Prerequisites for Explainable Machine Reading Comprehension: A Position Paper

Saku Sugawara[1], **Pontus Stenetorp**[2], **and Akiko Aizawa**[1]
[1] National Institute of Informatics
[2] University College London
{saku,aizawa}@nii.ac.jp
p.stenetorp@cs.ucl.ac.uk

Abstract

Machine reading comprehension (MRC) has received considerable attention in natural language processing over the past few years. However, the conventional task design of MRC lacks the explainability beyond the model interpretation, i.e., the internal mechanics of the model cannot be explained in human terms. To this end, this position paper provides a theoretical basis for the design of MRC based on psychology and psychometrics and summarizes it in terms of the requirements for explainable MRC. We conclude that future datasets should (i) evaluate the capability of the model for constructing a coherent and grounded representation to understand context-dependent situations and (ii) ensure substantive validity by improving the question quality and by formulating a white-box task.

1 Introduction

Evaluation of natural language understanding (NLU) is a long-standing goal of artificial intelligence. Machine reading comprehension (MRC) is a task that tests the ability of a machine to read and understand unstructured text, and may be the most suitable task for evaluating NLU because of its general formulation (Chen, 2018). Recently, many large-scale datasets have been recently proposed, and neural network systems have achieved human-level performances in some of these datasets.

However, analytical studies have shown that MRC models do not necessarily provide human-level understanding. For example, Jia and Liang (2017) used manually crafted adversarial examples to show that successful systems are easily distracted. Sugawara et al. (2020) also showed that a significant part of already solved questions is solvable even after shuffling the words in a sentence or dropping content words, and the complex understanding of the given text is not necessary. These studies proved that we cannot *explain* what type of understanding is required by the datasets and is actually acquired by models. Although the explainability of MRC is related to the intent behind questions and is critical to understand the behavior of a model and test hypotheses for reading comprehension, its theoretical foundation is lacking in the existing literature.

In this position paper, we examine the requirements for the explainability of MRC through the following two questions: (i) What is the actual meaning of reading comprehension? (ii) How can we correctly evaluate the reading comprehension ability? Our motivation is to provide a theoretical basis for the task that can be relied on by those who create MRC datasets and analyze model behaviors. In the context of explainability, Gilpin et al. (2018) indicated that interpreting the internals of a system is closed to only that system's architecture and is insufficient for explaining how the task is accomplished. This is because even if we could interpret models' internals, we cannot explain what is measured by the datasets. Therefore, our focus in this study is the explainability of the task and datasets rather than the interpretability of models.

We first overview MRC and existing datasets in Section 2. We also review the analytical literature that indicates that existing datasets might fail to correctly evaluate their intended behavior. Then, we visit the psychological study of human reading comprehension in Section 3 for the *what* question (i). We argue that the concept of *representation levels* could be served as a conceptual hierarchy for organizing existing technologies in MRC. Next, in Section 4, we refer to the study of psychometrics to discuss what is necessary for the task design of MRC, answering the *how* question (ii). Our aim is to introduce the concept of *construct validity*, which emphasizes how we can

Question	Foundation	Requirements	Future direction
What is reading comprehension?	Representation levels in human reading comprehension: (A) surface structure, (B) textbase, and (C) situation model.	(A) Linguistic-level understanding, (B) comprehensiveness of skills for inter-sentence understanding, and (C) evaluation of coherent and grounded representation.	(C) Dependence of context on defeasibility and novelty, and grounding to non-textual information with a long passage.
How can we evaluate reading comprehension?	Construct validity in psychometrics: (1) content, (2) substantive, (3) structural, (4) generalizability, (5) external, and (6) consequential aspects.	(1) Covering skills comprehensively, (2) ensuring the evaluation of the internal process, (3) structured metrics, (4) reliability of metrics, (5) comparison with external variables, and (6) accountability and robustness to adversarial attacks.	(2) Improving the question quality by filtering and ablation, and designing a task for visualizing the internal process.

Table 1: Overview of theoretical foundations, requirements, and future directions of MRC discussed in this paper.

validate the interpretation of models' performance in the task. Finally, in Section 5, we discuss future directions in MRC. For the *what* part, we indicate that datasets should evaluate the capability of the *situation model*, which refers to a coherent, grounded representation constructed when humans understand texts. For the *how* part, we argue that we need to ensure that there is *substantive validity*, which necessitates the verification of the internal process of comprehension.

Table 1 provides an overview of the theoretical bases, requirements, and the future directions of MRC discussed in this paper. Our conclusions for further development of MRC are as follows.

- MRC could be the most suitable task for evaluating NLU. Focusing on the situation model is a next frontier for evaluating and achieving the human-level language understanding.
- We should ensure the substantive validity for the explainability of the internal process of NLU by improving the question quality and designing a white-box task formulation.

2 Task Overview

This section briefly overviews recent datasets from different viewpoints and describes analytic studies that revealed an issue of datasets' explainability for reading comprehension.

2.1 Task Variations and Existing Datasets

MRC is a task in which a machine is given a document (which we refer to as the *context*) and answers questions about it. As a general definition of MRC, Burges (2013) suggests that *a machine comprehends a passage of text if, for any question regarding that text that can be answered correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question*. In the following section, we describe variations of different task aspects along with representative datasets. We list the existing datasets in Appendix A.

Context styles. The form of a given context can be different in its length, for example, a single paragraph (Rajpurkar et al., 2016), a set of paragraphs (Yang et al., 2018), a longer document (Kočiský et al., 2018), or open domain (Chen et al., 2017). In some datasets, a context includes non-textual information such as images (Yagcioglu et al., 2018).

Question styles. A question can be a natural question sentence (in most datasets), a fill-in-blank sentence (cloze) (Lai et al., 2017; Xie et al., 2018), or semi-structured words (e.g., knowledge-base entries (Welbl et al., 2018) and search engine queries (Nguyen et al., 2016)).

