Evaluating Natural Language Understanding
in Machine Reading Comprehension
（機械読解における自然言語理解の評価）

by

Saku Sugawara
菅原 朔

A Doctor Thesis
博士論文

Submitted to
the Graduate School of the University of Tokyo
on December 6, 2019
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Information Science and Technology
in Computer Science

Thesis Supervisor: Akiko Aizawa　相澤 彰子
Professor of Computer Science

ABSTRACT

Building machines that can understand human language is one of the long-standing challenges in natural language processing. This thesis tackles how to evaluate natural language understanding in machine reading comprehension—a task in which computer systems answer questions about given texts. Machine reading comprehension is an important testbed for jointly evaluating various aspects and components of language understanding. There are large-scale, various datasets presented recently on some of which proposed systems achieved human-level performance. However, we raise two major issues in machine reading comprehension. The first issue is about evaluation metrics. Because systems are evaluated with simple accuracy in most existing datasets, we cannot obtain fine-grained information about the reading comprehension ability of a system. Therefore, we cannot explain what the system achieved in terms of language understanding, which prevent us from improving the development of systems. The latter issue is about the quality of questions. Even if questions seem to require human-level understanding of given texts, they may be solved only by matching word patterns between a question and given texts. In this situation, we cannot conclude that the system achieved human-level language understanding even if it exhibits the performance comparable with humans.

In this thesis, we discuss how we can design a dataset of machine reading comprehension for precisely and correctly evaluating the capability of language understanding. This thesis consists of seven chapters. In Chapter 1, we introduce current issues in machine reading comprehension and our motivation. In Chapter 2, we overview machine reading comprehension datasets, systems, and related language understanding tasks. In Chapter 3, we consider evaluation metrics, namely, how to evaluate the performance of machines beyond simple accuracy. We propose new metrics comprised of requisite skills and text readability to highlight systems' abilities in detail. In Chapter 4, we address how to investigate the quality of questions so that they can correctly evaluate intended language understanding. We propose analysis methods to look into question difficulty and requisite skills. In Chapter 5, we present a methodology for automatically assessing the benchmarking capacity of machine reading comprehension datasets from language understanding skills. We combine our proposed skills and analysis methods and reveal what kind of skills is required for answering questions. In Chapter 6, we discuss the explainability of machine reading comprehension and provide theoretical foundations for reading comprehension and its evaluation. We inspect current machine reading comprehension using these foundations and list requirements for the explainability. In Chapter 7, we summarize conclusions and mention the future of machine reading comprehension. The analysis methods of machine reading comprehension datasets we proposed in this thesis contribute to facilitating the explainability of MRC in terms of what kind of language understanding is required. This is important for the verification of hypothesis testing in academic research and the accountability of practical applications such as assisting human intelligent activities.

論文要旨

自然言語処理分野では、人間のように文章を理解するシステムを構築することがひとつの大きな目標である。本論文では機械読解を用いて自然言語理解をどのように評価するかという課題に取り組む。機械読解は与えられた文章に対して質問に答える形式のタスクであり、言語理解の様々な側面や要素を同時に評価するために重要である。近年は大規模で多様なデータセットが提案されており、データセット上で人間に匹敵する性能を示しているシステムも提案されている。しかしながら大きく2つの課題がある。ひとつは評価指標の単純さである。ほとんどのデータセットは単純な精度のみを評価指標としており、システムがもつ言語理解能力の多様な側面を評価することができない。したがって実際にシステムが達成したことについての説明を与えることができず、着実な開発に結びつけづらい。もうひとつの課題はデータセットにおける問いの品質である。多くのデータセットにおいて人間らしい言語理解が要求されていると思われるような問いであっても、実際は単純なパターンマッチなどで正答できてしまうようなものが存在する。こうした状況では、仮にデータセット上の性能が高くても人間と同等な言語理解を実現していると言うことができない。

本論文では、自然言語理解を適切に評価するための機械読解タスクをどのように設計すればよいかを議論する。論文は7章からなる。1章では本論文の背景と動機を説明する。2章では機械読解タスクのデータセットとシステムを概観する。3章では評価指標、具体的には単なる精度だけではない仕方でどのようにシステムの性能を評価するかについて考える。新しい評価指標として読解に要求される能力と読みやすさを提案し、システムの性能をより詳細に明らかにする。4章では正確に言語理解を評価するために必要な問題の品質について調査する。具体的には、問題の難易度や要求される能力の観点からシステムを分析する手法を提案する。5章では機械読解データセットのベンチマーク性能を言語理解の能力の観点から自動的に評価する方法を提案する。2章と3章で提案した評価指標と分析手法を組み合わせ、これまで解かれている問いにおいてどのような能力が必要になっているか明らかにする。最後に6章では、機械読解タスクの説明性について着目し、読解そのものが何であるか、そしてそれをどのように検証するかという課題に必要な理論的背景を与える。現状の機械読解タスクをそれらの背景から分析しながら、説明性のためにタスクが満たすべき要件を列挙する。7章では各章の結論をまとめ、今後の展望について述べる。本論文における機械読解データセットの整備・評価手法の提案は、データセットがどのような言語理解を要求しているかという説明性を高めることに貢献するものであり、学術的な仮説検証を通して知見の蓄積をしていく営為や人間の知的活動を支援するような応用技術において非常に重要であると言える。

Acknowledgements

I spent nine years from 2011 to 2020 at the University of Tokyo. I do not feel like I am actually writing the acknowledgments of my Ph.D. thesis, but it means that my long university life is ending soon. When I entered the graduate school, I cannot write and speak English (although I do not think I can do it now) nor write academic papers. Although having been getting better now, I still always struggle to write and speak fluently. I still do not know what to write here—anyway, I would like to express my gratitude to many people.

First of all, I am deeply grateful to my supervisor, Akiko Aizawa. I remember that when I first met her, I talked about my graduate thesis that studied a topic in philosophy of language. Although my research was always ambiguous and opaque since then, she kindly listens to my mysterious words and gives me comments patiently. This thesis consists of topics that were hard to handle, but it finally reached completion thanks to her thoughtful supervision.

I also thank Yuki Arase, Takuya Matsuzaki, Masami Hagiya, Akihiko Takano, and Yusuke Miyao for being on my thesis committee. They gave me a lot of guidance, which helped me improve the early draft and the presentation of this thesis.

I was a research assistant at RIKEN Center of Advanced Intelligence Project since 2017. I also thank my collaborators—especially to Professors Satoshi Sekine and Kentaro Inui. Their thoughtful comments helped me decide where I should go. During my Ph.D., I did a research internship at Microsoft Research Montreal and a research visit at University College London. I thank Tong Wang, Eric Yuan, Tetsuro Oya, and Adam Trischler at Microsoft and thank Pascuale Minervini, Max Bartolo, Johannes Welbl, Partick Lewis, Yuxiang Wu, Marco Rudilosso, Sebastian Riedel, and Pontus Stenetorp at University College London. I really enjoyed my stays and their help made my great experience. I remind that it was so hot in the summer of Montreal and I met nice Asian cuisines in London.

Finally, I would like to thank my friends, family, and wife. I have survived until today thanks to their supports. I devote this thesis to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Teaching machines to understand human language (*natural language*) texts is a long-standing goal in the field of natural language processing (NLP). In activities related to language understanding, reading comprehension is the essential ability to process and comprehend texts. How can we develop machines that are capable of reading and comprehending texts? One way is to use a task by which we can teach and test machines. Machine reading comprehension (MRC) is one of such tasks that requires a machine reading a given text (context) and answering questions about it, as we humans do.

Figure 1.1 shows example MRC questions taken from MCTest (Richardson et al., 2013). This dataset consists of hundreds of children-level narratives and questions that are collected using crowd-sourcing methods. In the following, we overview what understanding is required for answering these questions.

- Question 1. You first need to understand what the *escape* in the question means. This refers to the fact that the princess (the answer to this question) left the high tower when her mother was sleeping, which is described in the third sentence in the given context. This reference needs to be made by **commonsense reasoning** because we might not refer it to *escape* if the person who left the high tower was someone else other than the princess.

- Question 2. In addition to understanding *escaping*, you also need to understand **coreference resolution** between *princess* and *she* (*she wandered out ...* and *she went into ...*) in the context. Then you look at the *after* in the question and recognize that there are two subsequent events; the princess first escaped and then wandered into the forest (to be precise, you also understand that the sequence of *wandered* and *went into the forest* means *wandered to the forest*). It means that you understand the **temporal relation** between these two events and focus on the latter event by indicating *after*.

- Question 3. You recognize that there is some relationship between *climb* and *see* mentioned in the question. Because the princess climbs up to the top of a tree, she was able to see the castle. Understanding this **causal relation** is required for answering this question.

- Question 4. The phrase *in the beginning*, similarly to Question 3, needs you to understand the temporal order of the events described in the context. Looking

ID	mc160.dev.29
Context	Once upon a time there was a princess who lived in a high tower and she was not allowed to leave because of her mean mother. One day she chose to leave but her mother would not let her. The princess climbed out the window of the high tower and climbed down the south wall when her mother was sleeping. She wandered out a good ways. Finally she went into the forest where there are no electric poles but where there are some caves. There she met a young man who was running. His name was John. John asked the princess why such a beautiful woman like her was out in the middle of a forest. She said that she had been trapped for her whole life by an evil woman who said she was her mother. The man said that he would take the princess to a castle that was near. He also said that he thought that she may be the missing princess. As they go through the forest they run into many problems. They see that they are lost and have no way of finding where to go. After several days pass, the princess climbs up to the top of a tree in order to find out where they are. She sees that the castle where they want to go is not that far away and near a mountain. After thinking of the best way to get there, John and the princess go to the castle where they live for the rest of their lives.
Question 1	Who escaped from the tower?
Answer	(A) Mother **(B) Princess** (C) Man (D) John
Question 2	Where did the princess wander to after escaping?
Answer	(A) Mountain **(B) Forest** (C) Cave (D) Castle
Question 3	What did the princess climb to see the castle?
Answer	(A) Electric pole (B) mountain **(C) Tree** (D) Castle
Question 4	Where does the princess live in the beginning?
Answer	(A) Castle (B) house (C) Cave **(D) High Tower**

Figure 1.1: An example of machine reading comprehension questions from MCTest (Richardson et al., 2013).

at the first sentence in the context, you implicitly process the relative clause (*a princess who lived*) and recognize that the princess lives in the high tower. However, **parsing complex sentences** sometimes might not be easy for machines.

Answering these questions above, therefore, involves different processings of texts. From the perspective of NLP, it also involves other fundamental technologies such as part-of-speech tagging and named entity recognition. A major advantage of MRC is that in principle we can comprehensively evaluate various processings of which the ability of reading comprehension is comprised. Moreover, its general form of a triplet (context, question, answer) makes it easier to apply MRC systems to other related tasks by converting their task form into the form of MRC.

Recent success in MRC is fueled by the advancement of neural-network models and large-scale datasets. As machine learning models including neural networks have developed, it is necessary to create a massive size of data to train large models. In MRC, it is resulting in dozens of datasets consisting of several hundreds of thousand of questions collected using automated generation and crowdsourcing methods. Neural reading comprehension models show great progress in such datasets. For example, in the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), the best systems that are pre-trained on the language modeling task with more than hundreds million parameters achieved a beyond-human performance.

Historically speaking, MRC can be seen as a method for evaluating language understanding, which is one of the biggest challenges in the field of artificial intelligence (AI). An initial conceptual attempt is made by Turing (Turing, 1950), in which he proposed the Turing Test to evaluate a machine's intelligent behaviors through testing whether it can mimic human responses in dialogue. Later, in NLP, several milestone tasks have been established to date:

- **Recognizing Textual Entailment**: Dagan et al. (2006) proposed a task that requires capturing textual entailment between two short texts (mostly sentence). Textual entailment includes not only logical entailment but also a variety of linguistic phenomena and knowledge reasoning.

- **Winograd Schema Challenge**: Levesque (2011) proposed a simple coreference resolution task that consists of twin sentences with a small difference. To resolve the ambiguity between the twin sentences, you need to perform commonsense reasoning triggered by the difference between them.

- **Reading Comprehension**: To our knowledge, Hirschman et al. (1999) initially introduced to NLP. Followed by Richardson et al. (2013) and Hermann et al. (2015), large-scale datasets started to be proposed where crowdsourcing was used for collecting a large number of questions to fuel neural network models.

In Chapter 2, we give an overview of MRC and discuss its differences with related tasks.

Our main goal in this thesis is to investigate how to design MRC for the purpose of *evaluating natural language understanding*. Our underlying motivation is two folds:

> (i) **Evaluation metrics.** We usually only use simple accuracy as evaluation metrics. As we introduced above, a major advantage of MRC is that the formulation of MRC allows us to comprehensively evaluate various aspects in the process of reading comprehension. However, the single metric does not tell us the fine-grained information on the performance of evaluated systems. Therefore, it is not straightforward to explain what the systems can understand or not on the current MRC datasets. In Chapter 3, we propose new evaluation metrics of MRC that are organized as a set of *skills* required for language understanding.
>
> (ii) **Question quality.** Questions of MRC datasets are usually collected using crowdsourcing or automatically generated. Just asking crowdworkers to write questions does not guarantee that sourced questions have enough quality to precisely evaluate human-level language understanding even if we have a variety

ID	RACE / high6527.txt / Question 1
Context	Having a great collection of books at home doesn ’ t really mean that you are a person who has a passion for literature and reading . It can be a family inheritance or it can be just to impress people around you by the fact that you are a person of culture . On the other hand , there are many persons who cannot afford to buy books , because some of them are quite expensive , but who usually go to libraries and spend hours reading something that interests them a lot , or just borrow books to home . From my point of view , literature is very important in our life . For example , reading is a means of gaining culture and enriching our knowledge in different areas . It can help us have a great imagination and it makes things easier when it comes to make compositions on different themes . [...]
Context after ablation	█ █ great collection █ books █ home █ ’ █ really mean █ █ █ █ person █ █ █ passion █ literature █ reading . █ █ █ █ family inheritance █ █ █ █ █ █ impress people around █ █ █ fact █ █ █ █ person █ culture . █ █ █ hand , █ █ many persons █ cannot afford █ buy books , █ █ █ █ █ quite expensive , █ █ usually go █ libraries █ spend hours reading something █ interests █ █ lot , █ █ borrow books █ home . █ █ point █ view , literature █ █ important █ █ life . █ example , reading █ █ means █ gaining culture █ enriching █ knowledge █ different areas . █ █ help us █ █ great imagination █ █ makes things easier █ █ comes █ make compositions █ different themes . [...]
Question	People who are fond of literature are those that ______ .
Answer	**(A) have much interest in reading** (B) keep many books (C) go to libraries every day (D) buy expensive books in the bookstore
System Prediction	(A) → (A)

Figure 1.2: Example of our ablation methodology. Even after dropping function words (stop words), an evaluated system correctly answers the question.

of good metrics. In other words, we cannot ensure that sourced questions correctly evaluate intended processings in reading comprehension. In Chapter 4, we propose simple heuristics for filtering low-quality questions out.

