Active Reading Comprehension: A dataset for learning the Question-Answer Relationship strategy

Diana Galvan
Tohoku University
Graduate School of Information Sciences
Sendai, Japan
dianags@ecei.tohoku.ac.jp

Abstract

Reading comprehension (RC) through question answering is a useful method for evaluating if a reader understands a text. Standard accuracy metrics are used for evaluation, where a high accuracy is taken as indicative of a good understanding. However, literature in quality learning suggests that task performance should also be evaluated on the undergone process to answer. The Question-Answer Relationship (QAR) is one of the strategies for evaluating a reader's understanding based on their ability to select different sources of information depending on the question type. We propose the creation of a dataset to learn the QAR strategy with weak supervision. We expect to complement current work on reading comprehension by introducing a new setup for evaluation.

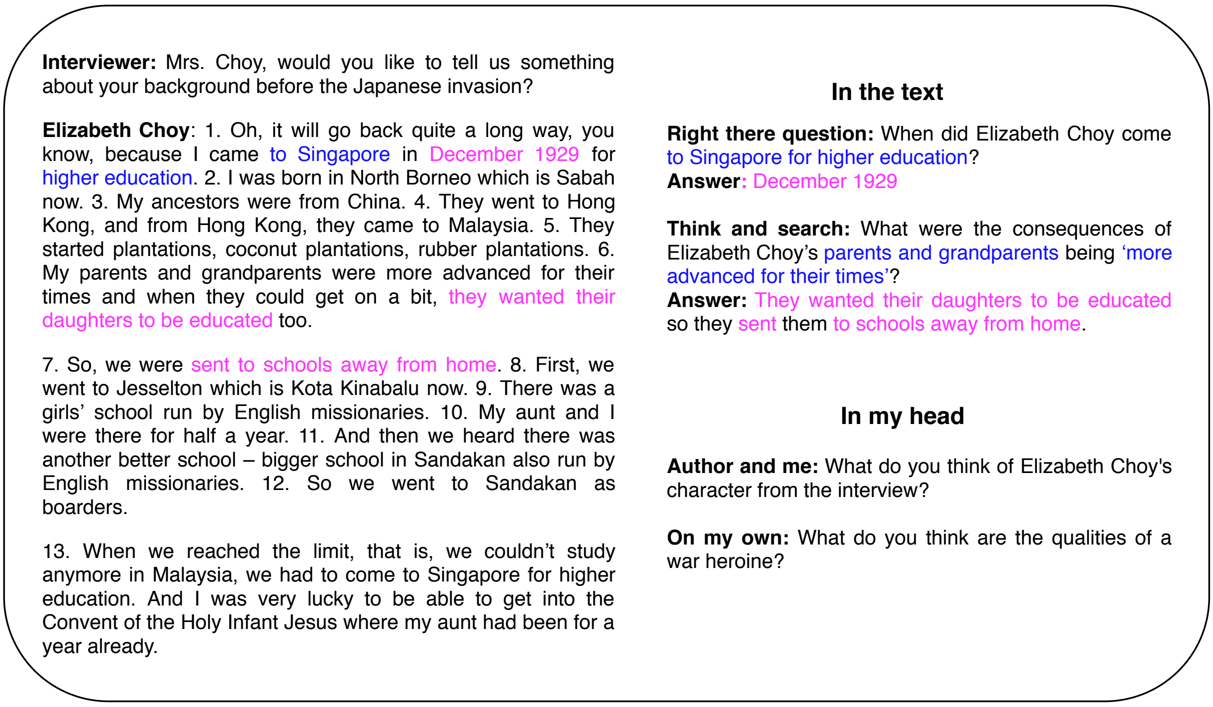
1 Introduction

Computer system researchers have long been trying to imitate human cognitive skills like memory (Hochreiter and Schmidhuber, 1997; Chung et al., 2014) and attention (Vaswani et al., 2017). These skills are essential for a number of Natural Language Processing (NLP) tasks including reading comprehension (RC). Until now, the method for evaluating a system's understanding imitated the common classroom setting where students are evaluated based on their number of correct answers. In the educational assessment literature this is known as *product-based* evaluation and is one of the performance-based assessments types (McTighe and Ferrara, 1994). However, there is an alternative form: *process-based* evaluation. Process-based evaluation does not emphasize the output of the activity. This assessment aims to know the step-by-step procedure followed to resolve a given task.