Answering styles. An answer is (i) chosen from a text span of the given document (*answer extraction*) (Trischler et al., 2017), (ii) chosen from a candidate set of answers (*multiple choice*) (Richardson et al., 2013), or (iii) generated as a free-form text (*description*) (Kočiský et al., 2018). Some datasets optionally allow answering by a *yes*/*no* reply (Clark et al., 2019).

Sourcing methods. Initially, questions in small-scale datasets were created by experts (Sutcliffe et al., 2013). Later, fueled by the development of neural network models, most published datasets have more than a hundred thousand questions that have been automatically created (Hermann et al., 2015), crowdsourced (Rajpurkar et al., 2016), and collected from student exams (Lai et al., 2017).

Domains. The most popular domain seems to be Wikipedia articles (Kwiatkowski et al., 2019). In addition, news articles are often used (Hermann et al., 2015; Onishi et al., 2016). Lai et al. (2017) used English exams for middle and high school students, which covers various topics. Suster and Daelemans (2018) and Pampari et al. (2018) proposed their datasets in the clinical domain. Saha et al. (2018) and Kočiský et al. (2018) used movie scripts as the context documents.

Skill focuses. Recently proposed datasets seem to be specialized for requiring specific skills including unanswerable questions (Rajpurkar et al., 2018), dialogue (Choi et al., 2018; Reddy et al., 2019; Sun et al., 2019), multiple-sentence reasoning (Khashabi et al., 2018), multi-hop reasoning (Welbl et al., 2018; Yang et al., 2018), mathematical and set reasoning (Dua et al., 2019), commonsense reasoning (Huang et al., 2019), and coreference resolution (Dasigi et al., 2019).

2.2 Explanation Issues

In some datasets, machines' performance already reached the human level performance. However, Jia and Liang (2017) indicated that models are easily fooled by manual injection of distracting sentences. They highlighted that existing models do not necessarily understand given passages precisely. Although this does not mean that machine learning models cannot solve such adversarial questions even when these questions are given in their training (Liu et al., 2019b), their study revealed that questions simply gathered by crowdsourcing without careful guidelines or constraints are insufficient to evaluate precise language understanding.

This argument is supported by further findings on existing datasets. For example, Min et al. (2018) found that more than 90% of the questions in SQuAD (Rajpurkar et al., 2016) require obtaining an answer from a single sentence despite being provided with a passage. Sugawara et al. (2018) showed that large parts of 12 datasets were easily solved only by looking at a few first question tokens and attending the similarity between the given questions and the context. Similarly, Feng et al. (2018) and Mudrakarta et al. (2018) demonstrated that models do not change their predictions even when question tokens are partly dropped in SQuAD. Kaushik and Lipton (2018) also observed that question- and passage-only models often perform well. More recently, Sugawara et al. (2020) observed that already solved questions in existing datasets can be solved even after shuffling sentence words or dropping content words, which indicates that these questions do not necessarily require complex understanding of the given texts. Min et al. (2019) and Chen and Durrett (2019) concurrently indicated that for the multi-hop reasoning datasets, the questions are solvable only with a single paragraph and thus do not necessarily require multi-hop reasoning over multiple paragraphs. For commonsense reasoning, Zellers et al. (2019b) reported that their dataset unintentionally contains stylistic biases in the answer options, which made the dataset fall short of requiring commonsense reasoning. These biases were embedded by a language-based model that generated answer options, and thus made the dataset fall short of requiring commonsense reasoning.

Overall, these investigations highlight a serious issue with the task design. That is, even if models show human-level scores, we cannot conclude that they successfully perform human-level reading comprehension. We admit that this issue is due to the low interpretability of black-box neural network models which are currently prevalent. However, we emphasize the importance of the explainability because even if we could interpret models' internals, we cannot explain what is measured by the datasets. We conjecture that the explainability issue in MRC can be analyzed by the following two points; (i) we do not have a comprehensive theoretical basis for specifying what we should ask of reading comprehension (Section 3) and (ii) we do not have a well-established methodology for creating a dataset and validating a model's performance on it (Section 4). In the remainder of this paper, we argue that these issues can be addressed by using insights from the psychological study of reading comprehension and the study of the validity in psychometrics.

3 Reading Comprehension from Psychology to MRC

3.1 Computational Model in Psychology

In psychology, there is a long history of the study on human text comprehension (Kintsch and Rawson, 2005; Graesser et al., 1994; Kintsch, 1988). They proposed connectionist and computational architectures including a mechanism pertinent to knowledge activation and memory storing.

Among the computational models, we adopt the construction–integration (CI) model, which is the most influential and provides a foundation in the field (refer to McNamara and Magliano (2009) for a comprehensive review). The CI model assumes that text comprehension is achieved by the following two steps. (i) The *construction* step involves reading words at the surface level and constructing propositions where a proposition represents a predicate and its arguments that denote a described event, often elaborated by a reader's background knowledge. (ii) The *integration* step refers to the process of associating the propositions and creating a network of them. These steps are not exclusive, that is, propositions are iteratively updated in accordance with the surrounding propositions with which they are linked.

Besides, the CI model assumes that these processes involve processing at three different representation levels as follows.

- *Surface structure* is the linguistic information of particular words, phrases, and syntax obtained by decoding the raw textual input.
- *Textbase* is a set of propositions in the text where the propositions are locally connected by inferences (*microstructure*).
- *Situation model* is a situational, coherent mental representation covering where the propositions are globally connected (*macrostructure*) and it is often grounded to not only texts but also the sound, imagery, and personal information.

