{(i)} \cdot log(\tilde{A}_j^{(i)}) \\ &+ (1 - A_j^{(i)}) \cdot log(1 - \tilde{A}_j^{(i)})] \end{aligned}$$

where the $\tilde{A}_j$ is the predict likelihood of the j-th candidate answer computed by applying softmax on four candidates' prediction scores $P(C_j|Q)$.

B. Evidence Retriever

Given a question Q consisting of l tokens $\{q_1, q_2, \ldots, q_l\}$ and a knowledge source page set P with n pages $\{p_1, p_2, \ldots, p_n\}$. The Evidence Retriever returns a fragment of a knowledge source, because we found that in most of the unstructured knowledge sources, even in one page there are still redundant information.

Page Retrieval We first calculate each word's TF-IDF score $S_{tf-idf}(w)$ in P:

$$\begin{aligned} S_{tf-idf}(w) = &\log[tf(w, p_i) + 1] \\ &* \log \frac{vocab - idf(w) + 0.5}{idf(w) + 0.5} \end{aligned}$$

where $tf(w, p_i)$ is the number of times that word w occurs in p_i, $idf(w)$ is the number of knowledge source pages containing word w and $vocab$ is the number of unique words in all knowledge source pages.

We then compute each word's importance score and choose m words with highest scores as keywords. We use a softmax function to get probability distributions of the keywords. The set of keywords is denoted as K consisting of m tokens $\{k_1, k_2, \ldots, k_m\}$.

$$S_{imp}(j) = \sum_{i=0}^{3} S_{ij},\ w_j \in Q \tag{3}$$

$$P(k_i) = \frac{e^{S_{key}(k_i)}}{\sum_{j=1}^{m} e^{S_{key}(k_j)}},\ k_i \in K$$

$$weight(w) = \begin{cases} \log_{0.9}(1 - P(w)),\ w \in K \\ 1,\ w \notin K \end{cases} \tag{4}$$

where S_{ij} is attention score of the j-th word in question computed with the i-th candidate. $S_{imp}(j)$ is the importance score of the j-th word in question which is the sum of j-th word's attention scores. $S_{key}(k_i)$ is the importance score of keyword k_i and its value is computed by Eq. (3) . $weight(w)$ is the importance degree.

Finally, we compute the relevant score of every paragraph by

$$\begin{aligned} S_{para}(P, Q) = \sum_{w \in \{P \cap Q\}} &[a * S_{tf-idf}(w) \\ &* weight(w)] + |P \cap Q| \end{aligned}$$

$|Q \cap P|$ is the number of words both in Q and P, and a is a parameter to be tuned. We choose the page which has the highest $S_{para}(P, Q)$. Then this page will be processed by the following module.

Fragment Retrieval: In this module, we use a sliding window to find the most relevant fragment. Relevant score of each fragment selected by the sliding window is computed by

$$S_{frag}(F, Q) = \sum_{w \in \{F \cap Q\}} weight(w)$$

where F is a continuous sequence with a fixed length, $weight$ is computed by Eq. (4). The final result is the fragment with the highest score.

Table II
THE STATISTICS OF THE GAOKAO DATASETS

	Gaokao-577 (Test)	Gaokao-744 (Test)	TIKU (Train)
Count	577	744	53709

Table III
THE STATISTICS OF THE MODIFIED RACE

	Dev middle	Dev high	Test middle	Test high
#passage set	1,436	3,451	1,436	3,498

C. Final Prediction

The final prediction combines the results of Question-Candidate Reader and Evidence Retriever. To use the fragment returned by Evidence Retriever, we replace the Q in (Q, Cs, A) mentioned before with F, which is the fragment of Q found by Evidence Retriever. Then we get the triple (F, Cs, A). The final prediction score is computed by

$$P(C_i|Q, F) = P(C_i|Q) + b \cdot P(C_i|F)$$

where $P(C_i|Q)$ and $P(C_i|F)$ is calculated by Eq. (2) with triple (Q, Cs, A) and (F, Cs, A) respectively and b is a hyperparameter to be tuned.

IV. EXPERIMENTS

A. Dataset

We evaluate the proposed model in both Chinese and English datasets.

