On Making Reading Comprehension More Comprehensive

Matt Gardner,[♠] **Jonathan Berant,**[♠,♣] **Hannaneh Hajishirzi,**[♠,♢]
Alon Talmor,[♣] **and Sewon Min,**[♢]
[♠]Allen Institute for Artificial Intelligence
[♣]Tel Aviv University
[♢]University of Washington
mattg@allenai.org

Abstract

Machine reading comprehension, the task of evaluating a machine's ability to comprehend a passage of text, has seen a surge in popularity in recent years. There are many datasets that are targeted at reading comprehension, and many systems that perform as well as humans on some of these datasets. Despite all of this interest, there is no work that systematically defines what reading comprehension *is*. In this work, we justify a question answering approach to reading comprehension and describe the various kinds of questions one might use to more fully test a system's comprehension of a passage, moving beyond questions that only probe local predicate-argument structures. The main pitfall of this approach is that questions can easily have surface cues or other biases that allow a model to shortcut the intended reasoning process. We discuss ways proposed in current literature to mitigate these shortcuts, and we conclude with recommendations for future dataset collection efforts.

1 Introduction

Getting machines to "understand" natural language text is a vast and long-standing problem, made more challenging by the fact that it is not even clear what it means to understand text, or how to judge whether a machine has achieved success at this task. Much recent research in the natural language processing community has converged on an approach to this problem called *machine reading comprehension*, where a system is given a passage of text and a natural language question that presumably requires some level of "understanding" of the passage in order to answer. While there have been many papers in the last few years studying this basic problem, as far as we are aware, there is no paper formally justifying this approach to "understanding", or discussing its drawbacks.[1]

In this work we aim to motivate question answering as a good, but potentially fraught, means of measuring a machine's comprehension of natural language text. We argue that current reading comprehension datasets, largely inspired by the Stanford Question Answering Dataset (Rajpurkar et al., 2016, SQuAD),[2] are a good start at measuring reading comprehension, but do not go far enough in probing systems' understanding capabilities. Most of these datasets simply require a basic understanding of local predicate-argument structure and entity typing; there is a lot more to understanding text than that, such as tracking entities through a discourse, understanding the implications of text that is read, and recovering the underlying world model that the author intended to convey.

Question answering is a natural format to use when probing these complex phenomena, but it comes with inherent challenges. In particular, it is very easy to write questions that seem like they require deep understanding of text to answer, but in fact give lexical or other cues to a machine that allow the system to bypass the intended reasoning when answering the question. When constructing reading comprehension datasets, it is essential to deal with this issue up front, designing mechanisms in the data collection process that combat these shortcuts. We give many examples of both the shortcuts themselves and methods people have used to mitigate them, such as having mismatched questions and passages, including "no answer" as a possible answer option, and creating adversarial

[1] Richardson et al. (2013) give a good overview of the early history of this approach, but provide only very little justification.

[2] Though SQuAD was not nearly the first reading comprehension dataset, its introduction of the span extraction format was innovative and useful, and most new datasets follow its design.

Proceedings of the Second Workshop on Machine Reading for Question Answering, pages 105–112
Hong Kong, China, November 4, 2019.

examples, among others.

We conclude with a discussion about gaps we see in the literature that should be addressed by future dataset collection efforts.

2 Defining Reading Comprehension

How does one define "understanding a passage of text"? The process which a human uses to recover some notion of meaning when reading a passage is not well understood computationally (Kendeou and Trevors, 2012), so while this would be an ideal benchmark for machine understanding, it is unavailable to us. The natural language processing community has long drawn on linguistic formalisms to represent pieces of this meaning, from syntax trees and word sense disambiguation to semantic roles and coreference resolution. These formalisms only take us so far, however, as there is no linguistic formalism that satisfactorily captures the full meaning of a paragraph.

Instead we turn to ideas that go back at least to Alan Turing's test for machine intelligence (Turing, 1950; Levesque, 2013)—it is through interacting in natural language that an entity can demonstrate their understanding of language. We begin with a postulate: **an entity (human or machine) *understands* a passage of text if it can correctly answer *arbitrary questions* about that text**. We claim that this is a *sufficient* condition for understanding, but not a *necessary* one; there are surely other ways of demonstrating understanding.