In summary, the CI model first decodes textual information (i.e., surface structure) from the raw textual input, then creates the propositions (i.e., textbase) and their local connections sometimes using the reader's knowledge, and finally constructs a coherent representation (i.e., situation model) that is coherently organized according to the five dimensions (space, causation, intentionality, objects, and time (Zwaan and Radvansky, 1998)) and globally explains the described events. Although a definition of successful reading comprehension can be different, Hernández-Orallo (2017) stated that the goal of text comprehension here is to create the situation model that best explains the given text and the reader's background knowledge. This definition also effectively explains that the situation model plays an important role in human reading comprehension.

Our aim in this section is to provide a basis for explaining what reading comprehension is, which needs *units* for the explanation (Doshi-Velez and Kim, 2018). In the computational model above, the levels of representations seem to be useful for organizing such units. Our goal in Section 3.2 is to ground existing natural language processing (NLP) technologies and tasks to the different representation levels.

3.2 Skill Hierarchy for MRC

In this section, we associate the existing NLP tasks with the three representation levels we introduced above. We consider that the biggest advantage of MRC is that it could be the most general task for evaluating NLU because of its general formulation. This emphasizes the importance of MRC comprehensively requiring various *skills*, which can be served as units for the explanation of reading comprehension. Therefore, our motivation is twofold: (i) to give an overview of them as a hierarchical taxonomy of *skills* and (ii) to highlight what is missing in existing MRC datasets for comprehensively covering these representation levels.

Existing taxonomies. To digest existing tasks and technologies, we first briefly overview existing taxonomies of *skills* in the context of NLU tasks. For recognizing textual entailment (Dagan et al., 2006), several studies classified types of reasoning and commonsense (Bentivogli et al., 2010; Sammons et al., 2010; LoBue and Yates, 2011). For science question answering (QA), Jansen et al. (2016) categorized knowledge and inference for an elementary-level dataset. Boratko et al. (2018) also similarly proposed types of knowledge and reasoning for science questions in MRC (Clark et al., 2018). A limitation of both studies is that proposed sets of knowledge and inference are specific to the elementary-level science domain. For MRC, although some of the existing datasets have their own classifications of skills, they are coarse and only cover a limited extent of typical processing in NLP (e.g., word matching and paraphrasing). Among them, multiple-sentence reasoning is too simplified for which there can be several types of sentence relations (Khashabi et al., 2018). In contrast, for more generalizable definitions, Sugawara et al. (2017) proposed a set of 13 skills for MRC. However, these skills are defined at a single level, which is not fully considered in multiple representation levels.

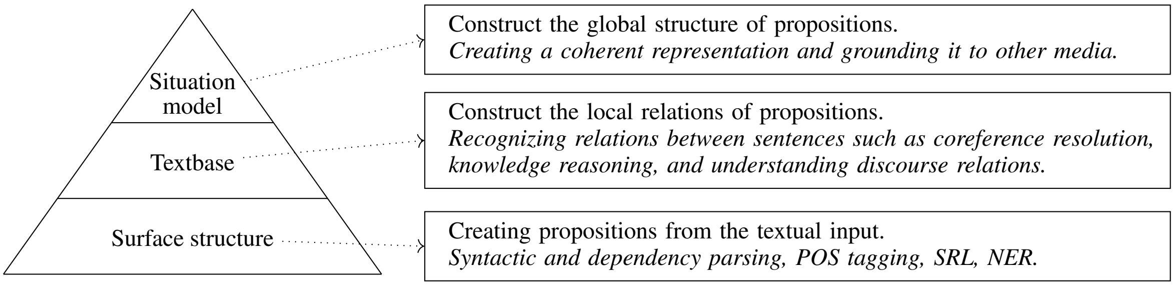
Situation model
Textbase
Surface structure
Construct the global structure of propositions.
Creating a coherent representation and grounding it to other media.
Construct the local relations of propositions.
Recognizing relations between sentences such as coreference resolution, knowledge reasoning, and understanding discourse relations.
Creating propositions from the textual input.
Syntactic and dependency parsing, POS tagging, SRL, NER.

Figure 1: Representation levels and corresponding skills.

As follows, we describe the three representation levels that basically follow the three representations of the CI model but are modified for MRC in Figure 1. We emphasize that we do not intend to create exhaustive and rigid definitions of skills. Rather, we aim to place them in a hierarchical organization as a foundation on which we can rely on and highlight what is missing in current MRC.

Surface structure. This level broadly covers the linguistic information and its semantic meaning that can be formed by the raw textual input. Although these features form a proposition in psychology, it seemingly should be viewed as sentence-level semantic representation in computational linguistics. This level includes part-of-speech tagging, syntactic parsing, dependency parsing, punctuation recognition, named entity recognition (NER), and semantic role labeling (SRL). Although recent pretraining-based neural language models can have the capability of these basic tasks (Liu et al., 2019a), these tasks are hardly required in NLU tasks including MRC. McCoy et al. (2019) indicated that the natural language inference (NLI) task (e.g., Bowman et al. (2015)) fails to ask the syntactic understanding of given sentences. White et al. (2017) and Kim et al. (2019) also proposed local-level tasks including probing tasks, as requiring sentence-level semantics and syntax. For MRC, Sugawara et al. (2020) also indicated that questions are solvable even after dropping function words. Although it is not obvious that we should include these basic tasks into MRC and it is not easy to circumscribe linguistic knowledge from concrete and abstract knowledge (cf., Zaenen et al. (2005) and Manning (2006)), we could say that we should always care about the capabilities of basic tasks when assessing a model's achievement (e.g., there may be adversarial noises that perturb the syntactic information of texts).