Chinese: Gaokao dataset We adopt two published datasets whose question-candidates pairs are collected from Gaokao (Chinese College Entrance Examinations) The first one [3] is published by Cheng [23], containing 577 multiple choice questions. This dataset is referred to Gaokao-577. The second one is published by Guo [16], containing 2 question sets. One is 744 questions from Gaokao, and the other is 53,709 practice questions[4] from TIKU . We refer these two question sets to Gaokao-744 and TIKU, respectively. All the questions in TIKU are used as the training data. Each question in Gaokao-744 has been manually divided into entity questions (EQs) and sentence questions (SQs). Entity questions are those whose candidates are all entities, and sentence questions are those whose candidates are all sentences. The statistics of these datasets are listed in Table II. During our experiments, we found that most of the question-candidates pairs in Gaokao-744 appears in TIKU and some question-candidates pairs are repeated several times. So we removed TIKU's questions of which 90% words are the same as a certain question in Gaokao-744, as a result, about 1,000 question-candidates pairs are removed and 10% of the rest questions are used as development data.

[3]http://ws.nju.edu.cn/gaokao/ijcai-16/GaokaoHistory577.xml
[4]https://github.com/IACASNLPIR/GKHMC/tree/master/data

The external knowledge resource we used as knowledge source for Gaokao-577 and Gaokao-744 is the contents of Chinese Wikipedia pages (dump on 2017/11/14) with a total of 299,383 pages. The latest version for Chinese Wikipedia pages is available here[5].

English: RACE RACE [8] (Reading Comprehension Dataset From Examinations) consists of 27,933 passages and 97,867 questions collected from English exams for middle and high school Chinese students. RACE is divided into two subsets RACE-M (from middle school exams) and RACE-H (from high school exams). Each question in RACE is followed with 4 candidate answers and only one of them is correct.

We choose RACE because it is a multiple choice style question answering dataset which is similar to the Gaokao datasets. The main difference between RACE and Gaokao multiple choice questions is that each question-candidates pair in RACE has a passage which contains the evidence for choosing the correct answer. This difference happens to cover the shortage of the Gaokao datasets that they do not have tagged evidence to directly evaluate the retrieval module's performance. We change the dataset to fit it to our system. We collect all the passages to build a knowledge source and the pre-given passage for a certain question is tagged as the best passage for the question, so that we can directly evaluate the retrieval module's performance. Table III lists the statistics of RACE dataset with our modification. The *#passage set* means how many passages a knowledge source contains in a sub-dataset

B. Reader Evaluation

Settings We use 1-layer bidirectional LSTM with 400 hidden units for both question and candidate answers encoding. For Gaokao-577 and Gaokao-744, We use pre-trained word embeddings by $word2vec$ [24] from 903M Wikipedia data in Chinese, and for RACE, we use GloVE [20]. The dimensions of embedding for all the datasets is 300. Our model is implemented using Tensorflow. ADAM optimizer is adopted for weight updating with a declining learning rate from 0.001 to 0.00005 during training and dropout rate of 0.9. A gradient clipping threshold [25] is applied to avoid gradient explosions. All the training samples are divided into batches of 32 samples each. For Chinese datasets, sentences need to be segmented into words firstly [26].

Effectiveness of top *k* As the answer prediction of Question-Candidate Reader is up to the sum of the k largest element in Eq. (1), the value of k must have a great impact to the final result. We test the accuracy of the Question-Candidate Reader with k varying from 1 to 6 on both Gaokao-577, Gaokao-744, and the result is showed in Figure 2, from which, Question-Candidate reader achieves best performance on Gaokao-577 when k = 4, on Gaokao-744 when k = 5.

Figure 2. Test accuracy against the value of K.

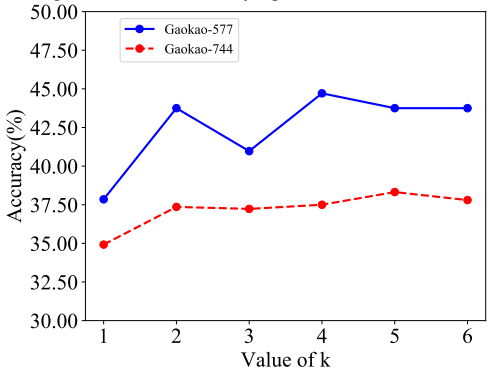

Table IV
ACCURACIES ON GAOKAO-577 (%)