After tackling these issues separately, we integrate our approaches into an evaluation methodology of MRC datasets, analyzing the capability of them for precisely evaluating language understanding in Chapter 5. This contributes to assessing the quality of questions with well-organized metrics. We propose ablation-based methods for analyzing what is needed for answering questions. Our intuition is that, as shown in an example in Figure 1.2, we could say that if a system can correctly answer a question even after ablating features (e.g., function words), the question does not

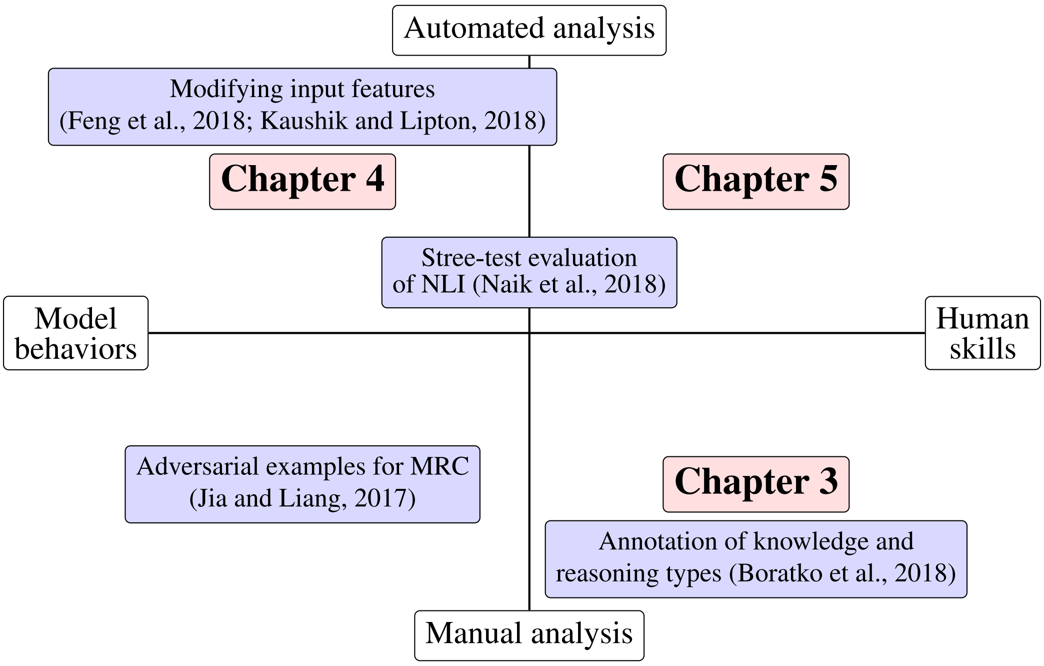
Automated analysis
Modifying input features
(Feng et al., 2018; Kaushik and Lipton, 2018)
Chapter 4
Chapter 5
Stree-test evaluation
of NLI (Naik et al., 2018)
Model
behaviors
Human
skills
Adversarial examples for MRC
(Jia and Liang, 2017)
Chapter 3
Annotation of knowledge and
reasoning types (Boratko et al., 2018)
Manual analysis

Figure 1.3: Relations among the main part of this thesis (Chapters 3, 4, and 5) and related work.

require understanding of the ablated information.

Figure 1.3 illustrates the relations among the main part of this thesis (Chapters 3, 4, and 5) and related work on the dichotomy between *from what perspective the analysis is performed* (the horizontal axis) and *how the analysis is performed* (the vertical axis). In Chapter 3, our defined metrics are mainly based on human skills and its annotation is manually conducted. Boratko et al. (2018) performed a similar annotation work with knowledge and reasoning types. On the other hand, our heuristics in Chapter 4 provide an automated method to analyze MRC questions through model behaviors, similarly to how existing studies (Feng et al., 2018; Kaushik and Lipton, 2018; Naik et al., 2018) analyzed their models. These attempts are inspired by finding unintended behaviors of natural language understanding models using manually created adversarial examples (Jia and Liang, 2017). Merging these two directions of *human skills* and *automated analysis*, we develop automated methods for analyzing the quality of MRC datasets in terms of human-oriented language understanding skills in Chapter 5.

We finally discuss our findings in this thesis, associating them with psychological foundations in Chapter 6. Our discussion focuses on *explainability* in MRC; it differs from the interpretability of internals of machine learning models that depends on their architecture (Gilpin et al., 2018). We define the explanation of MRC as a task-oriented description *in human terms* that specifies the process of reading comprehension (i.e., an agent reads a given document and answers a question about it). It is also intended to be provided in an understandable and agreeable manner among the research community. Then we argue that the design of existing MRC datasets may be insufficient for ensuring the explainability of reading comprehension, which is important for the verification of hypothesis testing in academic research and the accountability of practical applications such as assisting human intelligent activities.

1.2 Thesis Outline

The outline of this thesis is as follows:

Chapter 2. *What is the study of machine reading comprehension and its differences with other related language understanding tasks?* We first overview of the MRC task itself, listing datasets and types of systems in the field after we formally define the task. We then mention related tasks as listed above and discuss the difference between MRC and them.

Chapter 3. *What metrics can we use to evaluate machine reading comprehension?* Beyond simple accuracy, we adopt two classes of metrics for evaluating MRC datasets: *requisite skills* and *readability*. Our assumption is that the difficulty of answering questions can be separated from the difficulty of reading given documents. We first define a set of requisite skills according to existing studies in NLP and the psychological literature. Regarding readability, we use measures proposed in previous studies such as the average number of characters per word and the average length of sentences in the context. We apply these classes to six existing datasets, and highlight the characteristics of the datasets according to each metric and the correlation between the two classes. This chapter is based on our works (Sugawara and Aizawa, 2016; Sugawara et al., 2017a,b).

Chapter 4. *How can we ensure that questions require precise language understanding?* To analyze the quality of questions in MRC, we investigate what makes questions easier across recent 12 MRC datasets with three answering styles (answer extraction, description, and multiple choice). We propose to employ simple heuristics to split each dataset into *easy* and *hard* subsets and examine the performance of two baseline models for each of the subsets. We then qualitatively analyze questions sampled from each subset by annotating them with both validity and requisite reasoning skills to investigate which skills explain the difference between easy and hard questions. This chapter is based on our work (Sugawara et al., 2018).

Chapter 5. *How can we specify high-quality questions with organized metrics?* In order to assess the capabilities of datasets for benchmarking language understanding precisely, we propose a semi-automated, ablation-based methodology for this challenge; By checking whether questions can be solved even after removing features associated with a skill requisite for language understanding, we evaluate to what degree the questions do *not* require the skill. We exemplify our ablation-based methods along with newly defined 12 skills and analyze 10 existing MRC datasets, highlighting the characteristics of them in terms of requisite skills. This chapter is based on our work (Sugawara et al., 2020b).

Chapter 6. *In summary, what is necessary for the explainability of machine reading comprehension?* We finally discuss the explainability of MRC for the future development of the task. Our motivation is that, in the recent situation of the field, we cannot *explain* what models can do and how they internally work beyond the interpretation of models. We aim to provide theoretical bases of reading comprehension and its task design from psychology and psychometrics and then summarize them as requirements and future directions for explainable MRC. This chapter is based on our work (Sugawara et al., 2020a).

1.3 Contributions

The contributions of this thesis are summarized as follows:

- In Chapter 2, we give a comprehensive overview of the literature in the field of MRC. It also covers related tasks in NLP, distinguishing them from MRC by discussing their purpose and possible applications. We also indicate that the most important feature of MRC is its general formulation.

- In Chapter 3, our defined comprehensive set of requisite skills and readability measures help us know the quality of MRC datasets for the development of natural language understanding systems. Our dataset analysis suggests that there is only a weak correlation between the readability of given documents and the question difficulty and that we could create an MRC dataset that is easy to read but difficult to answer.

- In Chapter 4, our proposed heuristics contribute to collecting questions that require a sophisticated understanding of language to answer beyond recognizing superficial cues. Our analysis demonstrates that (i) the baseline performances for the hard subsets remarkably degrade compared to those of entire datasets, (ii) hard questions require knowledge inference and multiple-sentence reasoning in comparison with easy questions, and (iii) multiple-choice questions tend to require a broader range of reasoning skills than answer extraction and description questions. These observations imply that one might overestimate recent advances in MRC.

- In Chapter 5, our analysis methodology assesses the capabilities of datasets for benchmarking language understanding precisely. Experiments on 10 datasets show that, for example, the relative scores of a baseline model provided with content words only and with shuffled sentence words in the context are on average 89.2% and 78.5% of the original score, respectively. These results suggest that most of the questions already answered correctly by the model do not necessarily require grammatical and complex reasoning. It is implied that, for precise benchmarking, MRC datasets will need to take extra care in their design to ensure that questions can correctly evaluate the intended skills.

- In Chapter 6, we provide theoretical foundations MRC and discussed requirements for the explainability of MRC. We conclude that future datasets should evaluate the capability of a coherent, grounded representation and ensure the verification of evaluating the internal process of reading comprehension by collecting high-quality questions and designing a white-box task formulation.

Chapter 2

An Overview of Machine Reading Comprehension

In this chapter, we give an overview of machine reading comprehension. We first provide a formal definition of the task (Section 2.1), and then describe task components and representative existing datasets. Next, we introduce a short history of datasets and systems along with some trends in the field (Sections 2.2 and 2.3). We finally discuss the difference between MRC and other related natural language understanding tasks in Section 2.4.

2.1 Task Definition

2.1.1 General Definition

Machine reading comprehension (MRC) is a task in which a system is given a document (which we refer to the *context*) and answers questions about it. Historically, MRC can be seen as a way to evaluate natural language understanding (NLU) systems in terms of its behavior, similar to the original Turing test (Turing, 1950). As a general definition of MRC, Burges (2013) noted 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*. With a focus on the NLU competence required in the ideal reading comprehension task, Sutcliffe et al. (2013) also noted that *reading comprehension tests do not require only semantic understanding but they assume a cognitive process which involves using implications and presuppositions, retrieving the stored information, performing inferences to make implicit information explicit.* These *skills* are prerequisites that are common to all readers and are comprehensively required in reading comprehension. Accordingly, we can summarize the general objective of MRC as follows: MRC is an evaluation method for NLU systems in terms of their behavior, and tests a cognitive process that involves several skills, such as performing inferences using background knowledge, by letting the system answer questions about a given text. We consider that this *given text,* including multiple sentences and paragraphs longer than a single sentence, is an important feature of MRC (refer to Section 2.4 for comparison with other related tasks).

2.1.2 Formulation

In MRC, a system (which we call a *model* interchangeably in this thesis) M is given a context document d (which we often call *context*) and a question q about d, and return its output a. That is,

$$M : (d, q) \to a. \tag{2.1}$$

A dataset is defined as a collection of triples (d, q, a). According to variations of d, q, and a respectively, there are task variants of MRC. For example, d is a passage of text, q is a natural interrogative question written by crowdworkers, and a is a text span selected from d in SQuAD (Rajpurkar et al., 2016). In the following sections, we introduce their variants and existing representative datasets.

Context styles. The form of a given document d can be different in its length, for example, a single paragraph (Rajpurkar et al., 2016), a set of paragraphs (Yang et al., 2018), or a longer document (Kočiský et al., 2018). Chen et al. (2017) proposed open domain question answering, as an extension of single-paragraph reading comprehension, where all Wikipedia articles can be the context. 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 a bag of words that does not necessarily form a sentence (e.g., knowledge-base entries (Welbl et al., 2018) and web search queries (Nguyen et al., 2016)).

Answer styles. An answer is (i) chosen from a text span of the given document (*answer extraction*; e.g., Rajpurkar et al. (2016) and Trischler et al. (2017)), (ii) chosen from a candidate set of answers (*multiple choice*; e.g., Richardson et al. (2013) and Clark et al. (2018)), or (iii) generated as a free-form text (*description*; e.g., Kočiský et al. (2018) and Dua et al. (2019)). Some datasets optionally allow answering by a *yes/no* reply (Clark et al., 2019; Choi et al., 2018).

Formally, for the answer extraction, a is a text span in a given document $d = (d_1, d_2, \dots, d_{l_d})$ where l_d denotes the length of d:

$$a = (a_{start}, a_{end}) \text{ where } 1 \leq a_{start} \leq a_{end} \leq l_d. \tag{2.2}$$

Here we followed some notations by Chen (2018). For the multiple choice, a is chosen from a given set of k candidate answers:

$$a \in \{a_1, a_2, \dots, a_k\}, \tag{2.3}$$

where a can be a word, a phrase, or a sentence. For the free-form description style, a can be an arbitrary length of text.

Evaluation metrics. For the answer extraction style, the system performance is computed based on the overlap between predicted and gold answers. Exact match (EM) literally assigns 1.0 if the predicted and gold answered are completely matched. F1 score computes the word overlap between the predicted and gold answers using recall and precision. For the multiple choice style, a simple accuracy is computed.

For the generation style, BLEU (Papineni et al., 2002), Meteor (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004) scores against the gold answer are usually computed, which are commonly used in the natural language generation community.

2.2 Datasets

2.2.1 Chronological Trends

Before the machine learning era. To our knowledge, Hirschman et al. (1999) were the first to use NLP methods for reading comprehension. They posited reading comprehension by machines as a research problem and an evaluation method for language understanding systems. Their dataset comprised reading materials for grades 3–6 with simple 5W (*wh-*) questions.

Subsequent investigations into questions of natural language understanding focused on other formulations, such as question answering (Yang et al., 2015; Wang et al., 2007; Voorhees and Tice, 1999) and textual entailment (Bentivogli et al., 2010a; Sammons et al., 2010; Dagan et al., 2006). One of MRC datasets of the time was QA4MRE (Sutcliffe et al., 2013). The highest accuracy achieved for this dataset was 59% and the size of the dataset was very limited: there were only 224 gold-standard questions, which is insufficient for machine learning methods.

Large-scale sourcing. This means that an important issue for designing MRC datasets is their scalability. Richardson et al. (2013) presented MCTest, which is an open-domain narrative dataset for gauging comprehension at a child's level. This dataset was created by crowdsourcing and was based on a scalable methodology. Since then, additional large-scale datasets have been proposed with the development of machine learning methods in NLP. For example, the CNN/Daily Mail dataset (Hermann et al., 2015) and CBTest (Hill et al., 2016) have approximately 1.4M and 688K passages, respectively. These context texts and questions were automatically curated and generated from large corpora. However, Chen et al. (2016) indicated that approximately 25% of the questions in the CNN/Daily Mail dataset are either unsolvable or nonsensical. This dataset-quality issue highlights the demand for more stable and robust sourcing methods.

Several additional datasets were presented in the last half of 2016, involving large documents and sensible queries that were guaranteed by crowdsourcing or other human testing. They were intended to provide large and high-quality content for machine learning models. Nonetheless, as shown in the examples of Chapter 1, they were not offered with metrics that could evaluate NLP systems adequately with respect to the difficulty of questions and the surface features of texts.

Skill-oriented datasets. After around 2018, there seem to be proposed datasets that are specialized for requiring specific situations and skills including unanswerable questions (Rajpurkar et al., 2018), dialogue (Choi et al., 2018; Reddy et al., 2019; Sun et al., 2019a), multiple-sentence reasoning (Khashabi et al., 2018a), 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). However, as we will view in the following chapters, these datasets struggled to ensure high-quality questions.

