SMART: A Stratified Machine Reading Test*

Jiarui Yao[1], Minxuan Feng[2], Haixia Feng[3], Zhiguo Wang[4], Yuchen Zhang[1], and Nianwen Xue[1]

[1] Brandeis University
{jryao,yuchenz,xuen}@brandeis.edu
[2] Nanjing Normal University
fennel_2006@163.com
[3] Ludong University
haixia872@163.com
[4] Amazon Web Services
zgw.tomorrow@gmail.com

Abstract. We present a Stratified MAchine Reading Test (SMART) data set for Chinese in which each question is assigned a "level" that reflects the type of reasoning that is needed to answer the question. This data set consists of close to 40K question-answer pairs and its stratified design allows machine reading researchers to quickly focus in on areas that present the most challenge for a machine comprehension system. We further establish a baseline for future research with BERT, and present results that show the levels we have designed correspond well with the level of difficulty that BERT experiences in answering these questions, as reflected by the lower accuracy for higher levels. We have also collected human answers to the questions in the test portion of this data set, and show that humans and the machine have different challenges when answering these questions. This means that even though the machine is approaching human-level performance on this task, humans and the machine perform this task with very different mechanisms.

1 Introduction

Machine reading comprehension, or simply machine comprehension, is the task of asking the computer to read a text passage, and answer questions about the content of the text passage. This particular problem setup is very similar to human reading comprehension problems often seen in standard tests. To make this problem computationally tractable and within the reach of current computational techniques, machine comprehension dataset developers often impose limitations on where the answers can be found.

* We would like to thank the students from Ludong University, particularly Liang Jian (梁健), Xu Yuanyuan (许缘圆), Shang Guofeng (尚国凤), and students from Nanjing Normal University, particularly Liu Han (刘晗), Cao Ziyan (曹紫琰), Mao Xuefen (毛雪芬) for their assistance with data preparation. The second author would like to acknowledge the support from a National Language Committee project (YB135-23) and a Jiangsu Higher Institutions' Excellent Innovative Team for Philosophy and Social Sciences project (2017STD006). The third author would like to acknowledge the support of a National Language Committee "13th Five-Year" Research Plan project (ZD|135-22).

In the widely used machine comprehension data set SQuAD [14], answers to the questions have to be contiguous spans of text from a paragraph. Questions that require answers to be from multiple locations from the paragraph are not permitted in the data set.

Even with this restricted form of questions, the computer needs sophisticated reasoning capability to answer certain types of questions. For example, the computer needs to be able to know that two spans of text refer to the same entity, and it also needs to "understand" alternative expressions that mean the same thing. In some cases, more complicated reasoning capabilities are needed to understand causal, temporal, and other types of semantic or discourse relations. In (1), for example, in order to correctly answer the question (Q) "Why is Jenny able to escape death by zombies?", the system needs to be able to understand that "she" in the context (C) refers to "Jenny", and there is an implicit causal relationship between "she escapes" and she is "protected by an enchanted charm given to her by her mother".

(1) Adapted from MultiRC [8]:

Q: Why is Jenny able to escape death by zombies?

A: She is protected by an enchanted necklace charm given to her by her mother

C: The researchers on the island are killed by the newly risen zombies, except for *Jenny*, the daughter of a scientist couple. *She* escapes, *protected by an enchanted necklace charm given to her by her mother* shortly before her death ...

For complicated natural language processing problems like machine reading, the traditional approach has been one of divide and conquer, and the end application is decomposed into many subproblems which are tackled separately. For example, the problem of recognizing "she" and "Jenny" refer to the same entity is called "coreference resolution", an intermediate NLP task that has little practical value on its own, but is crucial to many end applications and has received a lot of attention over the years [17, 12, 13, 11]. The same thing can be said about paraphrase detection, the task of determining two expressions mean the same thing. Finally, a separate model may also be needed to recognize causal relations, a problem that has also been studied extensively in the context of classifying discourse relations [20, 21]. An advantage of this analytic approach to complex end applications like machine comprehension is that it is easy to find out the weakest link in the system and determine where to devote research effort and resources. The downside is that it is hard to put the different components of a complex system together without causing error propagation, the problem of errors propagating from one component of the system to the next, hurting the overall performance of the system.