Textbase. This level covers local relations of propositions in the computational model of reading comprehension. In the context of NLP, it refers to various types of relations linked between sentences. These relations cover not only typical sentence relations (discourse relations), but also the linking between entities. As a result, this level includes coreference resolution, causality, temporal relations, spatial relations, text structuring relations, logical reasoning, knowledge reasoning, including bridging and elaboration (refer to McNamara and Magliano (2009) for their distinction), commonsense reasoning, mathematical reasoning, and logical reasoning. We also include multi-hop reasoning (Welbl et al., 2018) at this level because it does not necessarily require a coherent global representation over a given context. Although we do not intend to give comprehensive definitions of knowledge and commonsense types here, non-textual types of reasoning and knowledge are not included in this level. For example, Davis and Marcus (2015) indicate that physical reasoning (e.g., geometric reasoning) is one of the most difficult domains in commonsense reasoning. For the generalizability of MRC, Fisch et al. (2019) proposed a shared task featuring training and testing on multiple in/out domains. However, because requisite skills are not identified, the task still lacks explainability. Beyond a dataset focusing on a single skill, we should create a dataset in which the skills at this level are comprehensively identified.

Situation model. This level targets the global structure of propositions in human reading comprehension. It includes a coherent, situational representation of a given context and its grounding to the non-textual information. A coherent representation has well-organized sentence-to-sentence transitions (Barzilay and Lapata, 2008), which is also vital for using procedural and script knowledge. However, most existing MRC datasets fail to target the situation model for coherent understanding of given texts and grounding to non-textual information. We elaborate future directions of this level in Section 5.1.

In summary, we propose that the following features are missing in the current datasets:

- Caring about the capabilities of basic understanding of the linguistic-level information.
- Ensuring that questions comprehensively specify and evaluate textbase-level skills.
- Evaluating the capability of the situation model in which propositions are coherently organized and are grounded to non-textual information such as sound and imagery.

4 MRC on Psychometrics

In this section, we aim to give a theoretical foundation about how MRC models can be evaluated in an explainable way. A key concept is *validity*; given that MRC measures the capability of reading comprehension, validating the measurement is important to obtain reliable and useful explanation. Therefore, we visit psychometrics—a field of study concerned with methods used to evaluate the quality of psychological measurement (Furr, 2018). Our assumption is that we can make use of insights in psychometrics for a better task design as psychological experiments rely on psychometrics for the validation of measurement. In Section 4.1, we first overview the concept of validity in psychometrics. Among various definitions, we use the concept of *construct validity* proposed by Messick (1995), which is the most widely adopted definition in the field. Then in Section 4.2, we discuss what aspects correspond to construct validity in MRC and then indicate what we need to achieve for the verification of the intended explanation for MRC in its task design.

4.1 Construct Validity in Psychometrics

In psychometrics, *construct validity* refers to what is necessary to validate the interpretation of outcomes of psychological experiments.[1] According to Messick (1995), the construct validity consists of the following six aspects shown in Table 2.

In the design of educational and psychological measurement, these aspects are taken together and provide verification questions that need to be answered in justifying test scores' interpretation and use. In this sense, the construct validation can be seen as an empirical evaluation of the meaning and consequence of measurement in psychology. Given that MRC is intended to capture the capability of reading comprehension, those who design the task need to consider these validity aspects as much as possible. Otherwise, users of the task cannot justify the score interpretation; we cannot say that successful systems actually perform intended reading comprehension.

4.2 Construct Validity in MRC

In this section, we associate these aspects with MRC and discuss what we need to ensure for the validation of score interpretation in MRC. We summarize the six aspects of the construct validity and their corresponding MRC features in Table 2. As follows, we discuss what is missing to achieve the construct validity of the current MRC.

Content aspect. As we discussed in Section 3, sufficiently covering the skills across all the representation levels is an important requirement for MRC. In this sense, it is desirable that an MRC model is simultaneously evaluated on various skill-oriented datasets (e.g., multi-hop reasoning and commonsense reasoning) rather than different domains of corpus. As for the content aspect of the construct validity in MRC, there are two important points: *coverage* and *representativeness*.

Substantive aspect. This aspect appraises the evidence for the consistency of model behaviors. We consider that this aspect is the most important in evaluating reading comprehension, a process that subsumes various, implicit, and complex steps. To obtain a consistent response from an MRC system, which is important for the explainability, we somehow need to ensure that questions correctly assess the internal steps of the process of reading comprehension. However, as we viewed in Section 2.2, most current datasets fails to verify that a question is solved by using an intended skill, which fails to justify that a successful system can actually perform intended reading comprehension. We will further discuss how we can tackle this substantive aspect in Section 5.2.

Structural aspect. Another issue in most current datasets is that they only provide simple ac-

[1] A construct in psychology means an abstract concept used to facilitate understanding of human behavior, e.g., vocabulary, skills, and comprehension.

Validity aspects	Definition in psychometrics	Correspondence in reading comprehension
1. Content	Evidence of content relevance, representativeness, and technical quality.	Questions require reading comprehension skills with a sufficient *coverage* and *representativeness* over the representation levels.
2. Substantive	Theoretical rationales for the observed consistencies in the test responses including task performance of models.	Questions correctly evaluate the intended intermediate process of reading comprehension and provide rationales to the interpreters.
3. Structural	Fidelity of the scoring structure to the structure of the construct domain at issue.	Correspondence between the task structure and the score structure.
4. Generalizability	Extent to which score properties and interpretations can be generalized to and across population groups, settings, and tasks.	Reliability of test scores in correct answers and model predictions, and applicability to other situations.
5. External	Convergent and discriminant evidence from multitrait-multimethod comparisons as well as evidence of criterion relevance and applied utility.	Comparison of the performance of a task with that of other tasks and measurements.
6. Consequential	Value implications of score interpretation as a basis for action as well as for the actual and potential consequences of test use, especially regarding the sources of invalidity related to issues of bias, fairness, and distributive justice.	Considering the model vulnerabilities to adversarial attacks and social biases of the model and the datasets to ensure the fairness of model outputs.