Domains. Finally, we briefly mention that there are several domains used for MRC. 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.

2.2.2 Existing Datasets

We list existing datasets in Tables 2.1, 2.2, and 2.3 with their answering style, dataset size, corpus used for sourcing, question sourcing methods, and focuses.

2.3 Systems

In this section, we briefly overview MRC systems. Since MRC datasets afford training and validation splits for developing machine learning models, a mainstream is twofold: feature-based learning models and neural network models. We survey these two approaches with representative studies.

2.3.1 Feature-based Machine Learning Pipelines

In the early days of MRC, feature-based machine learning systems were used because of the small size of datasets. For example, Wang et al. (2015) proposed a max-margin learning framework with features based on syntax, frame semantics, coreference, and word embeddings for MCTest (Richardson et al., 2013). Similarly for MCTest, Sachan et al. (2015) convert multiple choice questions to textual entailment and solve it by considering the word alignment between the context and a hypothesis. On the other hand, Berant et al. (2014) approached reading comprehension questions in the biological domain by modeling event relations such as causality. More recently, Khashabi et al. (2018b) formulated a constrained optimization problem in science questions as an integer linear program and optimized it using an off-the-shelf solver. They gathered semantic features from the predicate-argument structure, part-of-speech tags, and named entities. Note that these approaches are proposed for multiple choice datasets, they are not directly applicable to other answering styles.

2.3.2 Neural-network Models

Fueled by the advancement of neural-network models and large-scale datasets, neural network-based models have been proposed since 2015. For example, Hermann et al. (2015) proposed a long short-term memory (LSTM) model together with the CNN/Daily Mail dataset which has 1.4 million automatically generated questions. Their model first encodes the document and its query using separate bidirectional single layer LSTMs and predicts the final answer after applying a linear layer to the encoded representation. Chen et al. (2016) improved this model by modifying its activation function and using a simplified encoding layer. In addition, Seo et al. (2017) enhanced this LSTM-based approach by adding a character embedding layer and bidirectional attention layers over the encoding module. Their model finally outputs probability distributions over the context that are used to predict the start and end indices of a span in the given paragraph. This model achieved the highest performance with a large margin for SQuAD (Rajpurkar et al., 2016) at that time.

More recently, Radford et al. (2018) proposed the generative pre-trained transformer which consists of 12 transformer blocks (Vaswani et al., 2017) pre-trained for language modeling. They finetuned that model on varieties of target tasks including natural language inference, question answering, sentence similarity, and sentiment classification. Among these tasks, the model presented the highest accuracy on RACE (Lai et al., 2017) at that time. For RACE, the input is a sequence of tokens, in which the context text, the question, and one candidate answer are concatenated with a delimiter just after the question tokens. The model then processes that sequence and generates the probability of the final token of the input sequence. Finally, the candidate answer that gives the highest probability is selected.

Half a year later, Devlin et al. (2019) presented a new language representation model called BERT. Similarly to the pre-trained transformer by Radford et al. (2018), they pre-trained 24 transformer blocks for language modeling on BookCorpus (Zhu et al., 2015) and all Wikipedia articles. Their model can be seen as a general encoder module and thus is generally applicable to various NLP tasks including MRC. It also showed new state-of-the-art results on SQuAD. After the pre-training based models exhibited the high performance, MRC models are often constructed on top of them. Pre-training based models are also improved in the training paradigm and scalability (Dai et al., 2019).

On the other hand, as datasets focus on certain types of skills, proposed models become task-specific. For example, Ding et al. (2019) proposed graph neural network-based models are proposed for multi-hop reasoning datasets. Min et al. (2019b) also presented a model that decomposes questions into a few sub-questions for multi-hop reasoning. Andor et al. (2019) proposed a mathematical reasoning model that executes a program picked from among a predefined set of executable programs that encompass arithmetic operations. For procedural reasoning on science facts (Dalvi et al., 2018), Das et al. (2019) proposed to construct dynamic knowledge graphs. Their models use an off-the-shelf MRC model to extract entities to evolve knowledge graph representations. Although these approaches are specific to targeted tasks, we seem to combine them with general-purpose models (e.g., pre-trained encoders) in order to generalize MRC models to various kinds of tasks in which language understanding skills are comprehensively required.

2.4 Related Tasks and Comparison

In the field of NLP, there are several task formulations for testing NLU competence. This section compares MRC with datasets of other related tasks, including QA (Section 2.4.1), RTE (Section 2.4.2), and other topics including dialogue systems and computer vision (Section 2.4.3). For the first two tasks, we consider the differences between them from the viewpoint of *components*, *contents* and *knowledge use*.

2.4.1 Question Answering

Recently, QA datasets that are relevant to MRC have been proposed. These datasets require systems to find a correct word/phrase as an answer for a given question, from long or segmented documents. Although such datasets originally had limited numbers of questions, e.g., TREC-8 QA (Voorhees and Tice, 1999), recently proposed datasets have become open domain and large scale (Choi et al., 2017; Hewlett et al., 2016; Yang et al., 2015; Iyyer et al., 2014). In addition, Chen et al. (2017) proposed a task that combines the challenge of document retrieval with open-domain QA.

Components. The QA task is similar to MRC in that MRC is performed in the QA manner. However, the most important difference between them is that MRC has the context comprising explicit and limited text. Unlike traditional open-domain QA with a knowledge base (e.g., Cheng et al. (2017) and Fader et al. (2014)), a question and its answer in an MRC dataset can be context-sensitive. This means that the information obtained in a certain context may not be relevant to another context (e.g., a character's intention in a narrative).

Contents. Most QA tasks involve factoid questions. This seems to be because the QA task historically lays a foundation based on the Web and knowledge bases. However, questions in several MRC datasets can query the understanding of sequential events, which are historically called *scenes* in script theory (Schank and Abelson, 1977) or *story understanding* (Mueller, 2003).

Knowledge use. In QA, knowledge can appear as an answer in itself, whereas readers in MRC use knowledge for inference purposes in answering. Although the readers may use linguistic and world knowledge, answers to the questions in MRC do not necessarily reflect knowledge itself but can be derived from the context.

2.4.2 Recognizing Textual Entailment

An RTE task requires recognizing, given a text pair of premise and hypothesis, whether the meaning of the hypothesis can be entailed from the premise (Bowman et al., 2015; Dagan et al., 2006).

Components. The context in MRC indicates a similarity to RTE rather than QA. In fact, MRC can be considered as RTE via the following three steps: (i) observe sentences in a given text as multiple premises; (ii) combine a question and its candidate answer into a hypothesis; and (iii) compare the hypothesis over each candidate answer and select the one that is best entailed by the premises. There are actual systems that consider answering MRC questions as RTE by these steps (Yin et al., 2016; Sachan et al., 2015). Although a question and answer pair in QA can also be seen as a hypothesis by step (ii), the hypothesis has no premises; a reader must find correct evidence (i.e., a premise) from external resources such as a knowledge base. Besides, Bentivogli et al. (2010b) and Bentivogli et al. (2011) proposed the RTE-6 and RTE-7 tasks, respectively, in which systems are required to identify all sentences that entail a given hypothesis among the candidate sentences. These task formulations are similar to MRC.

Contents. RTE instances can be created from a variety of sources. Originally, the instances of RTE-1 (Dagan et al., 2006) were collected from the general news domain. Bowman et al. (2015) created an RTE dataset based on captions from the flickr30k corpus (Young et al., 2014). Their topics are therefore mainly visible objects and events.

Knowledge use. RTE is presented for testing inference skills on lexical, syntactic, logical, and world-knowledge questions. That is, it can be described as *the act of passing from one proposition, statement, or judgment considered as true to another whose truth is believed to follow from that of the former* (Dagan et al., 2013). This skill is also important for MRC, with Etzioni et al. (2006) indicating that MRC can be seen as a combination of multiple textual-entailment steps.

Together with datasets, annotation and evaluation methodologies for RTE have been proposed (e.g., Sammons et al. (2010)). LoBue and Yates (2011) and Bentivogli et al. (2010a) proposed classifications of entailment phenomena required in RTE. Although these may seem useful for MRC, we should note that there are differences between MRC and RTE. As Etzioni et al. (2006) indicated, most instances of RTE datasets involve single-sentence premises and hypotheses, with few datasets having multiple premises (e.g., Cooper et al. (1996)). Therefore, to utilize existing RTE methodologies, as Manning (2006) has argued, we would need to intentionally move MRC in the direction of RTE with longer context documents.

When we consider *contextualism* in linguistics (Recanati, 2004), we assume that understanding of natural language text is necessarily conditioned by its given context. By adopting MRC rather than others, we can make explicit the context for understanding the text and can exclude adjacent processes appearing in other tasks such as searching in QA. This enables us to concentrate solely on the challenge of NLU.

2.4.3 Other Tasks

In this section, we introduce dialogue systems and computer vision among other related topics.

Dialogue systems. Raising a question plays an important role in conversational communication. Understanding the content of that question is necessary for answering it. However, a challenging issue in the study of dialogue systems is how to evaluate the system's response. Although some studies propose evaluation metrics that focus on the appropriateness or confidence of the system's utterance, they do not primarily target whether the listener correctly understands the speaker's utterance (Lowe et al., 2017; Higashinaka et al., 2014). Studies of dialogue systems and MRC seemingly share the same problem, namely assessing to what extent a listener/reader correctly understands a speaker's utterance/question associated with a dialogue history/given text.

Visual question answering. Some tasks tackle the challenge of question answering with respect to visual data such as images and movies (Johnson et al., 2017; Kembhavi et al., 2017; Suhr et al., 2017; Tapaswi et al., 2016; Antol et al., 2015). In such a task, a system is given a pair of some visual material and a natural language question, and is required to answer that question. Here, we can consider the visual material as the context. For example, Kembhavi et al. (2017) proposed a textbook-style question-answering task, in which each local context has a text and supplementary images.

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	on Wikidata articles
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., 2017b)	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 2.1: Machine reading comprehension datasets published before 2017. *Ans* denotes answering styles where *MC* is multiple choice, *Desc* is description (free-form answering), *Ex* is answer extraction, i.e., selecting a span in the given context, *YN* is yes or no, and *NA* is no answer. $(^*)$ indicates a dataset-specific answer style. *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 in an automated manner, and *Q* represents 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., 2018a)	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 2.2: Machine reading comprehension datasets published in 2018. *Ans* denotes answering styles where *MC* is multiple choice, *Desc* is description (free-form answering), *Ex* is answer extraction, i.e., selecting a span in the given context, *YN* is yes or no, and *NA* is no answer. (*) indicates a dataset-specific answer style. *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 in an automated manner, and *Q* represents 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., 2019a)	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

Table 2.3: Machine reading comprehension datasets published in 2019. *Ans* denotes answering styles where *MC* is multiple choice, *Desc* is description (free-form answering), *Ex* is answer extraction, i.e., selecting a span in the given context, *YN* is yes or no, and *NA* is no answer. (*) indicates a dataset-specific answer style. *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 in an automated manner, and *Q* represents search-engine queries.

Chapter 3

Evaluation Metrics

Knowing the quality of machine reading comprehension datasets is important for the development of natural-language understanding systems. In this chapter, two classes of metrics are adopted for evaluating machine reading comprehension datasets: prerequisite skills and readability. We apply these classes to six existing datasets, including MCTest and SQuAD, and highlight the characteristics of the datasets according to each metric and the correlation between the two classes. Our dataset analysis suggests that the readability of MRC datasets does not directly affect the question difficulty and that it is possible to create an MRC dataset that is easy to read but difficult to answer.

3.1 Introduction

A major goal of natural language processing (NLP) is to develop agents that can understand natural language. Such an ability can be tested with the machine reading comprehension (MRC) task that requires the agent to read open-domain documents and answer questions about them. Constructing systems with the competence of reading comprehension is challenging because reading comprehension comprises multiple processes including parsing, understanding cohesion, and inference with linguistic and general knowledge.

Clarifying what a system achieves is important in the development of MRC systems. To achieve robust improvement, systems should be measured according to a variety of metrics beyond simple accuracy. However, a current problem is that most MRC datasets are presented only with superficial categories, such as question types (e.g., what, where, and who) and answer types (e.g., numeric, location, and person). In addition, Chen et al. (2016) noted that some questions in datasets may not be suited to the testing of MRC systems. In such situations, it is difficult to obtain an accurate assessment of the MRC system.

Norvig (1989) argued that questions that are easy for humans to answer often turn out to be difficult for machines. For example, consider the two MRC questions in Figure 3.1. The first example is from SQuAD (Rajpurkar et al., 2016), although the document is taken from a Wikipedia article and was therefore written for adults. The question is answerable simply by noticing one sentence, without needing to fully understand the content of the text. On the other hand, consider the second example from MCTest (Richardson et al., 2013), which was written for children and is easy to read. Here, answering the question involves gathering information from multiple sentences and utilizing a combination of several skills, such as understanding causal relations (*Sara wanted...* → *they went to...*), coreference resolution (*Sara* and *Her*

ID: SQuAD, United_Methodist_Church
Context: The United Methodist Church (UMC) practices infant and adult baptism. Baptized Members are those who have been baptized as an infant or child, but who have not subsequently professed their own faith.
Question: What are members who have been baptized as an infant or child but who have not subsequently professed their own faith?
Answer: Baptized Members

ID: MCTest, mc160.dev.8
Context: Sara wanted to play on a baseball team. She had never tried to swing a bat and hit a baseball before. Her Dad gave her a bat and together they went to the park to practice.
Question: Why was Sara practicing?
Answer: She wanted to play on a team

Figure 3.1: Examples of MRC questions from SQuAD (Rajpurkar et al., 2016) and MCTest (Richardson et al., 2013) (the Contexts are excerpts).

Dad = *they*), and complementing ellipsis (*baseball team* = *team*). These two examples show that the readability of the text does not necessarily correlate with the difficulty of answering questions about it. Furthermore, the accompanying categories of existing MRC datasets cannot help with the analysis of this issue.

In this chapter, our goal is to investigate how these two types of difficulty, namely "answering questions" and "reading text," are correlated in MRC. Corresponding to each type, we formalize two classes of evaluation metrics, *prerequisite skills* and *readability*, and analyze existing MRC datasets. Our intention is to provide the basis of an evaluation methodology of MRC systems to help their robust development.

Our two classes of metrics are inspired by the analysis in McNamara and Magliano (2009) of human text comprehension in psychology. They considered two aspects of text comprehension, namely "strategic/skilled comprehension" and "text ease of processing."

Our first class defines metrics for "strategic/skilled comprehension," namely the difficulty of comprehending the context when answering questions. We adopted the set of prerequisite skills that Sugawara et al. (2017b) proposed for the fine-grained analysis of reading comprehension capability. Their study also presented an important observation of the relation between the difficulty of answering MRC questions and prerequisite skills: the more skills that are required to answer a question, the more difficult is the question. Based on this observation, in this chapter, we assume that the number of skills required to answer a question is a reasonable indication of the difficulty of the question. This is because each skill corresponds to one of the functions of an NLP system, which has to be capable of that functionality.