The wide adoption of deep learning techniques in the field of NLP makes it possible to design end-to-end neural systems without explicitly addressing each of the subproblems, and this addresses the error propagation problem and leads to improved performance. In some cases, the improvement is even very dramatic. Specific to machine comprehension, a number of systems have reported levels of accuracy that match or

even surpass that of human performance on the SQuAD test set [5] by standard machine comprehension evaluation metrics of exact match or F_1 score. It is worth asking, however, if these state-of-the-art end-to-end deep learning systems have solved intermediate problems like coreference resolution, which has traditionally been considered to be very hard, when answering questions in machine comprehension challenges. It is also interesting to see if, by approaching human-level performance or even outperforming humans, the machine has achieved human-level intelligence.

To answer these questions, we have designed a Stratified MAchine Reading Test (SMART) data set for Chinese where each question is labeled with a "level" that indicates the type of reasoning that is needed to answer that question. We have defined four levels, and hypothesize that these four levels generally correspond with the levels of difficulty encountered by the system. For example, to answer Level 1 questions, the system only needs to perform string match on the question, its possible answer, and the provided context passage. To answer Level 4 questions, however, the system needs to perform multiple types of reasoning. Using BERT [4], a system that provides state-of-the-art results on the SQuAD data set as the baseline, we are able to confirm with experimental results that the four levels we have defined correspond well with the level of difficulty we expected current machine reading systems will encounter, and that state-of-the-art systems still have a lot of difficulty in answering questions that require complicated reasoning.

We also collected human answers to the test portion of the SMART data set. That allows us to not only to compare overall machine performance against human performance, but to see if humans and the machine have the same difficulty in answering questions at different levels. To ultimately make the claim that the machine has achieved human-level performance, it is not enough to simply show the system can answer some types of questions as well as or better than humans, but also to show that the system can answer all types of questions well when compared with humans. Our results show that while the machine can approach human performance in terms of overall accuracy, humans are better at answering questions that require complicated reasoning. This result shows that the machine has a ways to go before reaching human intelligence, a point that might not be too surprising for researchers of the field, but might often be lost in the AI hype.

Our contributions are as follows:

- We provide a large-scale Chinese machine reading data set and plan to make it publicly available to the research community[6].
- We present a novel design for machine comprehension data sets that makes it easier machine comprehension researchers to perform error analysis on system output and to quickly pinpoint weaknesses of the model.
- We establish a strong baseline on this data set with BERT, a system that produces state-of-the-art results on a whole host of NLP tasks that include machine comprehension.

[5] See the leadboard at https://rajpurkar.github.io/SQuAD-explorer/. On SQuAD 1.0, a number of systems have surpassed human performance, and on SQuAD 2.0, the state of the art systems is approaching human performance

[6] Data will be made available here: https://www.cs.brandeis.edu/~clp/smart

– We compare system performance with human performance at each level to identify questions that are particularly hard for humans and for the machine, and show that humans and the machine have different challenges even though their overall performance are comparable.

The remainder of the paper is organized as follows. In Section 2 we discuss related work. In Section 3, we discuss the design of this data set in detail. In Section 4, we describe the baseline system, and in Section 5 we discuss experimental results. We conclude the paper in Section 6.

2 Related work

In the section we briefly describe existing machine comprehension data sets for both English and Chinese, and discuss how they differ from the SMART data set.

2.1 Related English machine comprehension data sets

Existing English machine comprehension data sets fall into two broad categories based on how the questions need to be answered. They are either *span selection* questions where the answer is a span of text from a passage or *multiple choice* questions where the correct answers are among the provided (often four) choices.