Table 2: Aspects of the construct validity in psychometrics and corresponding features in reading comprehension.

curacy as a metric. Given that the substantive aspect necessitates evaluating the internal process of reading comprehension, the structure of metrics needs to reflect it. However, there are only a few attempts for providing a dataset with multiple metrics. For example, QuAC (Choi et al., 2018), a dialogue-based dataset, introduced a metric for the percentage of dialogues for which a system correctly answers every question in the dialogue. If consecutive questions in a dialogue are mutually dependent, it seems that this metric can evaluate the understanding of a given dialogue within accompanying questions. Another example is HotpotQA (Yang et al., 2018), which asks not only for answers to questions but also an indication of the evidence sentences (*supporting facts*). This metric can also evaluate the process of multi-hop reasoning whenever understanding the supporting sentences is really required in answering a question. Therefore, we need to care about both substantive and structural aspects simultaneously.

Generalizability aspect. We can discuss the generalizability in MRC from two perspectives: (i) the reliability of metrics and (ii) the reproducibility of findings.

For (i), an issue for the reliability may happen in the context of the given correct answers and a model's predictions, respectively. On the side of the correct answers, the model performance and its interpretation become unreliable when correct answers are unintentionally ambiguous or unanswerable. When sourcing a dataset, there could be unintentionally ambiguous or unanswerable questions. Because in most datasets the correct answers are just decided by a majority vote of crowd workers, it does not take the ambiguity of the answers into account. It might be useful to have such ambiguity reflected in the evaluation metrics (e.g., using the item response theory for RTE (Lalor et al., 2016)). On the side of a system's predictions, an issue is the reproducibility of results (Bouthillier et al., 2019), which means that a reimplementation of the system generates statistically similar predictions. As Dror et al. (2018) pointed out, it is rarely confirmed that produced results are statistically significant in NLP. For the reproducibility of models, we should use statistical testing methods in evaluating MRC models.

For (ii), Bouthillier et al. (2019) stressed the reproducibility of findings, that is, transferability of findings in a dataset to another dataset. In other words, there should be some units for the explanation that both datasets have in common. Such units are called *cognitive chunks* by Doshi-Velez and Kim (2018) in the context of the explainability of machine learning models. This generalizability aspect therefore highlights the importance of the content aspect.

External aspect. Although this aspect is important in psychometrics, it might be less important in

MRC because of the difference in their purposes (psychological measurement versus the development of systems). Nonetheless, to develop a general NLU system, it is necessary that we need to evaluate it on various NLU tasks such as not only MRC but also NLI, dialogue, and visual question answering. In addition, it is also necessary to associate the performance in MRC to other external measures such as the vocabulary size, problem-solving time, and memory consumption.

Consequential aspect. This aspect highlights the actual and potential consequences of test use. In MRC, this refers to using a successful model in actual situations other than tasks. Wallace et al. (2019) showed that existing NLP models have vulnerabilities to adversarial examples and thereby generate egregious outputs. We need to care about model robustness to adversarial attacks and accountability for unintended model behaviors.

5 Future Directions

This section discusses future directions of MRC in terms of *what* and *how* as introduced in Sections 3 and 4. In particular, the situation model and the substantive validity are considered as critical for developing human-level explainable MRC.

5.1 What side: Evaluating Situation Model

As we viewed in Section 3, existing datasets fail to assess the situation model in reading comprehension. For future directions, we indicate that the task should deal with two features of the situation model, namely, context dependency and grounding to non-textual information.

5.1.1 Context-dependent Situations

One of the vital features of the situation model is that it is conditioned on a given text. That is, a representation is constructed differently depending on the given context. In this paper, we call this property *context dependency*. We elaborate it by discussing the following two important features: defeasibility and novelty.

Defeasibility. The defeasibility of a constructed representation means that a reader can modify and revise it according to the information newly observed (Davis and Marcus, 2015; Schubert, 2015). Although the defeasibility in NLU is tackled in tasks of if-then reasoning (Sap et al., 2019), abductive reasoning (Bhagavatula et al., 2019), and counterfactual reasoning (Qin et al., 2019), there have been few attempts in MRC.

Novelty. An example showing the importance of contextual novelty is *Could a crocodile run a steeplechase?* by Levesque (2014). This question poses a novel situation where the answerer needs to combine multiple commonsense knowledge together to derive the correct reasoning. Such a novel situation seems to appear more easily in a longer MRC document rather than in a short sentence of NLI. Using only non-fiction documents such as newspaper and Wikipedia articles, some questions possibly just require reasoning of facts already known in web-based corpus and do not require novel reasoning. Therefore, fictional narratives would be a better source for creating a dataset of novel questions.

On a slide note, the dialogue-style MRC could enhance the context dependency in reading comprehension. Chiang et al. (2020) indicated that recent dialogue-based datasets may fail to evaluate a precise understanding of conversations beyond simple QA. This may be because the datasets do not evaluate the question-to-context dependency including the question history (See also Section 5.2.2). While the process of reading comprehension is assumed to be static in the current MRC, context-dependent situations need to be evaluated in a dynamic context, where a question triggers to update a given context, and the subsequent question requires an understanding of that update. Examples of the context would include users' intentions, non-textual worlds, and databases.

5.1.2 Grounding to Other Media

There are only a few MRC datasets for grounding texts to non-textual information. For example, Kembhavi et al. (2017) proposed a multiple-choice dataset on science textbooks which has questions with passages, diagrams, and images. Kahou et al. (2018) also proposed a figure-based QA dataset that requires understanding of figures including line plots and bar charts. Another approach is visual question answering (Antol et al., 2015) and visual commonsense reasoning (Zellers et al., 2019a) tasks. However, these approaches seem to have the following issues for evaluating language understanding deeply: (i) skills required for answering questions seem not to be identified; (ii) proposed models are likely to be domain- and task- specific, which lacks generalizability to other domains and tasks; and (iii) most datasets do not have long descriptions but short questions about images, which may cause flaws in evaluating precise understanding of given texts. Therefore, it might be important to create questions that, as an extension of MRC, have longer texts as a context and require understanding of the given texts by choosing correct images or their parts (refer to Kintsch and Rawson (2005) for an example of the relation between a situation model and a depiction).