When a reading comprehension system is not able to identify the correct answer, product-based evaluation can result in the false impression of weak understanding (i.e., misunderstanding of the text, the question, or both) or the absence of required knowledge. However, the system could have failed to arrive at the correct answer for some other reasons. For example, consider the reading comprehension task shown in Figure 1. For the question *"What were the consequences of Elizabeth Choy's parents and grandparents being 'more advanced for their times'?"* the correct answer is in the text but it is located in different sentences. If the system only identifies *"They wanted their daughters to be educated"* as an answer, it would be judged to be incorrect when it did not fail at finding the answer, it failed at connecting it with the fact *"we were sent to schools away from home"* (linking problem). Similarly, any answer the system infers from the text for the question *"What do you think are the qualities of a war heroine?"* would be wrong because the answer is not in the text, it relies exclusively on background knowledge (wrong choice of information source). We propose to adopt the thesis that reading is not a passive process by which readers soak up words and information from the text, but an active process[1] by which they predict, sample, and confirm or correct their hypotheses about the text (Weaver, 1988). One of these hypotheses is which source of information the question requires. The reader might think it is necessary to look in the text to then realize she could have answered without even reading or, on the contrary, try to think of an answer even though it is directly in the text. For this reason, Raphael (1982) devised the Question-Answer Relation (QAR) strategy, a technique to help the reader decide the most suitable source of information as well as the level of reasoning

[1] Not to be confused with *active learning*, a machine learning concept for a series of methods that actively participate in the collection of training examples. (Thompson et al., 1999)

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 106–112
Florence, Italy, July 28 - August 2, 2019.



Interviewer: Mrs. Choy, would you like to tell us something about your background before the Japanese invasion?

Elizabeth Choy: 1. Oh, it will go back quite a long way, you know, because I came to Singapore in December 1929 for higher education. 2. I was born in North Borneo which is Sabah now. 3. My ancestors were from China. 4. They went to Hong Kong, and from Hong Kong, they came to Malaysia. 5. They started plantations, coconut plantations, rubber plantations. 6. My parents and grandparents were more advanced for their times and when they could get on a bit, they wanted their daughters to be educated too.

7. So, we were sent to schools away from home. 8. First, we went to Jesselton which is Kota Kinabalu now. 9. There was a girls' school run by English missionaries. 10. My aunt and I were there for half a year. 11. And then we heard there was another better school – bigger school in Sandakan also run by English missionaries. 12. So we went to Sandakan as boarders.

13. When we reached the limit, that is, we couldn't study anymore in Malaysia, we had to come to Singapore for higher education. And I was very lucky to be able to get into the Convent of the Holy Infant Jesus where my aunt had been for a year already.

In the text

Right there question: When did Elizabeth Choy come to Singapore for higher education?
Answer: December 1929

Think and search: What were the consequences of Elizabeth Choy's parents and grandparents being 'more advanced for their times'?
Answer: They wanted their daughters to be educated so they sent them to schools away from home.

In my head

Author and me: What do you think of Elizabeth Choy's character from the interview?

On my own: What do you think are the qualities of a war heroine?

Figure 1: Example of reading comprehension applying the Question-Answer Relationship strategy to categorize the questions.

needed based on the question type.

In this work, we introduce a new evaluation setting for reading comprehension systems. We overview the QAR strategy as an option to move beyond a scenario where only the product of comprehension is evaluated and not the process. We discuss our proposed approach to create a new dataset for learning the QAR strategy using existing reading comprehension datasets.

2 Related work

Reading comprehension is an active research area in NLP. It is composed of two main components: text and questions. This task can be found in many possible variations: setups where no options are given and the machine has to come up with an answer (Yang et al., 2015; Weston et al., 2015; Nguyen et al., 2016; Rajpurkar et al., 2016) and setups where the question has multiple choices and the machine needs to choose one of them (Richardson et al., 2013; Hill et al., 2015; Onishi et al., 2016; Mihaylov et al., 2018). In either case, standard accuracy metrics are used to evaluate systems based on the number of correct answers retrieved; a product-based evaluation. In addition to this evaluation criteria, the current reading comprehension setting constrains systems to be trained on a particular domain for a specific type of reasoning. As a result, the good performance of a model drops when it is tested on a different domain. For example, the knowledgeable reader of Mihaylov and Frank (2018) was trained to solve questions from children narrative texts that require commonsense knowledge, achieving competitive results. However, it did not perform equally well when tested on basic science questions that also required commonsense knowledge (Mihaylov et al., 2018). Systems have been able to match human performance but it has also been proven by Jia and Liang (2017) that they can be easily fooled with adversarial distracting sentences that would not change the correct answer or mislead humans.