Following this postulate, we define *machine reading comprehension* to be a task aimed at understanding a single coherent passage of text, where a system is given a single passage and a single question about that passage, and must produce an answer. Our definition of "single coherent passage" is somewhat loose; we consider anything longer than, e.g., a typical Wikipedia page to be too long and not a single coherent passage, while single sentences are generally too short. This means that, while they are certainly relevant, we are not including in this strict definition tasks that involve retrieving paragraphs or answering multiple consecutive questions, as they require additional capabilities. The boundaries around "reading comprehension" and which capabilities are related to "reading" or something else are very fuzzy, however, as we will see throughout the rest of this paper. In order to talk formally about the problem, we must pick a concrete definition, and so this is the definition we choose, while admitting that it is not perfect.

Using natural language questions to test comprehension of natural language text seems like an obvious choice: the meaning of arbitrary open-domain text goes beyond any possible formalism. There are various attempts, such as open information extraction (Etzioni et al., 2011) and abstract meaning representations (Banarescu et al., 2013), to try to capture broad, open domain semantics and the meaning of entire sentences. However, leaving aside the difficulties in training annotators and collecting annotations for these formalisms, any attempt to normalize meaning across disparate surface forms will necessarily lose information that was present in the natural language. The flexibility inherent in natural language as an *annotation* and *query* format is necessary in order to test deep understanding of arbitrary passages.

However, using questions to judge understanding is itself somewhat problematic, as (1) it is not clear a priori what the scope of these questions should be, and (2) collecting these arbitrary questions is very challenging, as questions that seem to be probing a particular kind of understanding might have shortcuts that allow answering them correctly without actually understanding the text at the level that was intended. Section 3 explores the first of these issues, and Section 4 discusses the second, along with ways to mitigate it.

3 What kinds of questions?

Having claimed that the ability to answer arbitrary questions is a natural way for machines to demonstrate understanding of a passage of text, we turn to the obvious question: what exactly is included in "arbitrary questions"? Some questions one could ask about a passage have little to do with understanding the passage. For example, the question *What is the population of the country Trump visited?*, asked about a passage that mentions the country but not its population, does require understanding the passage, but also requires knowing an additional specific fact. Such a requirement of external background knowledge not relevant to the passage is not desirable in a test of reading comprehension.

In this section we attempt to enumerate the high-level phenomena that characterize the understanding of a passage of text, and which can be asked about in reading comprehension ques-

tions. This enumeration is by no means exhaustive, but it should be a decent starting place for researchers attempting to build reading comprehension datasets—very few of these phenomena are explicitly queried in existing reading comprehension datasets, and those that are have relatively little coverage. We implicitly assume that the number of high-level phenomena is small enough such that making headway on, say, a few dozen phenomena will substantially improve the ability of models to read and understand text.

There are fuzzy boundaries between all of these phenomena, and no dataset can possibly focus exclusively on one of them. Every dataset, even those that sample from naturally occurring questions, will have some bias in which phenomena are asked about. We advocate being intentional about this bias and trying to be comprehensive in the collection of datasets that we construct.

Sentence-level linguistic structure Most existing reading comprehension datasets implicitly target local predicate-argument structures. The incentives involved in the creation of SQuAD encouraged workers to create questions that were close paraphrases of some part of a paragraph, replacing a noun phrase with a question word. This, and other cloze-style question construction, encourages very local reasoning that amounts to finding and then understanding the argument structure of a single sentence. This is an important aspect of meaning, but one could construct much harder datasets than this. One direction to push on linguistic structure is to move beyond locating a single sentence. DROP (Dua et al., 2019) largely involves the same level of linguistic structural analysis as SQuAD, but the questions require combining pieces from several parts of the passage, forcing a more comprehensive analysis of the passage contents. A separate direction one could push on sentence-level linguistic structure in reading comprehension is to target other phenomena than predicate argument structure. There are many rich problems in semantic analysis, such as negation scope, distributive vs. non-distributive coordination, factuality, deixis, briding and empty elements, preposition senses, noun compounds, and many more. Many of these phenomena have well-defined formalisms that can be used for annotation and evaluation, but it would also be useful to carefully design reading comprehension datasets that require an implicit understanding of these phenomena.