Data sets that belong to the first category include SQuAD [14], SearchQA [5], TriviaQA [7], NewsQA [18], and QAngaroo [19], and they vary in size and the type of reasoning that is required to answer the questions in the data set. SQuAD consists of 100K crowdsourced questions collected from 536 English Wikipedia articles. NewsQA has about 120K crowdsourced question-answer pairs from 12,744 CNN news articles. Compared with SQuAD, the NewsQA data set attempts to include a larger portion of questions that require multi-sentence reasoning to answer, and multi-sentence reasoning questions account for about 21% of the questions in NewsQA. The TriviaQA data set contains over 95K question-answer pairs. Evidence documents are collected from Wikipedia and the Web, and multi-sentence reasoning questions account for 40% and 35% of the questions in the two domains respectively. Like the English machine comprehension data sets in this category, the questions in the SMART data also require answers that are selected from a text passage. Unlike these data sets, however, we explicitly label each question in the SMART data set that indicates the type of reasoning needed to answer the question, and this information can be used for machine comprehension researchers to identify weaknesses in their model more quickly.

Data sets that belong to the second category includ MCTest [15], RACE [10], ARC [2], and MultiRC [8], and the multiple choice questions in these data sets have one or more correct answers. Crucially, the answers may not be a span of text from the context passage, and thus often present more of a challenge to the machine. These data sets often differ in their sizes and genre. The MCTest contains 2,000 crowd-sourcing multi-choice questions from 500 fictional stories. The ARC data set has 7,787 natural science, grade-school questions. RACE is a much larger data set that has about 100K questions from English exams for middle and high school Chinese students, and about

26% of the questions in RACE involve multi-sentence reasoning. MultiRC is a smaller data set with about 6,000 questions that focuses on multi-sentence reasoning. Like the span selection questions in the first category, the type of reasoning that is involved in answering these questions is rarely explicitly labeled in these data sets, and machine comprehension researchers would have to characterize the reasoning type themselves if they want to identify the types of questions that are most challenging to their system. In contrast, the SMART data set has a more balanced distribution of the types of questions, and questions that involved complicated reasoning are explicitly labeled as Level 3 or Level 4 questions. The types of reasoning are characterized generally correspond to an intermediate NLP task rather than how many sentences are involved, but NLP tasks like coreference typically involves multi-sentence reasoning.

There are also a small number of datasets that do not fall nicely into those categories. For example, NarrativeQA [9] is a data set of questions about stories, and their answers are human generated and free formed. These questions with free-form answers are more difficult to evaluate, and they often need to be evaluated with metrics such as BLEU or Rouge-L that are harder to interpret.

2.2 Related Chinese machine comprehension data sets

There are relatively few data sets for machine reading for Chinese. [3] describes a cloze test style data set for Chinese which is generated by automatically masking certain words in the text, and thus do not require manual human annotation. Systems are tested to see if they can correctly recover the masked words, and given the powerful language models that are currently readily available, cloze tests are relatively easy to solve without requiring the system to actually "understanding" the text and do any reasoning.

Another Chinese Machine Reading data set is Du-Reader [6], which is collected from queries that real users submitted to the Baidu search engine. While user queries are more representative of real user needs, they present a different kind of challenge than span selection based machine reading data sets like SQuAD, where the correct answer is more objective and system accuracy can be measured with easy-to-interpret metrics like exact match and F_1 score.

Another Chinese machine reading data set is DRCD [16], a data set for traditional Chinese text. Like the SMART dataset, the raw data for DRCD is also from Wikipedia, and answers to questions in DRCD are also spans of text in a passage. The SMART data set differs from DRCD, however, in that the latter does not attempt to stratify the questions in the data set.

3 Constructing the SMART data set

In this section we describe how the SMART data set is constructed.

3.1 Source data preparation

The raw data we have selected for creating question answer pairs for is from Chinese Wikipedia. We extracted the plaintext from the wikipedia dump with wikiextractor[7],

[7] https://github.com/attardi/wikiextractor

and selected articles with a length of between 1,000 and 3,000 characters. We filtered out articles that have too many non-Chinese characters, or have content that is too specialized (e.g., articles on physics or chemistry topics), or are otherwise inappropriate for the machine comprehension task. After this filtering process, the articles we ended up using contain mostly factual information about non-scientific topics such as biographies.