The motivation behind introducing adversarial examples for evaluating reading comprehension is to discern to what extent systems truly understand language. Mudrakarta et al. (2018) followed the steps of Jia and Liang (2017) proposing a technique to analyze the sensitivity of a model to question words, with the aim to empower investigation of reading models' performance. With the same goal in mind, we propose a process-based evaluation that will favor a closer examination of the process taken by current systems to solve a reading comprehension task. In the educational assessment literature, this approach is recommended to identify the weaknesses of a student. If we transfer this concept to computers, we would be able to focus on the comprehension tasks a computer is weak in, regardless of the data in which the system has been trained.

3 Question-answer relationship

Raphael (1982) devised the Question-Answer Relationship as a way of improving children reading performance across grades and subject areas. This approach reflects the current concept of reading as an active process influenced by characteristics of the reader, the text, and the context within which the reading happens (McIntosh and Draper, 1995). Since its publication, several studies have explored its positive effects (Benito et al., 1993; McIntosh and Draper, 1995; Ezell et al., 1996; Thuy and Huan, 2018; Apriani, 2019).

QAR states that an answer and its source of information are directly related to the type of question being asked. It emphasizes the importance of being able to locate this source to then identify the level of reasoning the question requires. QAR defines four type of questions categorized in two broad sources of information:

In the text

- **Right There questions:** The answer can be literally found in the text.
- **Think and Search questions:** The answer can be found in several sentences in the text that need to be pieced together.

In my head

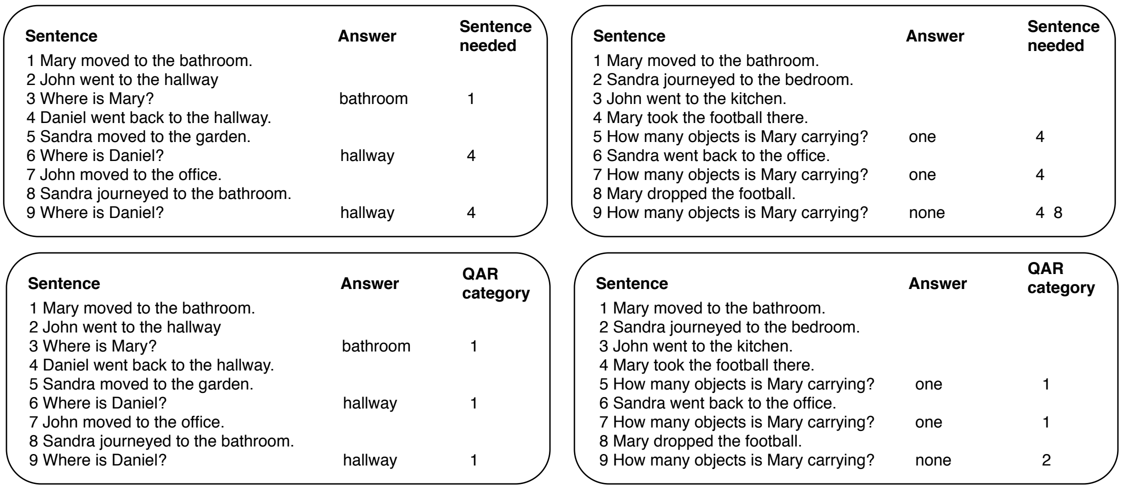
- **Author and Me questions:** The answer is not directly stated in the text. It is necessary to fit text information with background knowledge.
- **On My Own questions:** The answer can be given without reading the text. The answer relies solely on background knowledge.

Each one of the QAR categories requires a different level of reasoning. For *Right there* questions, the reader only needs to match the question with one of the sentences within the text. *Think and search* requires simple inference to relate pieces of information contained in different parts of the text. *In my head* questions introduce the use of background knowledge. Thus, deeper thinking is required to relate the information provided in the text with background information. Finally, *On my own* questions ask the reader to only use their background knowledge to come up with an answer. Figure 1 shows how QAR is applied to a reading comprehension task. Note that for both *In the text* questions, one can easily match the words in the question with the words in the text. However, *Think and search* goes beyond matching ability; the reader should be able to conclude that the information in sentences 6 and 7 are equally required to answer the question being asked. Thus, the correct answer is a combination of these two. For the *Author and me* question, the readers need to merge the information given in the text with their own background knowledge since the question explicitly asks for an opinion *"from the interview."* Without this statement, the question could be considered as *On my own* if the reader is already familiar with Elizabeth Choy. This is not the case in the last question, where even though the topic of the interview is related, the qualities of a war heroine are not in the text. The readers need to use their own background knowledge about heroes.