	Gaokao-577
Random	25.00
[23]	33.79
Our system (Q)	44.71
Our system (F)	30.03
Our system (Q+F)	**45.41**

Table V
ACCURACIES ON GAOKAO-744 (%)

Count	entity question 160	sentence question 584
[16]	45.63	45.72
Our system	31.25	39.04

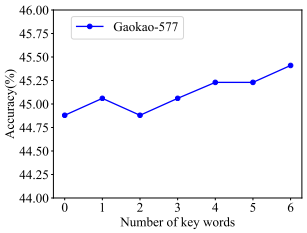

Figure 4. Test accuracy against the number of key words on Gaokao-577

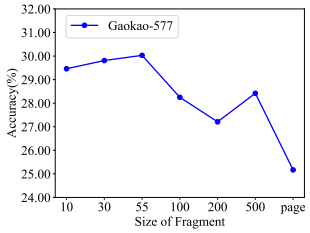

Figure 5. Test accuracy of different sizes of fragment on Gaokao-577. Fragment size is measured by the number of Chinese characters inside and 'page' indicates the entire page used as the evidence fragment.

C. System Evaluation

We examine the performance of our system with question only, evidence fragment only and combination of question and evidence fragment on Gaokao-577. The evaluation on Gaokao-744 is absent because the model in [16] used additional textbooks and resources other than Wikipedia as the external knowledge source. As the Wikipedia provides a great deal of informative clues, it makes incomparable results.

As we can see in Table IV, our system outperforms those systems that use Wikipedia as unique external knowledge on Gaokao datasets by a large margin, where 11.62% absolute improvements over [23] in Gaokao-577. DrQA does not perform well here, which is due to the relatively much more poor quality of Chinese Wikipedia pages. It can be also found that the system can achieve a relatively high accuracy only taking the question into consideration. The accuracy (30.03%) is low when we only use the evidence fragment on Gaokao datasets.

We also examine our best-performing system on Gaokao-744, in order to figure out our system performance on EQs and SQs, and the result of it is showed in Table V. We can see that our system performs better on SQs than EQs which proves that our system has a better ability to detect the relationships between sentences.

D. Evidence Retriever Evaluation

Effectiveness of key words and Fragment

To have a thorough investigation in the effectiveness of key words and their weighted scores given by Question-Candidate Reader, we first get the evidence fragment with n key words, $n \in \{0, 1, 2, 3, 4, 5, 6\}$, and then compute the accuracy of Question-Candidate reader only with the fragment. $n = 0$ means that we do not use the attention result of Question-Candidate Reader, and every words' weight is 1. n = 1 means only one key word is taken and its weight will be set to 15, as Eq. (4) can not calculate the weight at this situation. The results on Gaokao-577 are showed in Figures 4.

We see that the highest accuracy appears when n = 6 and the accuracy is low when n = 0, which proves that searching evidence without considering the importance of words performs poorly.

Different from the previous work [23], we do not use the whole page as evidence information to judge the possibility of a candidate answer. Instead, we use a slide window to pick a continuous sequence with a fixed length in the found page. To investigate whether this method works, we examine the best-performing system with the different fragment sizes, and the result is showed in Figures 5. The performance is poor when the fragment size is page size on Gaokao-577, this shows that using the entire page cannot bring about a better performance because of the redundant information in the entire page. Our system actually achieves the best performance when the fragment size is set to 55 on Gaokao-577.

Accuracy of retrieval results on modified RACE

After modifying the RACE dataset, we can directly figure out the accuracy of retrieval module. We build the TF-IDF index with the passages in the taining data of RACE and the best result is achieved when the number of keywords is set to 5. As is showed in Table VI, the accuracy increases after adding the keywords' weight. During our experiment we found that the number of keywords do not make much influence on the accuracy, it is mainly because the questions of RACE are short and usually do not contain information to choose the right answer, which makes the neural model hard to learn the attention weight for each question.

[5]https://dumps.wikimedia.org/zhwiki/latest/

Table VI
RETRIEVAL ACCURACY ON RACE (%)

	dev middle	dev high	test middle	test high
TF-IDF	0.533	0.390	0.487	0.406
TF-IDF + keywords	0.544	0.422	0.501	0.429

V. CONCLUSION

This work focuses on a challenging question answering type in real-word examination which does not have a pre-given document for answering the question. We proposed a system that uses external knowledge sources to help choose the right answer, and our system integrates neural model with attention and retrieval based model in a novel way. The effectiveness of the proposed approach has been verified on benchmark datasets. Only using Wikipedia as the unique knowledge source, our system outperforms previous state-of-the-art systems.