After this preprocessing step, we recruited college students who are Chinese majors from two Chinese universities to create question answer pairs for these articles. Following SQuAD, the articles are broken into smaller passages which consist of one or more paragraphs. The students are asked to create only questions that can be answered with a span of contiguous text in a passage of the article. The students are asked not to create questions that involve mathematical computation, because we believe answering such questions requires very different types of reasoning than questions asking for factual answers.

We depart from the SQuAD approach, however, in that we ask the annotators to also label the "level" of the question when they create these question-answer pairs. We provide the annotators with a set of guidelines in which these different levels are defined and illustrated with examples. We will discuss these levels next. We expect these levels to be broadly aligned with the level of difficulty for the machine, but the assignments of the levels are based on our *a priori* intuition, and they have not been tested empirically when these questions were created.

3.2 Stratified question and answer design

Each question in the SMART data set is labeled with one of four levels, based on the type of reasoning that is involved in answering these questions. The four levels are decided based on the level of challenge we expect the question to pose for a machine reasoning system, based on our understanding of how current machine reading systems work. For each level, we define the kind of reasoning that is needed to answer questions at the level, and ask the annotators to mark the level when the create the questions. The reasoning that is needed for each level are described below:

- Level 1: For questions of this level, the machine only needs to find the answer to a question based on string match.
- Level 2: To answer Level 2 questions, the system needs to be able to recognize paraphrases or syntactic variations.
- To answer Level 3 questions, the system needs to i) resolve the pronominal mention of entity to a named or nominal entity because the pronouns cannot be answers to questions themselves as they are not self-identifying, or ii) perform temporal or causal reasoning. The pronouns that need to be resolved include dropped pronouns, which are wide-spread as Chinese is a pro-drop language. For level 3 questions, the system only needs to perform one type of reasoning described above.
- To answer Level 4 questions, the system needs to perform multiple types of reasoning. For example, the system might need to perform coreference resolution as well as causal reasoning when answering a Level 4 question.

We illustrate each question level with examples. The example in (2) is a Level 1 question because to correctly answer this question, the system only needs to replace the question word/phrase 什么实验室 ("which laboratory") with 贝尔实验室 ("Bell Labs"), and the rest of the question matches the context sentence exactly.

(2) Level 1
Q: 1947年什么实验室发明晶体管已被列在IEEE里程碑列表中?
Which laboratory invented transistors in 1947, which has been listed in the IEEE Milestones?
A: 贝尔实验室
Bell Labs
C: 1947年贝尔实验室发明晶体管已被列在IEEE里程碑列表中
Bell Labs invented transistors in 1947, which has been listed in the IEEE Milestones

The example in (3) illustrates a Level 2 question. For Level 2 questions, replacing the question word/phrase in question with the answer does not lead to an exact match with the context due to use of synonymous words, variations in word order, or extra lexical material. In (3), replacing the question word 何时 ("when") with the answer 2016年7月 ("July, 2016") in the question does not lead to an exact match because of the change in word order and the extra lexical material in the context. Nevertheless, there is a partial match which provides a strong signal that 2016年7月 ("July 2016") is the correct answer.

(3) Level 2
Q: 南马都尔何时被联合国教科文组织认定为世界遗产?
When was Nan Madol recognized as a World Heritage by UNESCO?
A: 2016年7月
In July, 2016
C: 2016年7月，在土耳其伊斯坦布尔召开的第40届世界遗产委员会上，南马都尔被联合国教科文组织认定为世界遗产。
In July 2016, at the 40th Session of the World Heritage Committee held in Istanbul, Turkey, Nan Madol was recognized as a World Heritage by UNESCO.

The example in (4) illustrates a Level 3 question where replacing the question word 谁 ("who") with the pronoun 他 ("he") in the question would lead to an exact match with the context, but the pronoun needs to be resolved to a named entity 拉梅尔 ("R. J. Rummel") to answer the question as the pronoun itself is not self-identifying and does not serve as an informative answer.