Our second class defines metrics for "text ease of processing," namely the difficulty of reading the text. We regard it as readability of the text in terms of syntactic and lexical complexity. From among readability studies in NLP, we adopt a wide range of linguistic features proposed by Vajjala and Meurers (2012), which can be used for texts with no available annotations.

The contributions of this chapter are as follows.

1. We adopt two classes of evaluation metrics to show the qualitative features of MRC datasets. Through analyses of MRC datasets, we demonstrate that there is

only a weak correlation between the difficulty of questions and the readability of context texts in MRC datasets.

2. We revise a previous classification of prerequisite skills for reading comprehension. Specifically, skills of knowledge reasoning are organized by using insights of entailment phenomena in NLP and human text comprehension in psychology.

3. We annotate six existing MRC datasets, compared to the two datasets considered in Sugawara and Aizawa (2016), with our organized metrics being used in the comparison.

We should note that, in this study, MRC datasets with different task formulations were annotated with prerequisite skills under the same conditions. Annotators first saw a context, a question, and its answer. They selected the sentences required to provide the answer, and then annotated them with appropriate prerequisite skills. That is, the datasets were annotated from the point of view of whether the context entailed the hypothesis constructed from the pair of the question and answer. This means that our methodology cannot quantify the systems' competence in searching the context for necessary sentences and answer candidates. In other words, our methodology can be only used to evaluate the competence of understanding MRC questions as *contextual entailments*.

The remainder of this chapter is divided into the following sections. First, we overview the psychological study of human reading comprehension in Section 3.2. Next, we specify our two classes of metrics in Section 3.3. In Section 3.4, we annotate existing MRC datasets with the prerequisite skills. Section 3.5 gives the results of our dataset analysis and Section 3.6 discusses their implications.

3.2 Reading Comprehension in Psychology

In psychology, there is a rich tradition of research on human text comprehension. The construction–integration (C–I) model (Kintsch, 1988) is one of the most basic and influential theories. This model assumes a connectional and computational architecture for text comprehension. It assumes that comprehension is the processing of information based on the following two steps.

1. *Construction:* read sentences or clauses as inputs; form and elaborate concepts and propositions corresponding to the inputs.

2. *Integration:* associate the contents to understand them consistently (e.g., coreference, discourse, and coherence).

During these steps, three levels of representation are constructed (van Dijk and Kintsch, 1983): the *surface code* (i.e., wording and syntax), the *textbase* (i.e., text propositions with cohesion), and the *situation model* (i.e., mental representation). Based on these assumptions, McNamara and Magliano (2009) proposed two aspects of text comprehension, namely "strategic/skilled comprehension" and "text ease of processing." We adopted these assumptions as the basis of our two classes of evaluation metrics (Section 3.3).

In an alternative approach, Kintsch (1993) proposed two dichotomies for the classification of human inferences, including the knowledge-based inference assumed in the C–I model. The first dichotomy is between inferences that are *automatic* and those that are *controlled*. However, Graesser et al. (1994) indicated that this distinction is ambiguous, because there is a continuum between the two states that depends on individuals. Therefore, this dichotomy is unsuited to empirical evaluation, which is our focus. The second dichotomy is between inferences that are *retrieved* and those that are *generated*. *Retrieved* means that the information used for inference is retrieved entirely from the context. In contrast, when inferences are *generated*, the reader uses external knowledge that goes beyond the context.

A similar distinction was proposed by McNamara and Magliano (2009), namely that between *bridging* and *elaboration*. A bridging inference connects current information to other information that has been encountered previously. Elaboration connects current information to external knowledge that is not included in the context. We use these two types of inference in the classification of knowledge reasoning.

3.3 Evaluation Metrics for Datasets

Following the depiction of text comprehension by McNamara and Magliano (2009), we adopted two classes for the evaluation of MRC datasets: *prerequisite skills* and *readability*.

For the prerequisite skills class (Section 3.3.1), we refined skills that were proposed by Sugawara et al. (2017b) and Sugawara and Aizawa (2016). However, a problem in these studies is that their categorization of knowledge reasoning was provisional and with a weak theoretical background.

Therefore, in this study, we reorganized the category of knowledge reasoning in terms of textual entailment in NLP and human text comprehension in psychology. In research on textual entailment, several methodologies have been proposed for the precise analysis of entailment phenomena (Dagan et al., 2013; LoBue and Yates, 2011). In psychology research, as described in Section 3.2, McNamara and Magliano (2009) proposed a similar distinction for inferences: *bridging* versus *elaboration*. We utilized these insights in developing a comprehensive but not overly specific classification of knowledge reasoning.

Our prerequisite skills class includes the *textbase* and *situation model* (van Dijk and Kintsch, 1983). In our terminology, this means understanding each fact and associating multiple facts in a text, such as the relations of events, characters, or the topic of a story. The skills also involve knowledge reasoning, which is divided into several metrics according to the distinctions of human inferences. This point is discussed by Kintsch (1993) and McNamara and Magliano (2009). It also accords with the classification of entailment phenomena by Dagan et al. (2013) and LoBue and Yates (2011).

Readability metrics (Section 3.3.2) are quantitative measures used to assess the difficulty of reading, with respect to vocabulary and the complexity of texts. In this study, they measure the competence in understanding the first basic representation of a text, called the *surface code* (van Dijk and Kintsch, 1983).

3.3.1 Prerequisite Skills

Based on the 10 skills in Sugawara et al. (2017b), we identified 13 prerequisite skills, which are presented below. (We use * and † to indicate skills that have been modified/elaborated from the original definition or have been newly introduced in this study, respectively.)

1. **Object tracking***: jointly tracking or grasping of multiple objects, including sets or memberships (Clark, 1975). This skill is a version of the *list/enumeration* used in the original classification, renamed to emphasize its scope with respect to multiple objects.

2. **Mathematical reasoning***: we merged statistical and quantitative reasoning with mathematical reasoning. This skill is a renamed version of *mathematical operations*.

3. **Coreference resolution***: this skill has a small modification to include an anaphora (Dagan et al., 2013). It is similar to *direct reference* (Clark, 1975).

4. **Logical reasoning***: we identified this skill as the understanding of predicate logic, e.g., conditionals, quantifiers, negation, and transitivity. Note that this skill, together with *mathematical reasoning*, is intended to align with the offline skills described by Graesser et al. (1994).

5. **Analogy***: understanding of metaphors including metonymy and synecdoche (see LoBue and Yates (2011) for examples of synecdoche.)

6. **Causal relation:** understanding of causality that is represented by explicit expressions such as "why," "because," and "the reason for" (only if they exist).

7. **Spatiotemporal relation:** understanding of spatial and/or temporal relationships between multiple entities, events, and states.

8. **Ellipsis†**: recognizing implicit/omitted information (argument, predicate, quantifier, time, or place). This skill is inspired by Dagan et al. (2013) and the discussion in Sugawara et al. (2017b).

9. **Bridging†**: inference supported by grammatical and lexical knowledge (e.g., synonymy, hypernymy, thematic role, part of events, idioms, and apposition). This skill is inspired by the concept of *indirect reference* in the literature (Clark, 1975). Note that we exclude *direct reference* because it is covered by *coreference resolution* (pronominalization) and *elaboration* (epithets).

10. **Elaboration†**: inference using known facts, general knowledge (e.g., kinship, exchange, typical event sequence, and naming), and implicit relations (e.g., noun compounds and possessives) (see Dagan et al. (2013) for details). *Bridging* and *elaboration* are distinguished by the knowledge used in inferences being grammatical/lexical or general/commonsense, respectively.

11. **Meta-knowledge†**: using knowledge that includes a reader, writer, or text genre (e.g., narratives and expository documents) from meta-viewpoints (e.g., *Who are the principal characters of the story?* or *What is the main subject of this*

article?). Although this skill can be regarded as part of *elaboration*, we defined it as an independent skill because this knowledge is specific to reading comprehension. We were motivated by the discussion in Smith et al. (2015).

12. **Schematic clause relation**: understanding of complex sentences that have coordination or subordination, including relative clauses.

13. **Punctuation***: understanding of punctuation marks (e.g., parenthesis, dash, quotation, colon, or semicolon). This skill is a renamed version of *special sentence structure*. Concerning the original definition, we regarded "scheme" in figures of speech as ambiguous and excluded it. We defined *ellipsis* as a independent skill, and apposition was merged into *bridging*. Similarly, understanding of constructions was merged into the idioms in *bridging*.

Note that whereas the first 11 skills involve multiple items, the final pair of skills involve only a single sentence. The skills from 8 to 11 define the four categories by refining the "commonsense reasoning" category proposed originally in Sugawara et al. (2017b). In addition, we did not construct this classification to be dependent on particular MRC systems in NLP. This was because our methodology is intended to be general and applicable to many kinds of architectures. For example, we did not consider the dichotomy between *automatic* and *controlled* inferences because the usage of knowledge is not necessarily the same for all MRC systems.

3.3.2 Readability Metrics

In this study, we evaluated the readability of texts based on metrics in NLP. Several studies have examined readability in various applications, such as second-language learning (Razon and Barnden, 2015) and text simplification (Aluisio et al., 2010), and from various aspects, such as development measures in second-language acquisition (Vajjala and Meurers, 2012) and discourse relations (Pitler and Nenkova, 2008).

Of these, we adopted the classification of linguistic features proposed by Vajjala and Meurers (2012). This was because they presented a comparison of a wide range of linguistic features focusing on second-language acquisition and their method can be applied to plain text. The classification in Pitler and Nenkova (2008) is more suited to measuring text quality. However, we could not use their results because we could not use discourse annotations.

We list the readability metrics in Table 3.1, which were reported by Vajjala and Meurers (2012) as the top 10 features that affect human readability. To classify these metrics, we can identify three classes: lexical features (*NumChar*, *NumSyll*, *AWL*, *AdvVar*, and *ModVar*), syntactic features (*MLS*, *CoOrd*, *DC/C*, and *CN/C*), and traditional features (*Coleman*). Academis Word List was taken from `http://en.wikipedia.org/wiki/Academic_Word_List`. We applied these metrics only to sentences that needed to be read in answering questions.

However, because these metrics were proposed for human readability, they do not necessarily correlate with those used in MRC systems. Therefore, in any system analysis, ideally we would have to consult a variety of features.

1.	Ave. # of characters per word (*NumChar*)
2.	Ave. # of syllables per word (*NumSyll*)
3.	Ave. sentence length (*MLS*)
4.	Proportion of words in AWL (*AWL*)
5.	Modifier variation (*ModVar*)
6.	# of coordinate phrases per sentence (*CoOrd*)
7.	Coleman–Liau index (*Coleman*)
8.	Dependent clause-to-clause ratio (*DC/C*)
9.	Complex nominals per clause (*CN/C*)
10.	Adverb variation (*AdvVar*)

Table 3.1: Readability metrics. *AWL* refers to the Academic Word List.

3.4 Annotation of Reading Comprehension Datasets

We annotated six existing MRC datasets with the prerequisite skills. We explain the annotation procedure in Section 3.4.1 and the annotated MRC datasets in Section 3.4.2.

3.4.1 Annotation Procedure

We prepared annotation guidelines according to Sugawara et al. (2017b). The guidelines include the definitions and examples of the skills and annotation instructions.

Four annotators were asked to simulate the process of answering questions in MRC datasets, using only the prerequisite skills, and to annotate questions with one or more skills required in answering. For each task in the datasets, the annotators saw simultaneously the context, question, and its answer. When a dataset contained multiple-choice questions, we showed all candidate answers and labeled the correct one with an asterisk. The annotators then selected the sentences that needed to be read to be able to answer the question and decided on the set of prerequisite skills required.

The annotators were allowed to select *nonsense* for unsolvable or unanswerable questions (e.g., the “coreference error” and “ambiguous” questions described in Chen et al. (2016)) to distinguish them from any solvable questions that required no skills.

3.4.2 Datasets

As summarized in Table 3.2, the annotation was performed on six existing MRC datasets: QA4MRE (Sutcliffe et al., 2013), MCTest (Richardson et al., 2013), SQuAD (Rajpurkar et al., 2016), Who-did-What (Onishi et al., 2016), MS MARCO (Nguyen et al., 2016), and NewsQA (Trischler et al., 2017). We selected these datasets to enable coverage of a variety of genres, query sourcing methods, and task formulations. From each dataset, we randomly selected 100 questions. This number was considered sufficient for the degree of analysis of datasets performed by Chen et al. (2016). The questions were sampled from the gold-standard dataset of QA4MRE and the development sets of the other datasets. We explain the method of choosing questions for the annotation as follows.

QA4MRE (Sutcliffe et al., 2013): the gold-standard dataset comprised four different topics and four documents for each topic. We randomly selected 100 main and auxiliary questions so that at least one question for each document was included.

Dataset	Domain	Query sourcing	Answering style
QA4MRE (Sutcliffe et al., 2013)	technical documents	handcrafted by experts	multiple choice
MCTest (Richardson et al., 2013)	narratives	crowdsourced	multiple choice
SQuAD (Rajpurkar et al., 2016)	Wikipedia articles	crowdsourced	answer extraction
Who-did-What (Onishi et al., 2016)	news articles	automated	multiple choice
MS MARCO (Nguyen et al., 2016)	web snippets	search engine queries	description
NewsQA (Trischler et al., 2017)	news articles	crowdsourced	answer extraction

Table 3.2: Analyzed MRC datasets, their domains, query sourcing methods, and answering styles.

MCTest (Richardson et al., 2013): this dataset comprised two sets: MC160 and MC500. Their development sets had 80 tasks in total, with each containing context texts and four questions. We randomly chose 25 tasks (100 questions) from the development sets.

SQuAD (Rajpurkar et al., 2016): this dataset included Wikipedia articles involving various topics, with the articles being divided into paragraphs. We randomly chose 100 paragraphs from 15 articles and used only one question from each paragraph for the annotation.

Who-did-What (WDW) (Onishi et al., 2016): this dataset was constructed from the English Gigaword newswire corpus (v5). Its questions were automatically created using a different article from that used for context. In addition, questions that could be solved by a simple baseline method were excluded from the dataset.

MS MARCO (MARCO) (Nguyen et al., 2016): each task in this dataset comprised several segments, one question, and its answer. We randomly chose 100 tasks (100 questions) and only used segments whose attribute was *is_selected* = 1 as context.

NewsQA (Trischler et al., 2017): we randomly chose questions that satisfied the following conditions: *is_answer_absent* = 0, *is_question_bad* = 0, and *validated_answers* do not include *bad_question* or *none*.

For a variety of reasons, there were other datasets we did not annotate in this study. CNN/Daily Mail (Hermann et al., 2015) is anonymized and contains errors, according to Chen et al. (2016), making it unsuitable for annotation. We considered CBTest (Hill et al., 2016) to be devised as language-modeling tasks rather than MRC-related tasks. LAMBADA (Paperno et al., 2016) texts are formatted for machine reading, with all tokens in lower case, which would seem to disallow inferences based on proper nouns and render them unsuitable for human reading and annotation.

Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1. Tracking	**11.0**	6.0	3.0	8.0	6.0	2.0
2. Math.	**4.0**	**4.0**	0.0	3.0	0.0	1.0
3. Coref. resol.	32.0	**49.0**	13.0	19.0	15.0	24.0
4. Logical rsng.	**15.0**	2.0	0.0	8.0	1.0	2.0
5. Analogy	**7.0**	0.0	0.0	**7.0**	0.0	3.0
6. Causal rel.	1.0	**6.0**	0.0	2.0	0.0	4.0
7. Sptemp rel.	**26.0**	9.0	2.0	2.0	0.0	3.0
8. Ellipsis	13.0	4.0	3.0	**16.0**	2.0	15.0
9. Bridging	**69.0**	26.0	42.0	59.0	36.0	50.0
10. Elaboration	**60.0**	8.0	13.0	57.0	18.0	36.0
11. Meta	**1.0**	**1.0**	0.0	0.0	0.0	0.0
12. Clause rel.	**52.0**	40.0	28.0	42.0	27.0	34.0
13. Punctuation	**34.0**	1.0	24.0	20.0	14.0	25.0
Nonsense	10.0	**1.0**	3.0	27.0	14.0	**1.0**

Table 3.3: Frequencies (%) of prerequisite skills needed for the MRC datasets.

3.5 Results of the Dataset Analysis

We now present the results of evaluating the MRC datasets according to the two classes of metrics. In the annotation of prerequisite skills, the inter-annotator agreement was 90.1% for 62 randomly sampled questions. The evaluation was performed with respect to the following four aspects: (i) frequencies of prerequisite skills required for each dataset; (ii) number of prerequisite skills required per question; (iii) readability metrics for each dataset; and (iv) correlation between readability metrics and the number of required prerequisite skills.

(i) Frequencies of prerequisite skills (Table 3.3). QA4MRE had the highest scores for frequencies among the datasets. This seems to reflect the fact that QA4MRE involves technical documents that contain a wide range of knowledge, multiple clauses, and punctuation. Moreover, the questions are devised by experts.

MCTest achieved a high score for several skills (best for *causal relation* and *meta-knowledge* and second-best for *coreference resolution* and *spatiotemporal relation*), but a low score for *punctuation*. These scores seem to be because the MCTest dataset consists of narratives.

Another dataset that achieved notable scores is Who-did-What. This dataset achieved the highest score for *ellipsis*. This is because the questions of Who-did-What are automatically generated from articles not used as context. This methodology tends to avoid textual overlap between a question and its context, thereby requiring frequently the skills of *ellipsis*, *bridging*, and *elaboration*.

With regard to *nonsense*, MS MARCO and Who-did-What received relatively high scores. This appears to have been caused by the automated sourcing methods, which may generate a separation between the contents of the context and question

#Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
0	2.0	18.0	27.0	2.0	15.0	13.0
1	13.0	36.0	33.0	5.0	35.0	26.0
2	13.0	24.0	24.0	14.0	29.0	23.0
3	20.0	15.0	6.0	22.0	6.0	25.0
4	14.0	4.0	6.0	16.0	2.0	9.0
5	13.0	1.0	1.0	6.0	0.0	2.0
6	10.0	1.0	0.0	6.0	0.0	1.0
7	1.0	0.0	0.0	2.0	0.0	0.0
8	1.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0
10	3.0	0.0	0.0	0.0	0.0	0.0
Ave.	**3.25**	1.56	1.28	2.43	<u>1.19</u>	1.99

Table 3.4: Frequencies (%) of the number of required prerequisite skills for the MRC datasets.

(i.e., web segments and a search query in MS MARCO, and a context article and question article in Who-did-What). In contrast, NewsQA had no nonsense questions. Although this result was affected by our question sampling, it is important to note that the NewsQA dataset includes annotations of meta-information whether or not a question makes sense (*is_question_bad*).

(ii) Number of required prerequisite skills (Table 3.4). QA4MRE had the highest score. On average, each question required 3.25 skills. There were few questions in QA4MRE that required zero or one skill, whereas such questions were contained more frequently in other datasets. Table 3.4 also indicates that more than 90% of the MS MARCO questions required fewer than three skills according to the annotation.

(iii) Readability metrics for each dataset (Table 3.5). SQuAD and QA4MRE achieved the highest scores for most metrics. This reflects the fact that Wikipedia articles and technical documents usually require a high-grade level of understanding. In contrast, MCTest had the lowest scores, with its dataset consisting of narratives for children.

(iv) Correlation between numbers of required prerequisite skills and readability metrics (Figures 3.2 and 3.3, and Table 3.6). Our main interest was in the correlation between prerequisite skills and readability. To investigate this, we examined the relation between the number of required prerequisite skills and readability metrics. We used the Flesch–Kincaid grade level (Kincaid et al., 1975) as an intuitive reference for readability. This value represents the typical number of years of education required to understand texts based on counts of syllables, words, and sentences.

Figures 3.2 and 3.3 show the relation between two values for each dataset and for each question, respectively. Figure 3.2 shows the trends of the datasets. QA4MRE

Metrics		QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1.	NumChar	5.026	3.892	**5.378**	4.988	5.016	5.017
2.	NumSyll	1.663	1.250	**1.791**	1.657	1.698	1.635
3.	MLS	28.488	11.858	23.479	**29.146**	19.634	22.933
4.	AWL	0.067	0.003	**0.071**	0.033	0.047	0.038
5.	ModVar	0.174	0.114	**0.188**	0.150	0.186	0.138
6.	CoOrd	**0.922**	0.309	0.722	0.467	0.651	0.507
7.	Coleman	12.553	4.333	**14.095**	12.398	11.836	12.138
8.	DC/C	**0.343**	0.223	0.243	0.254	0.220	0.264
9.	CN/C	1.948	0.614	1.887	**2.310**	1.935	1.702
10.	AdvVar	**0.038**	0.035	0.032	0.019	0.022	0.019
	F–K	14.953	3.607	14.678	**15.304**	12.065	12.624
	Words	**1545.7**	174.1	130.4	253.7	70.7	638.4

Table 3.5: Results of readability metrics for the MRC datasets. *F–K* is the Flesch–Kincaid grade level (Kincaid et al., 1975). *Words* is the average word count of the context for each question.

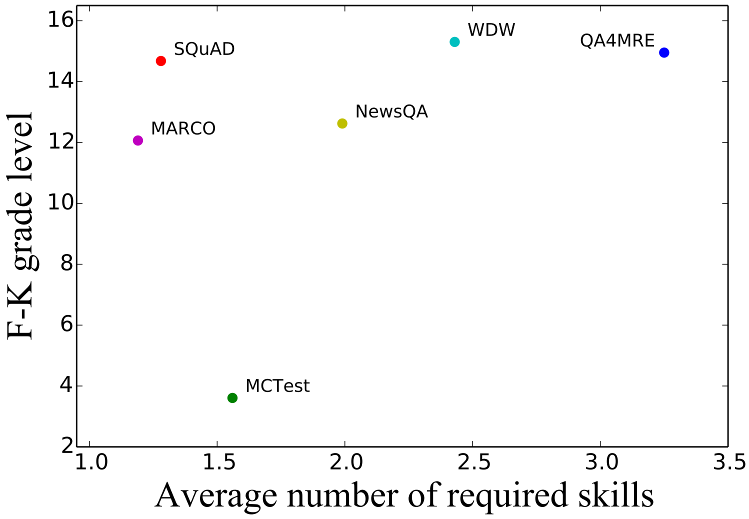

Figure 3.2: Flesch–Kincaid grade levels and average number of required prerequisite skills for the MRC datasets.

was relatively difficult both to read and to answer, whereas SQuAD was difficult to read but easy to answer. For further investigation, we selected three datasets (QA4MRE, MCTest, and SQuAD) and plotted all of their questions in Figure 3.3. Three separate domains can be seen.

Table 3.6 presents Pearson's correlation coefficients between the number of required prerequisite skills and each readability metric for all questions in the MRC datasets. Although there are weak correlations, from 0.025 to 0.416, these results demonstrate that there is not necessarily a strong correlation between the two values. This leads to the following two insights. First, the readability of MRC datasets does

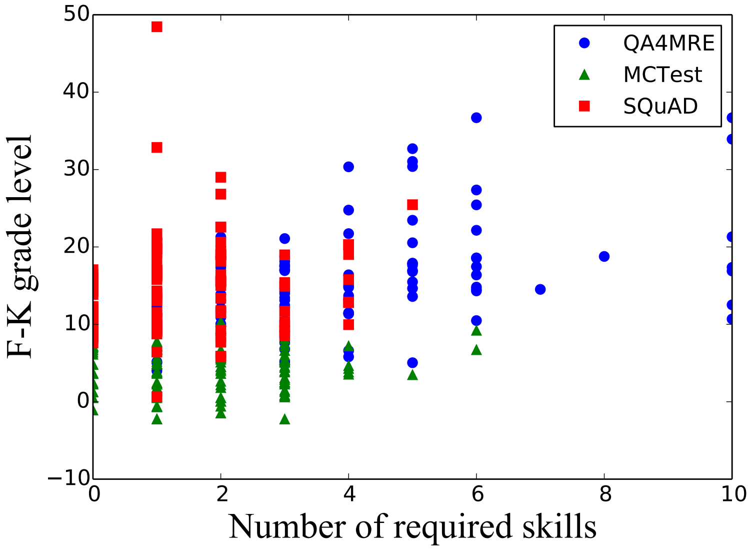

Figure 3.3: Flesch–Kincaid grade levels and number of required prerequisite skills for all questions in the selected MRC datasets.

	Metrics	r	p		Metrics	r	p
1.	NumChar	0.068	0.095	6.	CoOrd	0.166	0.000
2.	NumSyll	0.057	0.161	7.	Coleman	0.140	0.001
3.	MLS	0.416	0.000	8.	DC/C	0.188	0.000
4.	AWL	0.114	0.005	9.	CN/C	0.131	0.001
5.	ModVar	0.025	0.545	10.	AdvVar	0.026	0.515
	F–K	0.343	0.000		Words	0.355	0.000

Table 3.6: Pearson’s correlation coefficients (r) with the p-values (p) for the readability metrics and number of required prerequisite skills for all questions in the MRC datasets.

not directly affect the difficulty of their questions. That is, MRC datasets that are difficult to read are not necessarily difficult to answer. Second, it is possible to create difficult questions from the context that are easy to read. MCTest is a good example. The context texts in the MCTest dataset are easy to read, but the difficulty of its questions compares to that for the other datasets.

To summarize our results in terms of each dataset, we can make the following observations.

- **QA4MRE** is difficult both to read and to answer among the datasets analyzed. This would seem to follow its questions being devised by experts.
- **MCTest** is a good example of an MRC dataset that is easy to read but difficult to answer. We presume that this is because the corpus genre (i.e., narrative) reflects the trend in required skills for the questions.
- **SQuAD** is difficult to read, along with QA4MRE, but relatively easy to answer compared with the other datasets.

- **Who-did-What** performs well in terms of its query-sourcing method. Although its questions are created automatically, they are sophisticated in terms of knowledge reasoning. However, the automated sourcing method must be improved to exclude nonsense questions.
- **MS MARCO** is a relatively easy dataset in terms of prerequisite skills. However, one problem is that the dataset contained nonsense questions.
- **NewsQA** is advantageous in that it provides meta-information on the reliability of the questions. Such information enabled us to avoid using nonsense questions, as for the training of machine learning models.

3.6 Discussion

In this section, we discuss several issues regarding the construction of MRC datasets and the development of MRC systems using our methodology.

How to utilize the two classes of metrics for system development. One possible scenario for developing an MRC system is that it is first built to solve an easy-to-read and easy-to-answer dataset. The next step would be to improve the system so that it can solve an easy-to-read but difficult-to-answer dataset (or its converse). Finally, only after it can solve such datasets should the system be applied to difficult-to-read and difficult-to-answer datasets. The metrics of this study may be useful in preparing appropriate datasets for each step by measuring their properties. The datasets can then be ordered according to the grades of the metrics and applied to each step of the development, as in curriculum learning (Bengio et al., 2009) and transfer learning (Pan and Yang, 2010).

Corpus genre. Attention should be paid to the genre of the corpus used to construct a dataset. Expository documents such as news articles tend to require factorial understanding. Most existing MRC datasets use such texts because of their availability. On the other hand, narrative texts may have a closer correspondence to our everyday experience, involving the emotions and intentions of characters (Graesser et al., 1994). To build agents that work in the real world, MRC datasets may have to be constructed from narratives.

Question type. In contrast to factorial understanding, comprehensive understanding of natural language texts needs a better grasp of *global* coherence (e.g., the main point or moral of the text, the goal of a story, or the intention of characters) from the broad context (Graesser et al., 1994). Most questions in current use require only *local* coherence (e.g., referential relations and thematic roles) within a narrow context. An example of a question based on global coherence would be to give a summary of the text, as used in Hermann et al. (2015). It could be generated automatically by techniques of abstractive text summarization (Rush et al., 2015; Ganesan et al., 2010).

Annotation issues . We found questions for which there were disagreements regarding *nonsense* decisions. For example, some questions can be solved by external knowledge without even seeing their context. Therefore, we should clarify what constitutes a "solvable" or "reasonable" question for MRC. In addition, annotators

Sentence	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
# Required	1.120	**1.180**	1.040	1.110	1.080	1.170
Distance	**1.880**	0.930	0.090	0.730	0.280	0.540

Table 3.7: Average number and distance apart of sentences that need to be read to answer a question in the MRC datasets.

reported that the prerequisite skills did not easily treat questions whose answer was "none of the above" in QA4MRE. We considered these "no answer" questions difficult, in that systems have to decide not to select any of the candidate answers, and our methodology failed to specify them.

Competence in selecting necessary sentences. As mentioned in Section 3.1, our methodology cannot evaluate competence in selecting sentences that need to be read to answer questions. In a brief analysis, we further investigated sentences in the context of the datasets that were selected in the annotation. Analyses were performed in two ways. For each question, we counted the number of required sentences and their distance apart. The distance of sentences was calculated as follows. If a question required only one sentence to be read, its distance was zero. If a question required two adjacent sentences to be read, its distance was one. If a question required more than two sentences to be read, its distance was the sum of the distances of any two sentences. The first row of Table 3.7 gives the average number of required sentences per question for each MRC dataset. Although the scores are reasonably close, MCTest required multiple sentences to be read most frequently. The second row gives the average distance apart of the required sentences. QA4MRE required the longest distance because readers had to look for clues in the long context texts. In contrast, SQuAD and MS MARCO had lower scores. Most of their questions seemed to be answered by reading only a single sentence. Of course, the scores for distances will depend on the length of the context texts.

Metrics of reading comprehension for machines. Our underlying assumption in this study is that, in the development of interactive agents such as dialogue systems, it is important to make the systems behave in a human-like way. This has also become a distinguishing feature of recent MRC task design, and one that has never been explicitly considered in conventional NLP tasks. To date, the difference between human and machine reading comprehension has not attracted much research attention. We believe that our human-based evaluation metrics and analysis will help researchers to develop a method for the step-by-step construction of better MRC datasets and improved MRC systems.

3.7 Conclusion

In this study, we adopted evaluation metrics that comprise two classes, namely refined *prerequisite skills* and *readability*, for analyzing the quality of MRC datasets. We applied these classes to six existing datasets and highlighted their characteristics according to each metric. Our dataset analysis suggested that the readability of MRC datasets does not directly affect the difficulty of the questions and that it is possible to create an MRC dataset that is easy to read but difficult to answer.