Paragraph-level structure While the input to a reading comprehension dataset is a paragraph of text, most datasets do not explicitly target questions that require understanding the entire paragraph, or how the sentences fit together into a coherent whole. Some post-hoc analyses attempt to reveal the percentage of questions that require more than one sentence, but it is better to design the datasets from the beginning to obtain questions that look at paragraph- or discourse-level phenomena, such as entity tracking, discourse relations, or pragmatics. For example, Quoref (Dasigi et al., 2019) is a dataset that targets entity tracking and coreference resolution. There are few linguistic formalisms targeting structures larger than a paragraph, but those that do exist, such as rhetorical structure theory (Mann and Thompson, 1988), could form the basis of an interesting and useful reading comprehension dataset.

Grounding and background knowledge A key aspect of reading is understanding the text in terms of what you already know, either commonsense knowledge or more domain-specific factual knowledge. After reading a description of a room, for example, people can make commonsense inferences about the objects described, and a lot of training and background knowledge is required to really understand an abstract on PubMed. People exhibit varying levels of comprehension when reading a particular text, depending largely on their ability to situate that text in the context of the appropriate background knowledge. There is room for interesting datasets along these lines. Cosmos QA (Huang et al., 2019) is an attempt to make such a dataset, though the fact that it is multiple choice puts it outside of our strict definition of "reading comprehension".

Implicative reasoning Understanding text includes understanding the implications (or entailments) of that text on other text that might be seen. For example, understanding the text *Bill loves Mary. Mary was just diagnosed with cancer.* means also understanding that Bill will be sad. In some sense this can be seen as "grounding" the predicates in the text to some prior knowledge that includes the implications of that predicate, but it also includes the more general notion of reconstructing a model of the world being described by the text. There are two datasets that just scratch

the surface of this kind of reading: ShARC (Saeidi et al., 2018) requires reading rules and applying them to questions asked by users, though its format is not standard reading comprehension; and ROPES (Lin et al., 2019), which requires reading descriptions of causes and effects and applying them to situated questions.

Communicative aspects There are many communicative aspects of text that a human implicitly understands when reading, and which could be queried in reading comprehension datasets. For instance, is a text intended to be expository, narrative, persuasive, or something else? Did the author succeed in their communicative intent? Was there some deeper metaphorical point in the text? A dataset targeted at these sorts of phenomena could be incredibly interesting, and very challenging.

4 Ways to combat shortcuts

As discussed in the previous section, large-scale reading comprehension datasets where crowdworkers ask questions about the given passage have brought significant progress in the community. However, it is very easy to construct datasets where solving the task contributes little to genuine understanding of the text as intended. Chen et al. (2016) argues that 97% of answerable questions on CNNDAILYMAIL (Hermann et al., 2015) are solvable by superficial clues such as word or semantic overlap.[3] Jia and Liang (2017) find that models trained on SQUAD suffer significantly when adversarial input is injected despite no change in the semantics of the original text. Such findings indicate that there are certain shortcuts in solving reading comprehension tasks that allow a model to find the answer by superficial clues such as lexical overlap and entity types (Clark and Gardner, 2018; Sugawara et al., 2018). Accordingly, more recent reading comprehension datasets are constructed with several different approaches to prevent such shortcuts in order to foster natural language understanding.

4.1 Question / passage mismatch

One way to reduce lexical overlap between the question and passage is to expose the author of the question to a different passage that conveys a similar meaning. Examples include NARRATIVEQA (Kočiskỳ et al., 2018), where question authors were shown a summary of a movie script that will be used for answering questions, and DUORC (Saha et al., 2018), where questions are authored given a passage that is comparable to the one that will later be employed.

Another approach is to collect questions first, and then pair them with a passage, which was done in QUAC (Choi et al., 2018) or with a distantly collected relevant context, which was the method of choice in TRIVIAQA (Joshi et al., 2017).

Last, lexical overlap can be reduced if one has access to natural questions that have been posed by users who do not know the answers and are seeking information (Lee et al., 2019). NATURAL QUESTIONS (Kwiatkowski et al., 2019) and BOOLQ (Clark et al., 2019) are two examples for such datasets. However, access to such questions is usually limited for most researchers.

4.2 "No answer" option

Most of the reasoning shortcuts in existing datasets arise due to the fact that the system can assume that the answer is guaranteed to exist in the given passage. Removing this assumption and requiring the system to identify whether the question is even answerable from the passage can prevent such shortcuts.