(4) Level 3: Resolving an overt pronoun to its antecedent
Q: 谁在接下来15年里埋头于建构民主和平的理论?
Who is immersed in constructing the theory of democracy and peace in the next 15 years?
A: 拉梅尔
R. J. Rummel

C: 拉梅尔着作丰富，写下了24本学术书籍，并且在1975-1981年间出版的《认识冲突与战争》（"Understanding Conflict and War"）中记载了他研究的主要成果。他在接下来15年里埋头于建构民主和平的理论，不断加入各种新的资料和数据测试，对比其他人的研究成果，并对许多单独的战争案例进行研究
R. J. Rummel is rich in writing, has written 24 academic books, and recorded the main results of his research in "Understanding Conflict and War" published between 1975-1981. He has been immersed in constructing the theory of democracy and peace for the next 15 years, constantly adding new data and tests, comparing other people's research results with his own, and researching many individual war cases.

The example in (5) also illustrates a Level 3 question. In this case, replacing the question phrase 多少公里 ("how many kilometers") with the answer 120公里("120 kilometers") does not lead to a match. It is also necessary to resolve the dropped pronoun ∗ pro∗ to the named entity 巨石阵 ("Stonehenge"). Dropped pronouns are also known as zero pronouns, and are a phenomenon that have been explicitly studied in Chinese NLP [22, 1].

(5) Level 3: Resolving an implicit pronoun to its antecedent
Q: 巨石阵位于英国离伦敦大约多少公里一个叫做埃姆斯伯里的地方?
How many kilometers from London is the place in the United Kingdom called Amesbury where Stonehenge located?
A: 120公里
120 Kilometers
C: 巨石阵也叫做圆形石林，[∗ pro∗]位于英国离伦敦大约120公里一个叫做埃姆斯伯里的地方。
Stonehenge, also known as the Round Stone Forest, is located in a place in the United Kingdom about 120 kilometers from London, called Amesbury.

The question in (6) illustrates a Level 4 question. It takes multiple reasoning steps to correctly answer this question. First of all it needs to resolve the pronoun 他 ("he") to the named entity 聂鲁达 ("Neruda"), and then it needs to recognize not angering his father is the reason for using the pen name Neruda.

(6) Level 4
Q: 为什么聂鲁达以自己仰慕的捷克诗人扬·聂鲁达的姓氏为自己取了笔名"聂鲁达"?
Why did Neruda take the surname of the Czech poet Jan Neruda that he admired as his pen name "Neruda" ?
A: 为了避免引起父亲的不满
To avoid angering his father
C: 1920年，聂鲁达开始在塞尔瓦奥斯塔尔杂志上刊登短文和诗，为了避免引起父亲的不满，他以自己仰慕的捷克诗人扬·聂鲁达（Jan Neruda）的姓氏为自己取了笔名"聂鲁达"。
In 1920, Neruda began to publish short essays and poems in the magazine Selva Ostal. In order to avoid angering his father, he took the surname of his admired Czech poet Jan Neruda as his pen name "Neruda".

The examples above do not provide all possible forms of reasoning that are needed in order to answer machine comprehension questions, but they are the most frequently attested types of reasoning that are needed in our data set.

3.3 Key statistics of the data set

The SMART data set consists 39,408 question answer pairs from 564 Chinese Wikipedia articles, and we split the the whole data set into train/dev/test sets by taking articles as basic units (meaning all questions for an article will be in the same set), and setting the proportions of questions in the three sets to roughly 80%/10%/10% of the entire data set. Table 1 shows the distribution of questions across different levels and across different sets. As can be seen from the table, the number of questions is not evenly distributed across the four levels, with much more Level 1 and Level 3 questions than Level 2 and Level 4 questions. There are 15,476 level 3 and level 4 questions in the SMART data set and they account for about 40% of the questions in the data set.

Dataset	Level 1	Level 2	Level 3	Level 4	Overall
Train	15,181	3,945	10,868	1,402	31,396
Development	1,822	491	1,464	200	3,977
Test	1,987	506	1,349	193	4,035
Overall	18,990	4,942	13,681	1,795	39,408
Percentage	48.2%	12.5%	34.7%	4.6%	100%

Table 1. Number of instances for each level.