In the case of computers, *In my head* questions can be understood as *In a knowledge base*. We hypothesize that once the system establishes that the source of information is not in the text, it could trigger a connection to a knowledge base. For the time being, the type of knowledge needed is fixed for RC datasets by design (e.g., general domain, commonsense, elementary science) and the source is chosen accordingly in advance by the author (e.g., Wikipedia, ConceptNet). Automatically selecting the appropriate external resource for a reading comprehension task is a problem that we would like to explore in the future.

3.1 QAR use cases

As a process-based evaluation strategy, QAR can be used to understand a reader's ability in terms of the reasoning level applied and the elected source of information to answer a given question. In the case of humans, this outcome is later used as feedback to improve performance on a particular process. The incorporation of general reading strategies to a RC system has been recently proven effective by Sun et al. (2018) and we aim to explore QAR in the same way. However, our short-term objective is to test the QAR strategy as a complementary evaluation method for existing machine



Sentence	Answer	Sentence needed
1 Mary moved to the bathroom.		
2 John went to the hallway		
3 Where is Mary?	bathroom	1
4 Daniel went back to the hallway.		
5 Sandra moved to the garden.		
6 Where is Daniel?	hallway	4
7 John moved to the office.		
8 Sandra journeyed to the bathroom.		
9 Where is Daniel?	hallway	4

Sentence	Answer	Sentence needed
1 Mary moved to the bathroom.		
2 Sandra journeyed to the bedroom.		
3 John went to the kitchen.		
4 Mary took the football there.		
5 How many objects is Mary carrying?	one	4
6 Sandra went back to the office.		
7 How many objects is Mary carrying?	one	4
8 Mary dropped the football.		
9 How many objects is Mary carrying?	none	4 8

Sentence	Answer	QAR category
1 Mary moved to the bathroom.		
2 John went to the hallway		
3 Where is Mary?	bathroom	1
4 Daniel went back to the hallway.		
5 Sandra moved to the garden.		
6 Where is Daniel?	hallway	1
7 John moved to the office.		
8 Sandra journeyed to the bathroom.		
9 Where is Daniel?	hallway	1

Sentence	Answer	QAR category
1 Mary moved to the bathroom.		
2 Sandra journeyed to the bedroom.		
3 John went to the kitchen.		
4 Mary took the football there.		
5 How many objects is Mary carrying?	one	1
6 Sandra went back to the office.		
7 How many objects is Mary carrying?	one	1
8 Mary dropped the football.		
9 How many objects is Mary carrying?	none	2

Figure 2: Example of bAbI annotations for the *single supportive fact* task (left) and the *counting* task (right). Below, our proposed annotations with QAR category.

reading comprehension models, somewhat similar to PROTEST (Guillou and Hardmeier, 2016), a test suite for the evaluation of pronoun translation by Machine Translation systems.

In the next section, we discuss how the QAR strategy can be imported from the educational literature to the NLP domain by using existing reading comprehension datasets to create a new resource for active reading comprehension evaluation.

4 Research plan

4.1 Dataset

We propose to model QAR learning as a multiclass classification task with weak supervision. The dataset would contain labels corresponding to each one of the QAR categories and the annotation process will depend on the two sources of information Raphael (1982) defined.

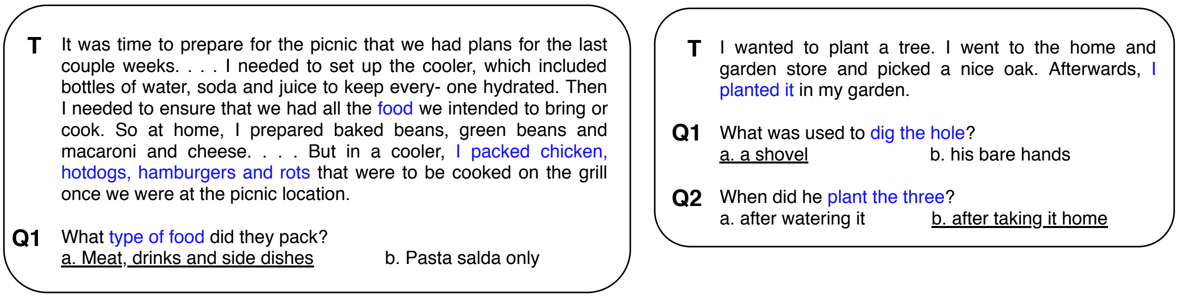
In recent years, we have seen a lot of effort from the NLP community in creating datasets to test different aspects of RC, like bAbI (Weston et al., 2015), SQuAD (Rajpurkar et al., 2016), NarrativeQA (Kočiskỳ et al., 2018), QAngaroo (Welbl et al., 2018), HotpotQA (Yang et al., 2018), MCScript (Ostermann et al., 2018), MultiRC (Khashabi et al., 2018) and CommonsenseQA (Talmor et al., 2018). In the following sections, we will briefly overview these datasets and explain how they can be adapted for our proposed task.