5.2 How side: Assuring Substantive Validity

The substantive validity requires ensuring that questions correctly assess the internal steps of reading comprehension (Section 4). Then, our question is how we can assure the substantive validity of MRC datasets and the explanation to provide. We discuss two approaches for this challenge: creating the *high-quality questions* and designing a *white-box task formulation*.

5.2.1 Collecting High-quality Questions

As Gururangan et al. (2018) revealed, NLU datasets may contain unintended biases embedded by annotators (*annotation artifacts*). If machine learning models exploit such biases for answering questions, we cannot evaluate models' precise language understanding. Therefore, we need to alleviate such biases by filtering out undesirable questions. Besides, for the explainability of MRC, we also need to identify what skills are required for answering questions. We introduce two directions: *removing unintended biases by filtering* and *identifying requisite skills by ablating input features*.

Removing unintended biases by filtering. Zellers et al. (2018) proposed a model-based adversarial filtering method that iteratively trains an ensemble of stylistic classifiers and uses them to filter questions out. Sakaguchi et al. (2020) also proposed filtering methods both by machines and humans to alleviate *dataset-specific* and *word-association* biases to create Winograd-schema questions (Levesque, 2011). A problem here is that we cannot truly distinguish knowledge from bias in a closed domain. When the domain is equal to a dataset, patterns that are true only in the domain are called *dataset-specific* biases (or annotation artifacts in the labeled data). When the domain covers larger corpora, the patterns (e.g., frequency) are called *word-association* biases. When the domain is our everyday experience, patterns are called *commonsense*. However, as we mentioned in Section 5.1, a certain type of commonsense is *defeasible*. This means that such knowledge can be false in unusual situations. Another type of commonsense is called the law of nature, which can be false in other distant possible worlds. Besides, when the domain is our real possible world, indefeasible patterns are called *factual knowledge*.

Therefore, the distinction of bias and knowledge depends on where we recognize that pattern. This means that a dataset should be created so that it can test reasoning on an intended kind of knowledge. For example, when we test defeasible reasoning, we have to filter out questions that are solvable only by usual commonsense. If we want to determine the reading comprehension ability independently from factual knowledge, we may have to ask them in counterfactual or fictional situations. This also supports the importance of testing the situation model as we discussed in Section 5.1.

Identifying requisite skills by ablating input features. Another approach is to verify the quality of questions by checking the human answerability of questions after ablating important features from them; our intuition is that, if a question is still answerable by humans even after removing the features, the question does not require understanding of ablated features as Sugawara et al. (2020) similarly pointed out for using machines. This method can be used for verifying that intended features are required for answering questions (e.g., checking the necessity of resolving pronoun coreference after replacing pronouns with dummy nouns). Although it is not easy to identify *necessary features* and this method is quite labor-intensive, the explainability needs to indicate textual features associated with certain skills as units for the explanation. In addition, answering a question is equal to choosing the correct answer from among the candidate answers. Necessary features are, therefore, *necessary* for discriminating between different but semantically similar candidate answers (Khashabi, 2019). In summary, the task design should take care of collecting these similar candidates while identifying critical features.

Because what kinds of skills we should organize might be a pragmatic problem, there is no definite answer. For practical use, those who develop a task need to invent a set of skills that is at least necessary to explain how the task works. Although the skill definition depends on the task and its purpose, it should be intuitive for explaining the internal processing. On the other hand, for the scientific study of language understanding, researchers may need to achieve some extent of agreement on what kinds of skills reading comprehension consists of. This agreement would be necessary to mutually understand subjects that researchers try to hypothesize and verify. Concretely, such skills may be derived from existing NLP tasks (e.g., parsing, tagging, commonsense reasoning, and discourse understanding). They may also need to be associated with psychological and cognitive theories of human text comprehension.

5.2.2 Designing White-box Task Formulation

Another approach for ensuring the substantive validity is to make the explanation in the task formulation explicit. We introduce two directions: (i) generating the introspective explanation and (ii) creating dependency between questions.

Generating the introspective explanation. Inoue et al. (2019) classified two types of explanation in the text comprehension; *justification explanation* and *introspective explanation*. while the *justification explanation* only provides a collection of supporting facts for making a certain decision, the *introspective explanation* provides a derivation for making the decision. Inoue et al. (2019) annotated the introspective explanation with multi-hop reasoning questions and proposed a task that required generating the derivation of the correct answer of a given question to improve the explainability. Similarly, Rajani et al. (2019) collect human explanations for commonsense reasoning and use them to improve a system's performance through modeling the generation of the explanation. Although gathering human explanations is costly, these approaches can enable us to verify a model's understanding in an explicit way.

Making the question dependency. Another approach for improving the substantive validity in the task formulation is to create dependency between questions. For example, Dalvi et al. (2018) proposed a dataset that requires a procedural understanding of science facts. In the dataset, a set of questions corresponds to the steps of the whole process of a science fact. Therefore, that set as a whole can be seen as a single question that requires understanding the process of that science fact. Yagcioglu et al. (2018) also proposed a dataset in the recipe domain in which a few types of questions required an understanding of cooking procedures, by choosing the correct order of the images to make a complete recipe. Dialogue-based datasets also have questions that are mutually dependent. However, one issue with such questions is that relations between questions are not identified. These approaches enables us to explicitly verify a model's understanding.

6 Conclusion

In this position paper, we overviewed issues and future directions of MRC. We focused specifically on the situation model in psychology for *what* we should ask of reading comprehension and the substantive validity in psychometrics for *how* we should correctly evaluate it. We conclude that future datasets should (i) evaluate the capability of the situation model for understanding context-dependent situations and for grounding to non-textual information and (ii) ensure the substantive validity by improving the question quality and designing a white-box task formulation.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2010. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 60–70. Association for Computational Linguistics.