Chapter 4

Question Quality

A challenge in creating a dataset for machine reading comprehension (MRC) is to collect questions that require a sophisticated understanding of language to answer beyond using superficial cues. In this chapter, we investigate what makes questions easier across recent 12 MRC datasets with three answering styles (answer extraction, description, and multiple choice). We propose to employ simple heuristics to split each dataset into *easy* and *hard* subsets and examine the performance of two baseline models for each of the subsets. We then manually annotate questions sampled from each subset with both validity and requisite reasoning skills to investigate which skills explain the difference between easy and hard questions. From this study, we observe that (i) the baseline performances for the hard subsets remarkably degrade compared to those of entire datasets, (ii) hard questions require knowledge inference and multiple-sentence reasoning in comparison with easy questions, and (iii) multiple-choice questions tend to require a broader range of reasoning skills than answer extraction and description questions. These results suggest that one might overestimate recent advances in MRC.

4.1 Introduction

Evaluating natural language understanding (NLU) systems is a long-established problem in AI (Levesque, 2014). One approach to doing so is the machine reading comprehension (MRC) task, in which a system answers questions about given texts (Hirschman et al., 1999). Although recent studies have made advances (Yu et al., 2018), it is still unclear to what precise extent questions require understanding of texts (Jia and Liang, 2017).

In this study, we examine MRC datasets and discuss what is needed to create datasets suitable for the detailed testing of NLU. Our motivation originates from studies that demonstrated unintended biases in the sourcing of other NLU tasks, in which questions contain simple patterns and systems can recognize these patterns to answer them (Gururangan et al., 2018; Mostafazadeh et al., 2017).

We conjecture that a situation similar to this occurs in MRC datasets. Consider the question shown in Figure 4.1, for example. Although the question, starting with *when*, requires an answer that is expressed as a moment in time, there is only one such expression (i.e., *November 2014*) in the given text (we refer to the text as the *context*). In other words, the question has only a single candidate answer. The system can solve it merely by *recognizing the entity type* required by *when*. In addition to this, even if another expression of time appears in other sentences, only one sentence

Article: Spectre (2015 film) on Wikipedia
Context: (s_1) In *November 2014*, **Sony Pictures** Entertainment was targeted by **hackers** who released details of confidential **e-mails** between **Sony** executives regarding [...]. (s_2) Included within these were several memos relating to the production [...]. (s_3) Eon Productions later issued a statement [...].
Question: When $_{(k=1)}$ did hackers get into the Sony Pictures e-mail system?
Prediction for the full question: *November 2014*
Prediction for the $k = 1$ question: *November 2014*
Uni-gram overlaps between s_i and the question: s_1: 5, s_2: 0, s_3: 0

Figure 4.1: Example from SQuAD (Rajpurkar et al., 2016). The baseline system can answer the token-limited question and, even if there are other candidate answers, it can easily attend to the answer-contained sentence (s_1) by watching word overlaps.

(i.e., s_1) appears to be related to the question; thus, the system can easily determine the correct answer by *attention*, that is, by matching the words appearing both in the context and the question. Therefore, this kind of question does not require a complex understanding of language—e.g., multiple-sentence reasoning, which is known as a more challenging task (Richardson et al., 2013).

In Section 4.3, we define two heuristics, namely *entity-type recognition* and *attention*. We specifically analyze the differences in the performance of baseline systems for the following two configurations: (i) questions answerable or unanswerable with the first k tokens; and (ii) questions whose correct answer appears or does not appear in the context sentence that is most similar to the question (henceforth referred to as *the most similar sentence*). Although similar heuristics are proposed by Weissenborn et al. (2017), ours are utilized for question filtering, rather than system development; Using these simple heuristics, we split each dataset into *easy* and *hard* subsets for further investigation of the baseline performance.

After conducting the experiments, we analyze the following two points in Section 4.4. First, we consider which questions are valid for testing, i.e., reasonably solvable. Second, we consider what reasoning skills are required and whether this exposes any differences among the subsets. To investigate these two concerns, we manually annotate sample questions from each subset in terms of validity and required reasoning skills, such as word matching, knowledge inference, and multiple sentence reasoning.

We examine 12 recently proposed MRC datasets (Table 4.1), which include answer extraction, description, and multiple-choice styles. We also observe differences based on these styles. For our baselines, we use two neural-based systems, namely, the Bidirectional Attention Flow (Seo et al., 2017) and the Gated-Attention Reader (Dhingra et al., 2017a).

In Section 4.5, we describe the advantages and disadvantages of different answering styles with regard to evaluating NLU systems. We also interpret our heuristics for constructing realistic MRC datasets.

Our contributions are as follows:

- This study is the first large-scale investigation across recent 12 MRC datasets with three answering styles.
- We propose to employ simple heuristics to split each dataset into *easy* and *hard* subsets and examine the performance of two baseline models for each of the subsets.

Answer extraction (select a context span)	
1.	SQuAD (v1.1) (Rajpurkar et al., 2016)
2.	AddSent (Jia and Liang, 2017)
3.	NewsQA (Trischler et al., 2017)
4.	TriviaQA (Wikipedia set) (Joshi et al., 2017)
5.	QAngaroo (WikiHop) (Welbl et al., 2018)
Description (generate a free-form answer)	
6.	MS MARCO (v2) (Nguyen et al., 2016)
7.	NarrativeQA (summary) (Kočiský et al., 2018)
Multiple choice (choose from multiple options)	
8.	MCTest (160 + 500) (Richardson et al., 2013)
9.	RACE (Middle + High) (Lai et al., 2017)
10.	MCScript (Ostermann et al., 2018)
11.	ARC Easy (ARC-E) (Clark et al., 2018)
12.	ARC Challenge (ARC-C) (Clark et al., 2018)

Table 4.1: Examined datasets.

- We manually annotate questions sampled from each subset with both validity and requisite reasoning skills to investigate which skills explain the difference between easy and hard questions.

We observed the following:

- The baseline performances for the hard subsets remarkably degrade compared to those of entire datasets.
- Our annotation study shows that hard questions require knowledge inference and multiple-sentence reasoning in comparison with easy questions.
- Compared to questions with answer extraction and description styles, multiple-choice questions tend to require a broader range of reasoning skills while exhibiting answerability, multiple answer candidates, and unambiguity.

These findings suggest that one might overestimate recent advances in MRC systems. They also emphasize the importance of considering simple answer-seeking heuristics when sourcing questions, in that a dataset could be easily biased unless such heuristics are employed.

4.2 Examined Datasets and Baselines

4.2.1 Datasets

We analyzed 12 MRC datasets with three answering styles: answer extraction, description, and multiple choice (Table 4.1). Our aim was to select datasets varying in terms of corpus genre, context length, and question sourcing methods. The ARC Easy and Challenge were collected using different methods; hence, we treated them as different datasets (see Clark et al. (2018) for further details). Other datasets that are not covered in our study, but can be analyzed using the same method (see Chapter 2).

4.2.2 Baseline Systems

We employed the following two widely used baselines.

- **Bidirectional Attention Flow (BiDAF)** (Seo et al., 2017) was used for the answer extraction and description datasets. BiDAF models bi-directional attention between the context and question. It achieved state-of-the-art performance on the SQuAD dataset.
- **Gated-Attentive Reader (GA)** (Dhingra et al., 2017a) was used for the multiple-choice datasets. GA has a multi-hop architecture with an attention mechanism. It achieved state-of-the-art-performance on the CNN/Daily Mail and Who-did-What datasets.

Why we used different baseline systems. The multiple-choice style can be transformed to answer extraction, as mentioned in Clark et al. (2018). However, in some datasets, many questions have no textual overlap to determine the correct answer span in the context. Therefore, in order to avoid underestimating the baseline performance of those datasets, we used the GA system which is applicable to multiple choice questions.

We scored the performance using exact match (EM)/F1 (Rajpurkar et al., 2016), Rouge-L (Lin, 2004), and accuracy for the answer extraction, description, and multiple-choice datasets, respectively (henceforth, we refer to these collectively as the *score*, for simplicity). For the description datasets, we determined in advance the answer span of the context that gives the highest Rouge-L score to the human-generated gold answer. We computed the Rouge-L score between the predicted span and the gold answer. We used the official evaluation scripts of SQuAD and MS MARCO to compute the EM/F1 and Rouge-L, respectively.

Reproduction of the baseline performance. We used the same architecture as the official baseline systems unless specified otherwise. All systems were trained on the training set and tested on the development/test set of each dataset. We show the baseline performance of both the official results and those from our implementations in Tables 4.4, 4.5 and 4.6. Our implementations outperformed or showed comparable performance to the official baseline on most datasets. However, in TriviaQA, MCTest, RACE, and ARC-E, our baseline performance did not reach that of the official baseline, due to differences in architecture or the absence of reported hyperparameters in the literature.

Hyperparameters. We used different hyperparameters for each dataset because of the different characteristics of the datasets, e.g., the context length. Tables 4.2 and 4.3 show the hyperparameters.

Dataset	b	h	d	q
SQuAD	60	100	400	20
AddSent	60	100	400	20
NewsQA	32	100	1000	20
TriviaQA	32	100	400	20
QAngaroo	16	50	4096	20
MARCO	20	40	1600	30
NarrativeQA	60	50	1000	20

Table 4.2: Hyperparameters (batch size b, hidden layer size h, document size threshold d, question size threshold q) of the Bidirectional Attention Flow (Seo et al., 2017) for each dataset. The other settings basically followed the original implementation. In TriviaQA, we followed a method for the dataset preparation used in Joshi et al. (2017).

Dataset	b	h	n	dr	lr
MCTest	10	32	1	0.5	0.01
RACE	32	128	1	0.2	0.1
MCScript	25	64	1	0.5	0.2
ARC-E	32	256	1	0.5	0.3
ARC-C	32	256	1	0.5	0.3

Table 4.3: Hyperparameters (batch size b, hidden layer size h, number of attention layers n, dropout rate dr, learning rate lr) of the Gated-Attentive Reader (Dhingra et al., 2017a) for each dataset. The other settings basically followed the implementation in Lai et al. (2017).

4.3 Two Filtering Heuristics

The first goal of this chapter is to determine whether there are unintended biases of the kind exposed in Figure 4.1 in MRC datasets. We examined the influence of the two filtering heuristics: (i) entity type recognition (Section 4.3.1) and (ii) attention (Section 4.3.2). We then investigated the performance of the baseline systems on the questions filtered by the defined heuristics (Section 4.3.3).

4.3.1 Entity Type-based Heuristic

The aim of this heuristic was to detect questions that can be solved based on (i) the existence of a single candidate answer that is restricted by expressions such as "wh-" and "how many," and (ii) lexical patterns that appear around the correct answer. Because the query styles are not uniform across datasets (e.g., MARCO uses search engine queries), we could not directly use interrogatives. Instead, we simply provided the first k tokens of questions to the baseline systems. We chose smaller values for k than the (macro) average of the question length across the datasets (= 12.2 tokens). For example, for $k = 4$ of the question *will I qualify for OSAP if I'm new in Canada* (excerpted from MARCO), we use *will I qualify for*. Even if the tokens do not have an interrogative, the system may recognize lexical patterns around the correct answer. Questions that can be solved by examining these patterns were also of interest when

Dataset	SQuAD	AddSent	NewsQA	TriviaQA	QAngaroo
Statistics					
Answering style (metrics)	answer extraction (exact match / F1)				
Question sourcing	reading context	reading context	reading headline	trivia / quiz	chaining knowledge[1]
Context genre	Wikipedia	Wikipedia	news	Wikipedia	Wikipedia
Split examined	dev	dev	test	dev[2]	dev
# questions	10570	3560	5126	430	5129
Avg. # context tokens	150.1	163.3	698.8	783.4	1545.5
Avg. # question tokens	11.8	12.3	8.0	19.0	3.6
Avg. # sents in context	5.2	5.8	30.3	28.5	57.2
Baseline performance					
Official baseline	67.7/77.3	28.2/34.3	34.1/48.2	47.5/53.7	42.9/-
Our BiDAF baseline	**67.9/77.2**	**42.6/50.4**	**40.2/56.4**	**44.0/49.3**	**43.8/49.3**
Q first tokens (k=4)	30.7/44.6	19.2/29.7	30.4/44.4	20.5/25.0	43.6/49.1
(k=2)	14.0/25.0	9.4/17.8	19.4/30.3	14.4/18.5	42.6/48.0
(k=1)	7.0/14.9	4.2/10.6	13.5/23.8	8.6/12.5	42.0/47.5
% of # Q ($\geq$0.5 for k=2)	22.4	15.8	29.7	20.0	49.8
Ans in sim sent	71.4/80.6	50.2/58.2	42.9/59.7	58.0/65.1	41.7/49.2
only with sim sent	73.3/82.8	71.4/81.1	52.8/70.9	64.8/72.7	66.7/74.2
Ans not in sim sent	56.6/66.4	28.1/35.5	37.8/53.5	40.4/45.2	43.9/49.3
% of # Q (ans in sim)	76.3	65.7	46.3	20.5	4.2
***Hard* subset**	**38.7/45.2**	**18.2/23.4**	**27.9/40.9**	**30.0/32.5**	**2.3/2.6**
% of *hard*	**15.7**	**25.4**	**30.0**	**59.8**	**36.9**

Table 4.4: Statistics from the answer extraction datasets and their baselines. *Dev* represents a development set. *Ans in sim sent* refers to questions whose answer appears in the sentence that is most similar to the question. [1]The questions are not complete sentences and may start with more specific words than interrogatives. [2]Verified set.

filtering.

Results. Tables 4.4, 4.5 and 4.6 present the results for $k = 1, 2, 4$. In addition, to know the exact ratio of the questions that are solved rather than the scores for the answer extraction and description styles, we counted questions with $k = 2$ that achieved the score ≥ 0.5. We considered that this threshold is sufficient to judge that the system attends to the correct span because of the potential ambiguity of these styles (see Section 4.4). As k decreased, so too did the baseline performance on all datasets in Tables 4.4 and 4.5 except QAngaroo. By contrast, in QAngaroo and the multiple-choice datasets, the performance did not degrade so strongly. In particular, the difference between the scores on the full and $k = 1$ questions in QAngaroo was 1.8. Because the questions in QAngaroo are not complete sentences, but rather knowledge-base entries that have a blank, such as *country_of_citizenship Henry VI of England,* this

Dataset	MARCO	NarraQA
Statistics		
Answering style (metrics)	description (Rouge-L)	
Question sourcing	search query[1]	reading summary
Context genre	web snippet	moviescript
Split examined	dev	test
# questions	55578[2]	10557
Avg. # context tokens	625.7	664.5
Avg. # question tokens	6.1	9.9
Avg. # sents in context	31.5	27.6
Baseline performance		
Official baseline	17.7[3]	36.30
Our BiDAF baseline	**36.42**[2]	**43.66**
Q first tokens (k=4)	32.61	25.23
(k=2)	25.13	13.00
(k=1)	21.67	8.45
% of # Q ($\geq$0.5 for k=2)	17.9	10.3
Ans in sim sent	38.96	45.17
only with sim sent	45.30	58.56
Ans not in sim sent	35.84	41.99
% of # Q (ans in sim)	18.6	52.6
***Hard* subset**	**15.42**	**39.61**
% of *hard*	**12.5**	**28.2**

Table 4.5: Statistics from the description datasets and their baselines. *Dev* represents a development set. *Ans in sim sent* refers to questions whose answer appears in the sentence that is most similar to the question. [1]The questions are not complete sentences and may start with more specific words than interrogatives. [2]No answer questions were removed. [3]The Passage Ranking model (Nguyen et al., 2016).

result implies that the baseline system can infer the answer merely by the first token of questions, i.e., the type of knowledge-base entry.