4.1.1 In the text questions

For this type of questions, we can rely on the bAbI dataset (Weston et al., 2015), a set of synthetically generated, simple narratives for testing text understanding. The dataset has several tasks with 1000 questions each for training and 1000 for testing. For our purposes, we will focus on the annotations of Task 8 and 7. Task 8 is a “single supporting fact” task that shows a small passage in which each sentence describes the location of a character (e.g. *“Mary moved to the bathroom. John went to the hallway.”*). After some sentences, there is a question asking where the character is (e.g. *“Where is Mary?”*) and the goal is to give a single word answer to it (e.g. *“bathroom”*). Task 7 is a “counting” task describing the same situation, but it aggregates a sentence where one of the characters either takes (e.g. *“Mary took the football there.”*) or drops (e.g. *“Mary dropped the football.”*) an object. This time, the question asks how many objects is the character carrying and the answer is also a single word (e.g. *“none”*). As shown in Figure 2, bAbI annotations enumerate each one of the sentences. The number next to the single word answer is the number of the sentence needed to answer the question. Instead of the number of the sentence, we will use as label the number of the QAR category. This can be done following this rule:

$$QARcategory = \left\{ \begin{array}{ll} 1, & \text{for } n = 1 \\ 2, & \text{for } n > 1 \end{array} \right\}$$

Where n is the number of sentences and the categories *1, 2* correspond to *Right there* and *Think and Search*, respectively. The bottom of Figure 2 shows how the new annotations will look like. This annotations can be generated automatically



T It was time to prepare for the picnic that we had plans for the last couple weeks. . . . I needed to set up the cooler, which included bottles of water, soda and juice to keep every- one hydrated. Then I needed to ensure that we had all the food we intended to bring or cook. So at home, I prepared baked beans, green beans and macaroni and cheese. . . . But in a cooler, I packed chicken, hotdogs, hamburgers and rots that were to be cooked on the grill once we were at the picnic location.

Q1 What type of food did they pack?
a. Meat, drinks and side dishes b. Pasta salda only

T I wanted to plant a tree. I went to the home and garden store and picked a nice oak. Afterwards, I planted it in my garden.

Q1 What was used to dig the hole?
a. a shovel b. his bare hands

Q2 When did he plant the three?
a. after watering it b. after taking it home

Figure 3: MCScript annotations for *text-based questions* (left) and *common sense questions* (right). In blue, key words and phrases necessary to arrive at the correct answer.

using a script that implements the aforementioned rule.

The same approach can be applied to HotpotQA and MultiRC. HotpotQA is a dataset with 113k Wikipedia-based question-answer pairs for which reasoning over multiple documents is needed to answer. Its annotations already identify the sentence-level supporting facts required for reasoning, making this dataset a perfect match for our subset of *Think and search* questions. SQuAD (100,000+ questions) has a very similar design and format, although questions are designed to be answered by a single paragraph. Since the correct answer is literally contained in one part of the text, questions will fall under the *Right there* category. The annotations only include the start-offset of the answer in the text, but we can easily use this information to identify the answer's position at a sentence level. In the same line of multiple-sentence reasoning, MultiRC presents $\sim 6k$ multiple-choice questions from paragraphs across 7 different domains. The additional challenge of this dataset is that multiple correct answers are allowed. Since the supporting sentences are already annotated, this dataset can be used entirely as a *Think and search* instance.

The multi-hop nature of QAngaroo and NarrativeQA questions also match the *Think and search* category. However, no span or sentence-level annotation is provided, making this datasets unsuitable for our approach.