Xavier Bouthillier, César Laurent, and Pascal Vincent. 2019. Unreproducible research is reproducible. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 725–734, Long Beach, California, USA. PMLR.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Christopher J.C. Burges. 2013. Towards the machine comprehension of text: An essay. Technical report, Microsoft Research Technical Report MSR-TR-2013-125.

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879. Association for Computational Linguistics.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4026–4032, Minneapolis, Minnesota. Association for Computational Linguistics.

Ting-Rui Chiang, Hao-Tong Ye, and Yun-Nung Chen. 2020. An empirical study of content understanding in conversational question answering. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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. *CoRR*, abs/1803.05457.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. 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 1595–1604. Association for Computational Linguistics.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5927–5934, Hong Kong, China. Association for Computational Linguistics.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Bhuwan Dhingra, Kathryn Mazaitis, and William W. Cohen. 2017. Quasar: Datasets for question answering by search and reading.

Finale Doshi-Velez and Been Kim. 2018. *Considerations for Evaluation and Generalization in Interpretable Machine Learning*, 1st edition. Springer International Publishing.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&A dataset augmented with context from a search engine.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

R Michael Furr. 2018. *Psychometrics: an introduction*. Sage Publications.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Arthur C. Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. 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 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

José Hernández-Orallo. 2017. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A novel large-scale language understanding task over wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545, Berlin, Germany. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *International Conference on Learning Representations*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2019. RC-QED: Evaluating natural language derivations in multi-hop reading comprehension.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, Osaka, Japan. The COLING 2016 Organizing Committee.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages

2011–2021. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611. Association for Computational Linguistics.

Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An annotated figure dataset for visual reasoning.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015. Association for Computational Linguistics.

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 *the IEEE Conference on Computer Vision and Pattern Recognition*.

Daniel Khashabi. 2019. *Reasoning-Driven Question-Answering for Natural Language Understanding*. Ph.D. thesis, University of Pennsylvania.

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. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review*, 95(2):163.

Walter Kintsch and Katherine A Rawson. 2005. Comprehension. *The Science of Reading: A Handbook*, pages 211–226.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Igor Labutov, Bishan Yang, Anusha Prakash, and Amos Azaria. 2018. Multi-relational question answering from narratives: Machine reading and reasoning in simulated worlds. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 833–844. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 796–805. Association for Computational Linguistics.

John Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

Hector J. Levesque. 2011. The winograd schema challenge. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*.

Hector J. Levesque. 2014. On our best behaviour. *Artificial Intelligence*, 212:27 – 35.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019b. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. 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 2039–2048, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher D. Manning. 2006. Local textual inference: It's hard to circumscribe, but you know it when you see it—and NLP needs it. Unpublished manuscript.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Danielle S McNamara and Joe Magliano. 2009. Toward a comprehensive model of comprehension. *Psychology of learning and motivation*, 51:297–384.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9):741.

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.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1725–1735. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51. Association for Computational Linguistics.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906. Association for Computational Linguistics.

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. *CoRR*, abs/1611.09268.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did What: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235. Association for Computational Linguistics.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117, Minneapolis, Minnesota. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset:

Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

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

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Matthew Richardson, Christopher J.C. 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. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WINOGRANDE: an adversarial winograd schema challenge at scale. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. 2010. "Ask not what textual entailment can do for you...". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1199–1208. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Lenhart K Schubert. 2015. What kinds of knowledge are needed for genuine understanding? In *IJCAI 2015 Workshop on Cognitive Knowledge Acquisition and Applications (Cognitum 2015)*.

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. Association for Computational Linguistics.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 806–817. Association for Computational Linguistics.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Simon Suster and Walter Daelemans. 2018. CliCR: a dataset of clinical case reports for machine reading comprehension. 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 1551–1563. Association for Computational Linguistics.

Richard Sutcliffe, Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2013. Overview of QA4MRE main task at CLEF 2013. *Working Notes, CLEF*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

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

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards AI-complete question answering: a set of prerequisite toy tasks. In *International Conference on Learning Representations*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, Michigan. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension.

Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162.

A Machine Reading Comprehension Datasets

This appendix lists existing machine reading comprehension datasets along with their answer styles, dataset size, type of corpus, sourcing methods, and focuses.

Name	Ans	Size	Corpus	Src	Focus
QA4MRE (Sutcliffe et al., 2013)	MC	240	technical document	X	exam-level questions
MCTest (Richardson et al., 2013)	MC	2.6K	written story	C	children-level narrative
bAbI (Weston et al., 2015)	Desc	10K * 20	generated text	A	toy tasks for prerequisite skills
CNN/ DailyMail (Hermann et al., 2015)	Ex	1.4M	news article	A	entity cloze
Children's Book Test (Hill et al., 2016)	Ex	688K	narrative	A	large-scale automated
SQuAD 1.1 (Rajpurkar et al., 2016)	Ex	100K	Wikipedia	C	large-scale crowdsourced
LAMBADA (Paperno et al., 2016)	Desc	10K	narrative	C	hard language modeling
WikiReading (Hewlett et al., 2016)	Desc	18M	Wikipedia	A	super large-scale dataset
Who did What (Onishi et al., 2016)	MC	200K	news article	A	cloze of person name
MS MARCO (Nguyen et al., 2016)	Desc	100K	web snippet	Q	description on web snippets
NewsQA (Trischler et al., 2017)	Ex	120K	news article	C	blindly created questions
SearchQA (Dunn et al., 2017)	Ex	140K	web snippet	C/X	49.6 snippets on average
RACE (Lai et al., 2017)	MC	100K	language exam	X	middle/high school English exam in China
Story Cloze Test (Mostafazadeh et al., 2017)	MC	3.7K	written story	C	98,159 stories for training
TriviaQA (Joshi et al., 2017)	Ex	650K	web snippet	C/X	trivia questions
Quasar (Dhingra et al., 2017)	Ex	80K	web snippet	Q	search queries
TextbookQA (Kembhavi et al., 2017)	MC	26K	textbook	X	with figures
AddSent SQuAD (Jia and Liang, 2017)	Ex	3.6K	Wikipedia	C	distracting sentences injected