In most multiple-choice datasets, the $k = 1$ scores were significantly higher than random-choice scores. Given that multiple-choice questions offer multiple options that are of valid entity/event types, this gap was not necessarily caused by the limited number of candidate answers, as in the case with the answer extraction datasets. Therefore, we inferred that in the solved questions, incorrect options appeared less than the correct option did or did not appear at all in the context (such questions were regarded as solvable exclusively using the word match skill, which we analyzed in Section 4.4). Remarkably, although we failed to achieve a higher baseline performance, the score for the complete questions in MCTest was lower than that of the $k = 1$ questions. This result showed that the MCTest questions were sufficiently

Dataset	MCTest	RACE	MCScript	ARC-E	ARC-C
Statistics					
Answering style (metrics)	multiple choice (accuracy)				
Question sourcing	reading context	English exam	script scenario	science exam	
Genre	narrative	various	narrative	textbook	
Split examined	test	test	dev	dev	dev
# questions	840	4934	1411	2376	1171
Avg. # context tokens	249.9	339.3	195.2	142.0	138.3
Avg. # question tokens	9.4	11.5	7.8	21.8	25.4
Avg. # sents in context	18.4	17.9	11.5	8.1	8.2
Baseline performance					
Random	25.0	25.0	50.0	25.0	25.0
Official baseline	43.2[1]	44.1	72.0	62.6[2]	20.3[2]
Our GA baseline	**34.3**	**42.7**	**75.5**	**43.9**	**30.1**
Q tokens (k=4)	36.1	38.4	73.7	38.8	30.6
(k=2)	33.9	37.7	71.1	37.0	29.0
(k=1)	34.9	36.4	70.9	35.3	28.6
Ans in sim sent	33.1	40.8	74.0	47.5	31.6
only w/ sim	32.4	40.4	74.4	48.5	28.9
Ans not in sim	34.9	43.3	75.8	40.4	29.4
% of # Q (in sim)	33.5	23.2	17.7	48.7	34.8
Hard **subset**	**4.3**	**23.5**	**28.7**	**20.6**	**15.6**
% of ***hard***	**64.2**	**58.8**	**27.1**	**53.9**	**66.4**

Table 4.6: Statistics from the multiple-choice datasets and their baselines. [1]The Attentive Reader (Hermann et al., 2015) from Yin et al. (2016). [2]An information retrieval system from Clark et al. (2018).

difficult such that it was not especially useful for the baseline system to consider the entire question statement.

4.3.2 Attention-based Heuristic

Next, we examined in each dataset (i) how many questions have their correct answers in the most similar sentence and (ii) whether a performance gap exists for such questions (i.e., whether such questions are easier than the others).

We used uni-gram overlap as a similarity measure. Although there are other similarity measures, we used this basic measure to obtain an intuitive result. We counted how many times question words appeared in each sentence, where question words were stemmed and stopwords were dropped. We then checked whether the correct answer appeared in the most similar sentence. For the multiple-choice datasets, we selected the text span that provided the highest Rouge-L score with the correct option as the correct answer.

Results. Tables 4.4, 4.5 and 4.6 show the results. Considering the average number of context sentences, most datasets contained a significantly high proportion of questions whose answers were in the most similar sentence.

In the answer extraction and description datasets, except QAngaroo, the baseline performance improved when the correct answer appeared in the most similar sentence, and gaps were found between the performances on these questions and the others. These gaps indicated that the dataset may lack balance for testing NLU. If these questions tend to require the word matching skill exclusively, attending the other portion is useful in studying a more realistic NLU, e.g., common-sense reasoning and discourse understanding. Therefore, we investigated whether these questions merely require word matching (see Section 4.4).

Meanwhile, in the first three multiple-choice datasets, the performance differences were marginal or inversed, implying that although the baseline performance was not especially high, the difficulty of these questions for the baseline system was not affected by whether their correct answers appeared in the most similar sentence.

We further analyzed the baseline performance after removing the context and leaving only the most similar sentence. In AddSent and QAngaroo, the scores remarkably improved (>20 F1). From this result, we can infer that on these datasets the baseline systems were distracted by other sentences in the context. This observation was supported by the results from the AddSent dataset (Jia and Liang, 2017), which contains manually injected distracting sentences (i.e., adversarial examples).

4.3.3 Performance on *Hard* Subsets

In the previous two sections, we observed that in the examined datasets (i) some questions were solved by the baseline systems merely with the first k tokens and/or (ii) the baseline performances increased for questions whose answers were in the most similar sentence. We were concerned that these two will become dominant factors in measuring the baseline performance using the datasets; Hence, we split each development/test set into *easy* and *hard* subsets for further investigation.

***Hard* subsets.** A *hard* subset comprised questions (i) whose score is not positive when $k = 2$ *and* (ii) whose correct answer does not appear in the most similar sentence. The easy subsets comprised the remaining questions. We aimed to investigate the gap of the performance values between the *easy* and *hard* subsets. If the gap is large, the dataset may be strongly biased toward questions that are solved by recognizing entity types or lexical patterns and may not be suitable for measuring the system's ability for complex reasoning.

Results and clarification. The bottom row of Tables 4.4, 4.5 and 4.6 shows that the baseline performances on the *hard* subset remarkably decreased in almost all examined datasets. These results revealed that we may overestimate the ability of the baseline systems previously perceived. However, we clarify that our intention is not to remove the questions solved or mitigated by our defined heuristics to create a new *hard* subset because this may generate new biases as indicated in Gururangan et al. (2018). Rather, we would like to emphasize the importance of the defined heuristics when sourcing questions. Indeed, ill attention to these heuristics can lead to unintended biases.

4.4 Annotating Question Validity and Required Skills

4.4.1 Annotation Specifications

Objectives. To complement the observations in the previous sections, we annotated sampled questions from each subset of the datasets. Our motivation can be summarized as follows: (i) How many questions are valid in each dataset? That is, the *hard* questions may not in fact be hard, but just unsolvable, as indicated in Chen et al. (2016). (ii) What kinds of reasoning skills explain the *easy*/*hard* questions? (iii) Are there any differences among the datasets and the answering styles?

We annotated the minimum skills required to choose the correct answer among other candidates. We assumed that the solver knows what type of entity or event is entailed by the question.

Annotation labels. Our annotation labels (Table 4.7) were inspired by previous works such as Chen et al. (2016), Trischler et al. (2017), and Lai et al. (2017). The major modifications were twofold: (i) detailed question validity, including a number of reasonable candidate answers and answer ambiguity, and (ii) posing multiple-sentence reasoning as a skill compatible with other skills.

Reasoning types indeed have other classifications. For instance, Lai et al. (2017) defined five reasoning types, including *attitude analysis* and *whole-picture reasoning*. We incorporated them into the *knowledge* and *meta/whole* classes. Clark et al. (2018) proposed detailed knowledge and reasoning types, but these were specific to science exams and, thus, omitted from our study.

Independent of the abovementioned reasoning types, we checked whether the question required multiple-sentence reasoning to answer the questions. As another modification, we extended the notion of "sentence" in our annotation and considered a subordinate clause as a sentence. This modification was intended to deal with the internal complexity of a sentence with multiple clauses, which can also render a question difficult.

Settings. For each subset of the datasets, 30 questions were annotated. Therefore we obtained annotations for $30 \times 2 \times 12 = 720$ questions. The annotation was performed by the authors. The annotator was given the context, question, and candidate answers for multiple-choice questions along with the correct answer. To reduce bias, the annotator did not know which *easy* or *hard* subset the questions were in, and was not told the predictions and scores of the respective baseline systems.

4.4.2 Annotation Results

Tables 4.8, 4.9, and 4.10 show the annotation results.

Validity

1. *Unsolvable*
   The context coupled with the question does not reasonably give the answer.
2. *Single candidate*
   The question does not have multiple candidate answers.
3. *Ambiguous*
   The question does not have a unique, decidable answer, or, multiple possible answers are not covered by the gold answers.

Reasoning skill

4. *Word matching*
   Matching the context and question words.
5. *Paraphrasing*
   Using lexical and grammatical knowledge.
6. *Knowledge*
   Inference using commonsense and/or world knowledge.
7. *Meta/Whole*
   Understanding meta terms, such as "author" and "writer," and comprehending the general context.
8. *Math/Logic*
   Using mathematical and logical knowledge, includeing multiple-choice questions that ask "which option is not true."

Multiple-sentence reasoning

9. (i) *Coreference* (ii) *Causal relation* (iii) *Spatial–temporal relations* (iv) *None*
   Gathering cues from multiple sentences/clauses.

Table 4.7: Annotation labels. One of the reasoning skills is annotated with the questions that are "no" in all validity labels. Multiple sentence reasoning is independent of reasoning skills and annotated with all valid questions.

Validity. TriviaQA, QAngaroo, and ARCs revealed a relatively high *unsolvability*, which seemed to be caused by the unrelatedness between the questions and their context. For example, QAngaroo's context was gathered from Wikipedia articles that were not necessarily related to the questions. Nonetheless, it is remarkable that even though the dataset was automatically constructed, the remaining valid *hard* questions were difficult for the baseline system. The context passages in ARCs were curated from textbooks that may not provide sufficient information to answer the questions. Our analysis was not intended to undermine the quality of these questions. We refer readers to Clark et al. (2018). Note that it is possible for unsolvable questions to be permitted, and that the system must indicate them in some datasets, such as QA4MRE, NewsQA, MARCO, and SQuAD (v2.0).

However, for *single candidate*, we found that few questions had only single-candidate answers. Furthermore, there were even fewer single-candidate answers in AddSent than in SQuAD. This result supported the claim that the adversarial examples

	Dataset	SQuAD		AddSent		NewsQA		TriviaQA		QAngaroo	
	Subset	easy	hard	easy	hard	easy	hard	easy	hard	easy	hard
	F1	80.9	37.6	61.5	29.5	52.7	30.3	70.6	33.4	71.1	3.5
Validity	Unsolvable	0.0	0.0	0.0	0.0	0.0	6.7	16.7	16.7	33.3	43.3
	Single cand.	23.3	10.0	6.7	3.3	10.0	3.3	3.3	6.7	6.7	3.3
	Ambiguous	3.3	13.3	3.3	13.3	43.3	30.0	13.3	13.3	13.3	20.0
	Valid	73.3	76.7	90.0	83.3	46.7	60.0	66.7	63.3	46.7	33.3
Skill	Word match	59.1	21.7	55.6	24.0	42.9	66.7	45.0	26.3	35.7	20.0
	Paraphrasing	18.2	26.1	11.1	36.0	21.4	11.1	5.0	10.5	7.1	20.0
	Knowledge	22.7	47.8	33.3	40.0	35.7	22.2	50.0	63.2	57.1	60.0
	Meta/Whole	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Math/Logic	0.0	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Relation	Multi sent.	22.7	17.4	25.9	36.0	35.7	16.7	35.0	36.8	57.1	80.0
	Coreference	18.2	17.4	14.8	32.0	21.4	16.7	35.0	31.6	50.0	50.0
	Causal	0.0	0.0	3.7	4.0	0.0	0.0	0.0	0.0	0.0	0.0
	Space/Temp.	4.5	0.0	7.4	0.0	14.3	0.0	0.0	5.3	7.1	30.0

Table 4.8: Annotation results for the answer extraction datasets.

ID	./cnn/stories/4ca29639845a40551a62d10212a46aec7caf3369.story-2
Context	[...] This plot of land is scheduled to house the permanent United Airlines Flight 93 memorial. [...]
Question	What was the name of the flight?
Answer	93
Possible answers	United Airlines Flight 93, Flight 93

Figure 4.2: Example of an *ambiguous* question from NewsQA (Trischler et al., 2017).

augmented the number of possible candidate answers, thereby degrading the baseline performance.

In our annotation, *ambiguous* questions were found to be those with multiple correct spans. Figure 4.2 shows an example. In this case, several answers aside from “93” were correct. Ambiguity is an important feature insofar because it can lead to unstable scoring in EM/F1.

The multiple-choice datasets mostly comprised valid questions, with the exception of the unsolvable questions in the ARC datasets.

Reasoning skills. We can see that *word matching* was more important in the *easy* subsets, and *knowledge* was more pertinent to the *hard* subsets in 10 of the 12 datasets. These results confirmed that the manner by which we split the subsets was successful at filtering questions that were relatively easy in terms of reasoning skills. However, we did not observe this trend with *paraphrasing*, which seemed difficult to distinguish from *word matching* and *knowledge*. With regard to *meta/whole* and *math/logic*, we can see that these skills were needed less in the answer extraction and description datasets. They were more pertinent to the multiple-choice datasets.

	Dataset	MARCO		NarraQA	
	Subset	easy	hard	easy	hard
	Rouge-L	49.4	21.5	54.9	51.2
Validity	Unsolvable	0.0	0.0	0.0	0.0
	Single cand.	0.0	0.0	6.7	0.0
	Ambiguous	6.7	3.3	0.0	0.0
	Valid	93.3	96.7	93.3	100.0
Skill	Word match	89.3	44.8	46.4	43.3
	Paraphrasing	0.0	10.3	25.0	20.0
	Knowledge	10.7	44.8	28.6	33.3
	Meta/Whole	0.0	0.0	0.0	3.3
	Math/Logic	0.0	0.0	0.0	0.0
Relation	Multi sent.	7.1	13.8	28.6	46.7
	Coreference	7.1	13.8	14.3	33.3
	Causal	0.0	0.0	14.3	6.7
	Space/Temp.	0.0	0.0	0.0	6.7

Table 4.9: Annotation results for the description datasets.

	Dataset	MCTest		RACE		MCScript		ARC-E		ARC-C	
	Subset	easy	hard	easy	hard	easy	hard	easy	hard	easy	hard
	Accuracy	83.3	13.3	76.7	30.0	93.3	26.7	60.0	16.7	43.3	10.0
Validity	Unsolvable	0.0	0.0	0.0	0.0	0.0	0.0	3.3	30.0	46.7	33.3
	Single cand.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Ambiguous	0.0	0.0	3.3	0.0	0.0	0.0	3.3	0.0	3.3	3.3
	Valid	100.0	100.0	96.7	100.0	100.0	100.0	93.3	70.0	50.0	63.3
Skill	Word match	56.7	46.7	17.2	6.7	36.7	46.7	71.4	52.4	33.3	15.8
	Paraphrasing	6.7	10.0	13.8	6.7	20.0	6.7	14.3	19.0	20.0	31.6
	Knowledge	30.0	26.7	34.5	43.3	20.0	36.7	14.3	23.8	40.0	42.1
	Meta/Whole	3.3	3.3	31.0	33.3	20.0	10.0	0.0	0.0	0.0	0.0
	Math/Logic	3.3	13.3	3.4	10.0	3.3	0.0	0.0	4.8	6.7	10.5
Relation	Multi sent.	46.7	73.3	58.6	76.7	0.0	30.0	7.1	14.3	0.0	10.5
	Coreference	33.3	56.7	44.8	60.0	0.0	16.7	7.1	9.5	0.0	0.0
	Causal	6.7	6.7	3.4	13.3	0.0	3.3	0.0	0.0	0.0	0.0
	Space/Temp.	6.7	10.0	10.3	3.3	0.0	10.0	0.0	4.8	0.0	10.5

Table 4.10: Annotation results for the multiple-choice datasets.