The four-level system is obviously still a very coarse-grained classification, and a more fine-grained classification is possible. In the meantime, a more fine-grained classification might put too much burden on student annotators, and we felt that the four-level classification is a good initial trade-off. We did look further into Level 3 and Level 4 questions, and found that most of the Level 3 questions involve coreference resolution.

4 Establishing a baseline

To evaluate how well a machine comprehension system can perform on the SMART dataset, we leverage the state-of-the-art BERT model [4] as our baseline model. For a given question $Q = (q^1, ..., q^{|Q|})$ and the corresponding context/passage $P = (p^1, ..., p^{|P|})$, where $q^i \in Q$ and $p^j \in P$ are words, we concatenate the question and the context P into a new sequence "[CLS] $p^1, ..., p^{|P|}$ [SEP] $q^1, ..., q^{|Q|}$ [SEP]", then apply the BERT model to encode this sequence. Then the vector representation of each word position from BERT encoder is fed into two separate dense layers to predict the start and end probabilities. During training, the log-likelihood of the correct start and end positions is optimized. During inference, the BERT model evaluates scores for each answer span by multiplying the start and end probabilities, and then the highest scoring span is selected as the final answer. In our experiment, we leverage the pre-trained Chinese BERT-base model with default hyper-parameters.

5 Experiments

We train the BERT model on the training from the SMART data set and evaluate the model on the development and test sets. We use the exact match (EM) and F_1 scores introduced in [14] as the evaluation metrics. However, we have to modify how the F_1 score is computed by viewing answers and predictions as a sequence of characters rather than a sequence of words, since there is no natural word delimiting white space between words in Chinese. This change inflates the F_1 score somewhat as a word in Chinese can have more than one character. The alternative would be using an automatic word segmenter to segment the answers into words but that would complicate the computation. In addition, the word segmenter would not be 100% accurate. We computed the overall EM and F_1 score as well as the EM and F_1 scores for each level for both the development and test set, and the results are presented in Table 2.

Data Set	Level	exact_match	F_1
Test	l1	82.5	91.9
	l2	79.8	91.0
	l3	72.6	87.7
	l4	64.2	84.8
	Overall	78.0	90.0
Dev	l1	82.8	91.5
	l2	79.0	89.3
	l3	73.6	87.4
	l4	58.5	79.0
	Overall	77.7	89.1

Table 2. System performance for each level.

As we only use the default parameters in BERT so that others can easily replicate our result, we do not strictly speaking need a separate development set for system development purposes. However, having both a development and test set helps to show that higher level questions are consistently more difficult for BERT than lower level questions in both the development and test sets, and this result bears out our expectation about the level of difficulty for questions in different levels. It is also worth noting that while there is a precipitous drop in accuracy from Level 2 to Level 3, and from Level 3 to Level 4, the drop in accuracy from Level 1 to Level 2 is more modest, indicating the system is getting very good at handling periphrastic expressions due to syntactic variations or the use of synonyms.

We also collected human answers to the questions in the test set from a group of college students in China (separate from the group who created the questions and answers). We collected three answers for each question, and computed the average accuracy for those answers. The human performance is presented in Table 3. Several observations can be made from this table. First, in contrast with the machine, questions in the higher level are not necessarily more difficult to answer for humans, as indicated by the higher EM and F_1 scores for Level 2 than Level 1. If we look at just the F_1 scores, Level 3 questions are not more difficult than Level 2 questions either, and the variation in accu-

Data Set	Level	exact_match	F_1
Test	l1	79.5	93.8
	l2	82.1	94.3
	l3	71.2	91.5
	l4	65.7	91.2
	Overall	76.3	92.8

Table 3. Human performance for each level on test set. The results are the average of three groups of students.

racy across all four levels is rather small, indicating humans can handle these different types of reasoning with relative ease, in contrast with the machine.

We also investigated the rather large discrepancy between the EM scores and F_1 scores, and found that humans are not particularly precise when selecting a span of the text as answers to the questions. While they get roughly the correct answer, they might include extra material or missing some detail. For example, humans might choose 主编 李大同 ("Editor-in-Chief Li Datong") rather the correct answer 李大同 ("Li Datong"), but it is essentially the correct answer, even though the EM score would be zero in this case. In contrast, the machine often makes the mistake of not producing an answer at all, or a totally incorrect answer, ending up with a zero score for both EM and F_1.