REFERENCES

[1] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *NIPS 2015*, 2015, pp. 1693–1701.

[2] O. Tafjord, P. Clark, M. Gardner, W. Yih, and A. Sabharwal, "Quarel: A dataset and models for answering questions about qualitative relationships," *CoRR*, vol. abs/1811.08048, 2018.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *EMNLP 2016*, 2016, pp. 2383–2392.

[4] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," in *NIPS 2016*, vol. 1773, 2016.

[5] D. A. Ferrucci, E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. M. Prager, N. Schlaefer, and C. A. Welty, "Building watson: An overview of the deepqa project," *AI Magazine*, vol. 31, no. 3, pp. 59–79, 2010.

[6] A. Lally, S. Bagchi, M. Barborak, D. W. Buchanan, J. Chu-Carroll, D. A. Ferrucci, M. R. Glass, A. Kalyanpur, E. T. Mueller, J. W. Murdock, S. Patwardhan, and J. M. Prager, "Watsonpaths: Scenario-based question answering and inference over unstructured information," *AI Magazine*, vol. 38, no. 2, pp. 59–76, 2017.

[7] S. Zhang, H. Zhao, Y. Wu, Z. Zhang, X. Zhou, and X. Zhou, "Dual co-matching network for multi-choice reading comprehension," *CoRR*, vol. abs/1901.09381, 2019. [Online]. Available: http://arxiv.org/abs/1901.09381

[8] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. H. Hovy, "RACE: large-scale reading comprehension dataset from examinations," in *EMNLP 2017*, 2017, pp. 785–794.

[9] S. He, Z. Li, H. Zhao, and H. Bai, "Syntax for semantic role labeling, to be, or not to be," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 2018, pp. 2061–2071.

[10] P. Clark, "Elementary school science and math tests as a driver for AI: Take the Aristo challenge!" in *IAAI 2015*, 2015, pp. 4019–4021.

[11] A. Fujita, A. Kameda, K. Ai, and Y. Miyao, "Overview of Todai Robot Project and evaluation framework of its NLP-based problem solving," in *IREC 2014*, 2014, pp. 2590–2597.

[12] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on Freebase from question-answer pairs," in *EMNLP 2013*, 01 2013, pp. 1533–1544.

[13] D. Ahn, V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, and S. Schlobach, "Using wikipedia at the TREC QA track," in *TREC 2004, November 16-19*, 2004.

[14] M. Sachan, A. Dubey, and E. P. Xing, "Science question answering using instructional materials," in *ACL 2016*, 2016, pp. 467–473.

[15] X. Zhang, J. Wu, Z. He, X. Liu, and Y. Su, "Medical exam question answering with large-scale reading comprehension," in *AAAI 2008*, 2018, pp. 5706–5713.

[16] S. Guo, X. Zeng, S. He, K. Liu, and J. Zhao, "Which is the effective way for gaokao: Information retrieval or neural networks?" in *EACL 2017*, 2017, pp. 111–120.

[17] Z. Zhang and H. Zhao, "One-shot learning for question-answering in gaokao history challenge," *CoRR*, vol. abs/1806.09105, 2018.

[18] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *ACL 2017*, 2017, pp. 1870–1879.

[19] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," in *ACL 2017*, 2017, pp. 593–602.

[20] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP 2014. October 25-29*, 2014, pp. 1532–1543.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[22] M. Sundermeyer, R. Schlter, and H. Ney, "LSTM neural networks for language modeling," in *ICA 2012*, 2012, pp. 194–197.

[23] G. Cheng, W. Zhu, Z. Wang, J. Chen, and Y. Qu, "Taking up the gaokao challenge: an information retrieval approach," in *IJCAI 2016*, 2016, pp. 2479–2485.

[24] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS 2013*, vol. 26, 2013, pp. 3111–3119.

[25] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *ICML 2013*, 2013, pp. III–1310.

[26] H. Zhao and C. Kit, "Integrating unsupervised and supervised word segmentation: The role of goodness measures," *Inf. Sci.*, vol. 181, no. 1, pp. 163–183, 2011.