One example of this kind of dataset construction is SQUAD 2.0(Rajpurkar et al., 2018), which asked annotators to read the given passage and write a question which the passage does not contain the answer to but contains a plausible negative answer. A drawback of this approach is that annotators see the passage when asking the question, which can introduce its own biases and shortcuts. An alternative is to combine a "no answer" option with the approach the previous section, where an annotator writes questions without knowing the answer, and another annotator verifies whether they are answerable by the paired passage. Example datasets include NEWSQA (Trischler et al., 2016)[4], QUAC (Choi et al., 2018) and NATURAL QUESTIONS (Kwiatkowski et al., 2019).

4.3 Dialog

Questions that require additional context to be understood, such as conversation state, are another

[3] They found 75% of questions are answerable, and among them, 73% are solvable by exact match, paragraph and partial clues (word/concept overlap).

[4] Non-answerable questions are provided as the extra challenge apart from answerable portions.

potential means of avoiding reasoning shortcuts. A person is not able to answer a simple question such as *How many?* without the additional context of a prior question describing what is being counted. Care needs to be taken with this method, however, as some datasets are amenable to input reduction (Feng et al., 2018), where there is only one plausible answer to such a short question. If done well, however, this method provides additional challenges such as clarification, coreference resolution, and aggregation of pieces scattered across conversation history. QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019) are two datasets that focus on such setting.

4.4 Complex reasoning

Tasks which require more advanced forms of reasoning are proposed to prevent answering the question from superficial clues. Examples include tasks requiring discrete and arithmetic reasoning (Dua et al., 2019), textbook question answering which requires understanding various forms of knowledge (Clark et al., 2018; Kembhavi et al., 2017) and multi-hop question answering which requires reading multiple distinct pieces of evidence (Talmor and Berant, 2018; Yang et al., 2018). Despite these attempts, it was found that shortcuts still exist in complex reasoning tasks such as multi-hop QA (Min et al., 2019; Jiang and Bansal, 2019), so careful construction of the dataset is necessary.

One novel method that may by applied to combat such shortcuts and enforce multi-hop reasoning is to check the semantic relations present in the question. In questions requiring a conjunction to be performed, functional or pseudo functional relations (Lin et al., 2010), such as *father* or *founder*, may facilitate arriving at the correct answer by solving only the functional relation and not the full conjunction. On the other hand such relations are desired when requiring a composition to be solved in a question. For example, in the question *What is the capital of the largest economy in Europe?* we would like *the largest economy in Europe* to be one answer we can use to modify the question to *what is the capital of* ***Germany***.

4.5 Context construction

Shortcuts in solving a reading comprehension questions may also occur when the context is not diverse with respect to the question. (Min et al., 2019) Functional relations and entity types in the question can give away the location of the correct answer when only one such function relation or entity type exists in the context. For instance when asked *What year... ?* having only one available year in the context enable models to easily locate the correct answer, without requiring the rest of the question. One option to avoid these shortcuts is to carefully select or construct the contexts that are used, and various methods of entity and relation type counting in the context may be employed.

4.6 Adversarial construction

One promising means of removing reasoning shortcuts is to encode those shortcuts into a learned system, and use that system to filter out questions that are too easy during dataset construction. DROP (Dua et al., 2019) and Quoref (Dasigi et al., 2019) used a model trained on SQuAD 1.1 (Rajpurkar et al., 2016) as an "adversarial" baseline when having crowd workers write questions. Because the people could see when the system answered their questions correctly, they learned to ask harder questions.

This kind of adversarial construction can introduce its own biases, however, especially if the questions being filtered are generated by machines instead of humans (Zellers et al., 2018). This also makes a dataset dependent on another dataset and model in complex ways, which has both positive and negative aspects to it. In some sense, it is a good thing to get a diverse set of reading comprehension questions, and encoding one dataset's biases into a model to enforce a different distribution for new datasets helps in collecting diverse datasets. If crowd workers end up simply wordsmithing their questions in order to pass the adversary, however, this seems unsatisfying. Overall, however, we believe this is a good method that could be used more widely when collecting reading comprehension datasets.

4.7 Minimal question pairs

ROPES (Lin et al., 2019) borrowed the idea of "minimal pairs" from linguistic analysis in its construction. In order to avoid subtle biases around which entity appears first in a question or paragraph, or simple lexical association biases between question and passage words, crowd workers were instructed to make minimal changes to the questions they wrote in order to change the answer. For example, a question such as *Which city would have more trees?* might be changed to

Which city would have fewer trees?. This method is not applicable in all reading comprehension scenarios, but where it is it can be an effective means of reducing shortcuts—a single question in isolation might exhibit the characteristics of a shortcut, but presumably the other question in the minimal pair would *also* have the same shortcut, leading to a system that relies on the shortcut getting at least one of them wrong.