4.1.2 In my head questions

For these questions we will use the MCScript dataset (Ostermann et al., 2018). This dataset is intended to be used in a machine reading comprehension task that requires reasoning over *script knowledge*, sequences of events describing stereotypical human activities. MCScript contains 2,100 narrative texts annotated with two types of questions: *Text-based* questions and *commonsense* questions with 10,160 and 3,827 questions each. *Text-based* questions match *Author and me* category since the answer is not directly contained within the text; it is necessary to combine the text information with background knowledge (script knowledge). *Commonsense* questions, on the other hand, depend only on background knowledge. Thus, there is no need to read the text to answer if the script activity is known.

Consider the example annotations shown in Figure 3. For the text on the left, the reader cannot give an answer even if it has knowledge of types of foods. It is necessary to read the text to identify the types of food the characters in the text packed. In contrast, the questions for the text on the right can be answered if the reader is familiar with the scenario of planting a tree.

The MCScript training annotations identify the correct answer and whether this can be found in the text or if commonsense knowledge is needed. All questions where commonsense is required can be assumed to be *On my own* questions. However, there are some *Text-based* questions in which the answer is explicitly contained in the text. It would be necessary to review these questions to manually annotate the *Author and me* QAR type. This could be achieved in a crowd-sourcing process, instructing the annotators on script knowledge and asking them to label a question as *Author and me* if they first are not able to answer without reading the text.

With a major focus on background knowledge, CommonsenseQA shifts from the common text-question-answer candidates setting to only question-answer candidates. This dataset could in principle complement the *On my own* questions type, but the absence of a passage makes CommonsenseQA inconsistent for a RC task.

To ensure the integrity of our resulting dataset, we will take a subset for manual inspection.

5 Summary

We introduced process-based evaluation as a new setting to evaluate systems in reading comprehension. We propose to model QAR learning as a weak supervision classification task and discussed how existing RC datasets can be used to generate new data for this purpose. Our work is inspired by the findings of the educational assessment field and we expect it to complement current work in reading comprehension. We will leave the details on how to use the QAR classification task for a RC model's evaluation performance to future work.

Acknowledgements

I would like to thank my supervisors Jun Suzuki, Koji Matsuda and Kentaro Inui for their feedback, support and helpful discussions. I am also thankful to Benjamin Heinzerling, Kyosuke Nishida and the anonymous reviewers for their insightful comments. This work was supported by JST CREST Grant Number JPMJCR1513, Japan.

References

Luthfiyah Apriani. 2019. The use of question-answer relationship to improve students' reading comprehension. In *International Seminar and Annual Meeting BKS-PTN Wilayah Barat*, volume 1.

Yolande M Benito, Christy L Foley, Craig D Lewis, and Perry Prescott. 1993. The effect of instruction in question-answer relationships and metacognition on social studies comprehension. *Journal of Research in Reading*, 16(1):20–29.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Helen K Ezell, Stacie A Hunsicker, Maria M Quinque, and Elizabeth Randolph. 1996. Maintenance and generalization of qar reading comprehension strategies. *Literacy Research and Instruction*, 36(1):64–81.

Liane Guillou and Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In *LREC*.

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*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. 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)*, pages 252–262.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association of Computational Linguistics*, 6:317–328.

Margaret E McIntosh and Roni Jo Draper. 1995. Applying the question-answer relationship strategy in mathematics. *Journal of Adolescent & Adult Literacy*, 39(2):120–131.

Jay McTighe and Steven Ferrara. 1994. Performance-based assessment in the classroom. *Pennsylvania Educational Leadership*, pages 4–16.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. *arXiv preprint arXiv:1805.07858*.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? *arXiv preprint arXiv:1805.05492*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. *arXiv preprint arXiv:1608.05457*.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. Mcscript: a novel dataset for assessing machine comprehension using script knowledge. *arXiv preprint arXiv:1803.05223*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Taffy E Raphael. 1982. Question-answering strategies for children. *Reading Teacher*.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. 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*, pages 193–203.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2018. Improving machine reading comprehension with general reading strategies. *arXiv preprint arXiv:1810.13441*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Cynthia A Thompson, Mary Elaine Califf, and Raymond J Mooney. 1999. Active learning for natural language parsing and information extraction. In *ICML*, pages 406–414. Citeseer.

Nguyen Thi Bich Thuy and Nguyen Buu Huan. 2018. The effects of question-answer relationship strategy on efl high school studentsreading comprehension. *European Journal of English Language Teaching*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Constance Weaver. 1988. *Reading process and practice: From socio-psycholinguistics to whole language*. ERIC.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association of Computational Linguistics*, 6:287–302.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.