A comparison between human results and system results suggests that humans and the machine, in this case BERT, might use very different mechanisms. While the machine seems to be very good at answering questions that involve low-level reasoning (e.g., Level 1 questions), humans are better at answering questions that involve high-level reasoning (Level 3 and Level 4 questions). On the other hand, when the machine can answer a question, it can often answer it more precisely than humans, as indicated by the slightly higher EM scores achieved by the system.

6 Conclusion and Future Work

We presented SMART, a large-scale machine comprehension data set for Chinese. We show the stratified design of the questions in the data set allows machine comprehension researchers to quickly focus in on the type of questions that are most challenging for the system. We also present results on how humans answer the same questions and our results show that when we compare system and human performance, our analysis needs to be more nuanced than just to say the system is approaching or outperforming humans. Our results show humans and the machine have different strengths and suggest that humans and the machine, as represented by current state of the art, use very different mechanisms when answering reading comprehension questions.

References

1. Chen, C., Ng, V.: Chinese zero pronoun resolution: Some recent advances. In: Proceedings of the 2013 conference on empirical methods in natural language processing (2013)
2. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have solved question answering? try arc, the AI2 reasoning challenge. CoRR **abs/1803.05457** (2018), http://arxiv.org/abs/1803.05457
3. Cui, Y., Liu, T., Chen, Z., Wang, S., Hu, G.: Consensus attention-based neural networks for chinese reading comprehension. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (2016)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Dunn, M., Sagun, L., Higgins, M., Güney, V.U., Cirik, V., Cho, K.: Searchqa: A new q&a dataset augmented with context from a search engine. CoRR **abs/1704.05179** (2017), http://arxiv.org/abs/1704.05179

6. He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., Liu, X., Wu, T., Wang, H.: Dureader: a chinese machine reading comprehension dataset from real-world applications. In: Proceedings of the Workshop on Machine Reading for Question Answering. pp. 37–46 (2018)
7. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada (July 2017)
8. Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., Roth, D.: Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 252–262 (2018)
9. Kocisky, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K.M., Melis, G., Grefenstette, E.: The narrativeqa reading comprehension challenge. Transactions of the Association for Computational Linguistics **6**, 317–328 (2018)
10. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: Race: Large-scale reading comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (2017)
11. Lee, K., He, L., Lewis, M., Zettlemoyer, L.: End-to-end neural coreference resolution. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark (2017)
12. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th annual meeting on association for computational linguistics (2002)
13. Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.: A multi-pass sieve for coreference resolution. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (2010)
14. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (2016)
15. Richardson, M., Burges, C.J., Renshaw, E.: Mctest: A challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013)
16. Shao, C., Liu, T., Lai, Y., Tseng, Y., Tsai, S.: DRCD: a chinese machine reading comprehension dataset. CoRR **abs/1806.00920** (2018), `http://arxiv.org/abs/1806.00920`
17. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Computational linguistics **27**(4), 521–544 (2001)
18. Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., Suleman, K.: Newsqa: A machine comprehension dataset. In: Proceedings of the 2nd Workshop on Representation Learning for NLP (2017)
19. Welbl, J., Stenetorp, P., Riedel, S.: Constructing datasets for multi-hop reading comprehension across documents. Transactions of the Association for Computational Linguistics **6**, 287–302 (2018)
20. Xue, N., Ng, H.T., Pradhan, S., Prasad, R., Bryant, C., Rutherford, A.: The conll-2015 shared task on shallow discourse parsing. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task. pp. 1–16 (2015)
21. Xue, N., Ng, H.T., Pradhan, S., Rutherford, A., Webber, B., Wang, C., Wang, H.: Conll 2016 shared task on multilingual shallow discourse parsing. In: Proceedings of the CoNLL-16 shared task (2016)
22. Zhao, S., Ng, H.T.: Identification and resolution of chinese zero pronouns: A machine learning approach. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (2007)