Table 3: Machine reading comprehension datasets published before 2017. *Ans* denotes answer styles where *MC* is multiple choice, *Desc* is description (free-form answering), and *Ex* is answer extraction by selecting a span in the given context. *Size* indicates the size of the whole dataset including training, development, and test sets. *Src* represents how the questions are sourced where *X* means questions written by experts, *C* by crowdworkers, *A* by machines with an automated manner, and *Q* are search-engine queries.

Name	Ans	Size	Corpus	Src	Focus
ARCT (Habernal et al., 2018)	MC	2.0K	debate article	C/X	reasoning on argument
QAngaroo (Welbl et al., 2018)	Ex	50K	Wikipedia, MEDLINE	A	multi-hop reasoning
CLOTH (Xie et al., 2018)	MC	99K	various	X	cloze in exam text
NarrativeQA (Kočiský et al., 2018)	Desc	45K	movie script	C	summary/full story tasks
MCScript (Ostermann et al., 2018)	MC	30K	written story	C	commonsense reasnoing, script knowledge
CliCR (Suster and Daelemans, 2018)	Ex	100K	clinical case text	A	cloze style queries
ARC (Clark et al., 2018)	MC	8K	science exam	X	retrieved documents from textbooks
DuoRC (Saha et al., 2018)	Ex	186K	movie script	C	commonsense reasoning, multi-sentence reasoning
ProPara (Dalvi et al., 2018)	Ex	2K	science exam	A	procedural understanding
DuReader (He et al., 2018)	Desc	200K	web snippet	Q/C	Chinese, Baidu Search/Knows
MultiRC (Khashabi et al., 2018)	MC	6K	various documents	C	multi-sentence reasoning
Multi-party Dialog (Ma et al., 2018)	Ex	13K	TV show transcript	A	1.7k crowdsourced dialogues, cloze query
SQuAD 2.0 (Rajpurkar et al., 2018)	Ex/NA	100K	Wikipedia	C	unanswerable questions
ShARC (Saeidi et al., 2018)	YN*	32K	web snippet	C	reasoning on rules taken from government documents
QuAC (Choi et al., 2018)	Ex/YN	100K	Wikipedia	C	dialogue-based, 14k dialogs
Textworlds QA (Labutov et al., 2018)	Ex	1.2M	generated text	A	simulated worlds, logical reasoning
SWAG (Zellers et al., 2018)	MC	113K	video captions	M	commonsense reasoning
emrQA (Pampari et al., 2018)	Ex	400K	clinical documents	A	using annotated logical forms on i2b2 dataset
HotpotQA (Yang et al., 2018)	Ex/YN	113K	Wikipedia	C	multi-hop reasoning
OpenbookQA (Mihaylov et al., 2018)	MC	6.0K	textbook	C	commonsense reasoning
RecipeQA (Yagcioglu et al., 2018)	MC*	36K	recipe script	A	multimodal questions
ReCoRD (Zhang et al., 2018)	Ex	120K	news article	C	commonsense reasoning, cloze query

Table 4: Machine reading comprehension datasets published in 2018. *Ans* denotes answer styles where *MC* is multiple choice, *Desc* is description (free-form answering), and *Ex* is answer extraction by selecting a span in the given context. *Size* indicates the size of the whole dataset including training, development, and test sets. *Src* represents how the questions are sourced where *X* means questions written by experts, *C* by crowdworkers, *A* by machines with an automated manner, and *Q* are search-engine queries.

Name	Ans	Size	Corpus	Src	Focus
CoQA (Reddy et al., 2019)	Ex/YN	127K	Wikipedia	C	dialogue-based, 8k dialogs
Commonsense QA (Talmor et al., 2019)	MC	12K	ConceptNet	C	commonsense reasoning
Natural Questions (Kwiatkowski et al., 2019)	Ex/YN	323K	Wikipedia	Q/C	short/long answer styles
DREAM (Sun et al., 2019)	MC	10K	language exam	X	dialogue-based, 6.4k multi-party dialogues
DROP (Dua et al., 2019)	Desc	96K	Wikipedia	C	discrete reasoning
BoolQ (Clark et al., 2019)	YN	16K	Wikipedia	Q/C	boolean questions, subset of Natural Questions
MSCript 2.0 (Ostermann et al., 2019)	MC	20K	narrative	C	commonsense reasoning, script knowledge
HellaSWAG (Zellers et al., 2019b)	MC	70K	web snippet	A	commonsense reasoning, WikiHow and ActivityNet
Quoref (Dasigi et al., 2019)	Ex	24K	Wikipedia	C	coreference resolution
CosmosQA (Huang et al., 2019)	MC	36K	narrative	C	commonsense reasoning
PubMedQA (Jin et al., 2019)	YN	273.5K	PubMed	X/A	biomedical domain, 1k expert questions
QuAIL (Rogers et al., 2020)	MC	15K	various	C	prerequisite real tasks

Table 5: Machine reading comprehension datasets published in 2019. *Ans* denotes answer styles where *MC* is multiple choice, *Desc* is description (free-form answering), and *Ex* is answer extraction by selecting a span in the given context. *Size* indicates the size of the whole dataset including training, development, and test sets. *Src* represents how the questions are sourced where *X* means questions written by experts, *C* by crowdworkers, *A* by machines with an automated manner, and *Q* are search-engine queries.