4.8 Free-form answers

Shortcuts almost always arise because of a limited output space that can be searched over to find simple biases that lead to the correct answer. The problem is largely, though not entirely, with multiple choice answers. This includes span extraction formats, which is still effectively multiple choice with on the order of 100 choices (or many fewer, if the system can reasonably model likely answer candidates from the passage). Requiring free-form answers, especially where the answer is not found in the paragraph, would dramatically reduce the occurrence of reasoning shortcuts. This introduces a separate problem of evaluating the free-form answers, however, which is a pressing problem in reading comprehension research. If we had a good means of automatically evaluating free-form answers, much of this section on designing datasets to avoid reasoning shortcuts would be unnecessary, and we could build much more interesting and challenging datasets.

4.9 Multi-task evaluation

Given the myriad datasets created for reading comprehension, a natural method to reduce the effects of shortcuts is to evaluate models on multiple datasets. Assuming shortcuts are often dataset-specific means that a model that succeeds on all datasets is likely to have better text understanding.

But evaluation on multiple datasets goes even beyond shortcut mitigation. In Section 3 we proposed to enumerate the phenomena required for reading comprehension and build datasets that highlight each category. A possible shortcoming of this approach is that researchers will develop models for specific datasets that do not generalize to other datasets. This will result in a collection of models, none of which fully understands text. Evaluating models on multiple reading comprehension datasets (Talmor and Berant, 2019) will ensure that progress is made towards comprehensive understanding of text.

4.10 Explainability

A possible way to reduce the effect of shortcuts is to demand some sort of explanation for the final answer provided by a reading comprehension model. In that vein, Yang et al. (2018) evaluate in HOTPOTQA not only QA accuracy but also whether the relevant supporting sentences are identified by a reading comprehension model.

5 Recommendations for future research

As evidenced by this survey, reading comprehension datasets have a long way to go before they approach a comprehensive test of a system's ability to read. Future datasets should try to improve coverage by focusing on phenomena that have been thus far neglected. Section 3 lists many possible phenomena that would make for very interesting reading comprehension datasets.

The challenge of creating a dataset without shortcuts has recently emerged as a fundamental one for progress in natural language understanding. Many datasets that have been created at great expense in an attempt to stress-test the abilities of existing models have been found to be simpler than expected due to the shortcuts that lie within them. Developing scientific methods for dataset collection that circumvent such shortcuts is instrumental for making sure the collective effort of our community actually leads to models that better understand text. For example, one possible method may be dropping out parts of a question as a means of insuring the question is not redundant and the model is not learning spurious shortcuts. Questions may be filtered using this technique, and models for shortcut checking may be trained on part of the questions to check if indeed no significant redundancy exists in them, and the model cannot solve the example with, say, only one word in the question (Feng et al., 2018).

In our opinion, evaluating reading comprehension models on many datasets is a promising direction that will prevent over-fitting to the statistical biases in a single dataset, but preventing bias *a priori*, as well as detecting bias and constructing adversarial examples are also important directions for future research.

References

L. Banarescu, C. B. S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and

N. Schneider. 2013. Abstract meaning representation for sembanking. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *ACL*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke S. Zettlemoyer. 2018. QuAC: Question answering in context. In *EMNLP*.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *ACL*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *NAACL-HLT*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Pradeep Dasigi, Nelson Liu, Ana Marasovic, Noah Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *EMNLP*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL*.

O. Etzioni, A. Fader, J. Christensen, S. Soderland, and Mausam. 2011. Open information extraction: the second generation. In *IJCAI*.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *EMNLP*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NeurIPS*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *EMNLP*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *ACL*.

Mandar S. Joshi, Eunsol Choi, Daniel S. Weld, and Luke S. Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*.

Panayiota Kendeou and Gregory Trevors. 2012. *Quality learning from texts we read: What does it take?*, pages 251–275. Cambridge University Press.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *TACL*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *TACL*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

H. J. Levesque. 2013. On our best behaviour. In *IJCAI*.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. *arXiv preprint arXiv:1908.05852*.

Thomas Lin, Mausam, and Oren Etzioni. 2010. Identifying Functional Relations in Web Text. In *EMNLP*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *ACL*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *TACL*.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *ACL*.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *NAACL-HLT*.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *ACL*.

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

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.