Multiple-sentence reasoning. Multiple-sentence reasoning was more correlated with the *hard* subsets in 10 of the 12 datasets. Although NewsQA showed the inverse tendency for *word matching*, *knowledge*, and *multiple-sentence reasoning*, we suspect that this was caused by annotation variance and filtering a large portion of ambiguous

Label	r	p
Single cand (BiDAF)	0.150	0.002
Ambiguous (BiDAF)	0.098	0.044
Word matching (BiDAF)	0.266	0.000
Knowledge (BiDAF)	-0.288	0.000
Multi sent (BiDAF)	-0.120	0.035
Unsolvable (GA)	-0.119	0.039

Table 4.11: Pearson's correlation coefficients (r) between the annotation labels and the baseline scores with $p < 0.05$.

questions. For relational types, we did not see a significant trend in any particular type.

Correlation of labels and baseline scores. Across all examined datasets, we analyzed the correlations between the annotation labels and the scores of each baseline system in Table 4.11. In spite of the small size of the annotated samples, we derived statistically significant correlations for six labels. These results confirmed that BiDAF performed well for the *word matching* questions and relatively poorly with the *knowledge* questions. By contrast, we did not observe this trend in GA.

4.5 Discussion

In this section, we discuss the advantages and disadvantages of the answering styles. We also interpret the defined heuristics in terms of constructing more realistic MRC datasets.

Differences among the answering styles. The biggest advantage to the answer extraction style is its ease in generating questions, which enables us to produce large-scale datasets. In contrast, a disadvantage to this style is that it rarely demands *meta/whole* and *math/logic* skills, which can require answers not contained in the context. Moreover, as observed in Section 4.4, it seems difficult to guarantee that all possible answer spans are given as the correct answers. By contrast, the description and multiple-choice styles have the advantage of having no such restrictions on the appearance of candidate answers (Kočiský et al., 2018; Khashabi et al., 2018a). Nonetheless, the description style is difficult to evaluate because the Rouge-L and BLEU scores are insufficient for testing NLU. Whereas it is easy to evaluate the performance on multiple-choice questions, generating multiple reasonable options requires considerable effort.

Interpretation of our heuristics. When we regard the MRC task as recognizing textual entailment (RTE) (Dagan et al., 2006), the task requires the reader to construct one or more premises from the context and form the most reasonable hypothesis from the question and candidate answer (Sachan et al., 2015). Thus, easier questions are those (i) where the reader needs to generate only one hypothesis, and (ii) where the premises directly describe the correct hypothesis. Our two heuristics can also be seen as the formalizations of these criteria. Therefore, to make questions more realistic, we need to create multiple hypotheses that require complex reasoning to be distinguished. Moreover, the integration of premises should be complemented by external knowledge to provide sufficient information to verify the correct hypothesis.

4.6 Related Work

Our heuristics and annotation were motivated by *unintended biases* (Levesque, 2014) and *evaluation overfitting* (Whiteson et al., 2011), respectively.

Unintended biases. The MRC task tests a reading process that involves retrieving stored information and performing inferences (Sutcliffe et al., 2013). However, constructing datasets that comprehensively require those skills is difficult. As Levesque (2014) discussed as a desideratum for testing AI, we should avoid creating questions that can be solved by matching patterns, using unintended biases, and selectional restrictions. For the unintended biases, one suggestive example is the Story Cloze Test (Mostafazadeh et al., 2016), in which a system chooses a sentence among candidates to conclude a given paragraph of the story. A recent attempt at this task showed that recognizing superficial features in the correct candidate is critical to achieve the state of the art (Schwartz et al., 2017).

Similarly, in MRC, Weissenborn et al. (2017) proposed *context/type matching heuristic* to develop a simple neural system. Min et al. (2018) observed that, in SQuAD, 92% of answerable questions can be answered only using a single context sentence. In visual question answering, Agrawal et al. (2016) analyzed the behavior of models with the variable length of the first question words. Khashabi et al. (2018a) more recently proposed a dataset with questions for multi-sentence reasoning.

Evaluation overfitting. The theory behind evaluating AI distinguishes between task- and skill-oriented approaches (Hernández-Orallo, 2017a). In the task-oriented approach, we usually develop a system and test it on a specific dataset. The developed system sometimes lacks generality but achieves the state of the art for that specific dataset. Further, it becomes difficult to verify and explain the solution to tasks. The situation in which we are biased to the specific tasks is called evaluation overfitting (Whiteson et al., 2011). By contrast, with the skill-oriented approach, we aim to interpret the relationships between tasks and skills. This orientation can encourage the development of more realistic NLU systems.

As one of our goals was to investigate whether easy questions are dominant in recent datasets, it did not necessarily require a detailed classification of reasoning types. Nonetheless, we recognize there are more fine-grained classifications of the required skills for NLU. For example, Weston et al. (2015) defined 20 skills as a set of toy tasks. LoBue and Yates (2011) and Sammons et al. (2010) analyzed entailment phenomena using detailed classifications in RTE. For the ARC dataset, Boratko et al. (2018) proposed knowledge and reasoning types.

4.7 Conclusion

This study examined MRC questions from 12 datasets to determine what makes such questions easier to answer. We defined two heuristics that limit candidate answers and thereby mitigate the difficulty of questions. Using these heuristics, the datasets were split into easy and hard subsets. We further annotated the questions with their validity and the reasoning skills needed to answer them. Our experiments revealed that the baseline performance degraded with the hard questions, which required knowledge inference and multiple-sentence reasoning compared to easy questions. These results suggested that one might overestimate the ability of the baseline systems. They also emphasized the importance of analyzing and reporting the properties of new datasets when released. One limitation of this chapter was the heavy cost of the annotation. It is necessary to explore a method for automatically classifying reasoning types. This will enable us to evaluate systems through a detailed organization of the datasets.

Chapter 5

Assessment of the Benchmarking Capacity

Existing analysis work in machine reading comprehension (MRC) is largely concerned with evaluating the capabilities of systems. However, the capabilities of datasets are not assessed for benchmarking language understanding precisely. We propose a semi-automated, ablation-based methodology for this challenge; By checking whether questions can be solved even after removing features associated with a skill requisite for language understanding, we evaluate to what degree the questions do *not* require the skill. Experiments on 10 datasets (e.g., CoQA, SQuAD v2.0, and RACE) with a strong baseline model show that, for example, the relative scores of the baseline model provided with *content words only* and with *shuffled sentence words* in the context are on average 89.2% and 78.5% of the original scores, respectively. These results suggest that most of the questions already answered correctly by the model do not necessarily require grammatical and complex reasoning. For precise benchmarking, MRC datasets will need to take extra care in their design to ensure that questions can correctly evaluate the intended skills.

5.1 Introduction

Machine reading comprehension (MRC) is a testbed for evaluating natural language understanding (NLU), by letting machines answer questions about given texts (Hirschman et al., 1999). Although MRC could be the most suitable task for evaluating NLU (Chen, 2018) and the performance of systems is comparable to humans on some existing datasets Devlin et al. (2019), it has been found that the quality of existing datasets might be insufficient for requiring precise understanding (Jia and Liang, 2017). Whereas these analyses are useful to investigate the performance of *systems*, however, it is still necessary to verify the fine-grained capabilities of *datasets* for benchmarking NLU.

In the design of MRC datasets, it is implicitly assumed that questions test a cognitive process of language understanding (Sutcliffe et al., 2013). As various aspects of such a process, we can use *requisite skills* for answering questions such as coreference resolution and commonsense reasoning (Chapter 3). Considering skills as metrics would be useful for analyzing datasets. However, for most datasets, the skills required to answer existing questions are not identified, or significant human annotation is needed.

In this chapter, we propose a semi-automated, ablation-based methodology to analyze the capabilities of MRC datasets to benchmark NLU. Our motivation is to investigate to what extent a dataset allows unintended solutions that do not need requisite

Original context

Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared ***to*** Saint Bernadette Soubirous ***in 1858***. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.

Anonymized context

@adv1 @prep5 @other0 @noun17 @verb2 @other0 @noun20 @punct0 @other1 @adj3 @noun21 @prep1 @noun22 @other2 @noun23 @period0 @other3 @verb2 @other1 @noun24 @prep1 @other0 @noun20 @prep6 @noun25 @punct0 @noun26 @wh0 @other0 @noun7 @noun8 @adv3 @verb4 ***@prep4*** @noun27 @noun28 @noun29 ***@prep2*** ***@num0*** @period0 @prep6 @other0 @noun30 @prep1 @other0 @adj4 @noun31 @punct3 @other2 @prep2 @other1 @adj5 @noun32 @wh1 @verb5 @prep7 @num1 @noun6 @other2 @other0 @noun4 @noun5 @punct4 @punct0 @verb2 @other1 @adj6 @punct0 @adj7 @noun33 @noun6 @prep1 @noun8 @period0

Question

To whom did the Virgin Mary allegedly appear ***in 1858*** in Lourdes France?

Anonymized question

@prep4 @wh2 @verb6 @other0 @noun7 @noun8 @adv4 @verb4 ***@prep2*** ***@num0*** @prep2 @noun25 @noun26 @period1

Baseline model's prediction before / after anonymization

Saint Bernadette Soubirous / noun27 @noun28 @noun29

Figure 5.1: Example of an ablation test that anonymizes context and question words, applied to a question from SQuAD v1.1 (Rajpurkar et al., 2016) with the correct answer in underscored. We found that the baseline model can achieve 61.2% F1 on SQuAD v1.1 even after the anonymization.

skills. This leads to the following intuition: if a question is correctly answered (or *solvable*) even after removing features associated with a given skill, the question does not require the skill. We show an example of our ablation method in Figure 5.1. Suppose we wish to analyze a dataset's capacity to evaluate understanding of texts beyond the information of part-of-speech (POS) tags. To this end, we replace context and question words with POS tags and ID numbers. If a model can still correctly answer this modified question, the question does not necessarily require deep understanding of texts but matching word patterns only. Questions of this kind might be insufficient for developing a model that understands texts deeply as they may reduce models to recognizing superficial word overlaps.

Our methodology uses a set of requisite skills and corresponding ablation methods. Inspired by the computational model of reading comprehension (Kintsch, 1988), we exemplify 12 skills on two classes: reading and reasoning (Section 5.3). Then, we present a large-scale analysis over 10 existing datasets using a strong baseline model (Section 5.4). In Section 5.5, we perform a complementary inspection of questions with our ablation methods in terms of the solvability of questions and the reconstructability of ablated features. Finally we discuss, in Section 5.6, two requirements for developing MRC to benchmark NLU: the control of question solvability and the comprehensiveness of requisite skills.

Our contributions are as follows:

- We propose a semi-automated methodology to analyze the benchmarking capacity of MRC datasets in terms of requisite skills for answering questions.
- With an example set of 12 skills and corresponding input-ablation methods, we use our methodology and examine 10 existing datasets with two answering styles.
- Our analysis shows that the relative performance on questions with *content words only*, *shuffled sentence words*, and *shuffled sentence order* averaged 89.2%, 78.5%, and 95.4% of the original performance, indicating that the questions might be inadequate for evaluating grammatical and complex reasoning.

These results suggest that most of the questions currently *solved* in MRC may be insufficient for evaluating various skills. A limitation of our method is that it can not draw conclusions regarding questions that remain *unsolved*, and thus we need to assume a reasonable level of performance for existing models on the dataset to be analysed. Given our findings, we posit that MRC datasets should be carefully designed, e.g., by filtering questions using methods such as the ones we propose, so that their questions correctly benchmark the intended NLU skills.

5.2 Related Work

We briefly survey existing interpretation methods and skill-based analyses for NLU tasks.

Interpretation methods. A challenge with the MRC task is that we do not know the extent to which a successful model precisely understands natural language. To analyze a model's behavior, existing studies mainly proposed modification of the input. For example, Jia and Liang (2017) showed that the performance of existing models on SQuAD (Rajpurkar et al., 2016) significantly degrades when manually verified distracting sentences are added to the given context. In addition, Feng et al. (2018) demonstrated that MRC models do not necessarily change their predictions even when most question tokens are dropped. Kaushik and Lipton (2018) also proposed dropping the whole context or question to ascertain whether models' predictions depend only on either the context or the question. Likewise, for the natural language inference task, Gururangan et al. (2018) proposed to hide the premise and to evaluate a model using only the hypothesis. One of similar attempts is performed by Naik et al. (2018), in which they proposed automatically constructed stress tests to examine the ability of natural language inference models. These kinds of analyses are helpful for detecting biases that are unintentionally included in datasets. Nonetheless, to assure that a dataset can evaluate various aspects of NLU, more fine-grained detail is needed than what is allowed by inspection using existing methods.

Skills as units of interpretation. In the topic of interpretable machine learning, Doshi-Velez and Kim (2018) defined the concept of *cognitive chunks* as the basic units of explanation. In the MRC task, we consider that *requisite skills* to answer questions are appropriate as such units. A skill-based analysis was conducted by Boratko et al. (2018), who proposed classifications of knowledge and reasoning. Prior to this, we also defined a set of 13 requisite skills in Chapter 3. However, there are two main

Comprehension skill s_i	Ablation method σ_i
Reading-class	
1. Recognizing question words excluding interrogatives	Drop all words except interrogatives (*wh*-words and *how*) in a question.
2. Recognizing content words	Drop content words in the context.
3. Recognizing function words	Drop function words in the context.
4. Recognizing vocabulary	Anonymize context and question words with their part-of-speech tag.
5. Attending the whole context other than similar sentences	Keep the sentences that are the most similar to the question in terms of unigram overlap and drop the other sentences.
6. Recognizing the word order	Randomly shuffle all words in the context.
Reasoning-class	
7. Grasping sentence-level compositionality	Randomly shuffle the words in all the sentences except the last token.
8. Understanding of discourse relations	Randomly shuffle the order of the sentences in the context.
9. Performing basic arithmetic operations	Replace numerical expressions (CD tag) with random numbers.
10. Explicit logical reasoning	Drop logical terms such as *not*, *every*, and *if*.
11. Resolving pronoun coreferences	Drop personal and possessive pronouns (PRP and PRP$ tags).
12. Reasoning about explicit causality	Drop causal terms/clauses such as *because* and *therefore*.

Table 5.1: Example set of requisite skills $\{s_i\}$ and corresponding ablation methods $\{\sigma_i\}$. f is a model and (x, y) is a pair consisting of an input instance and its gold-standard answer. We interpret that for x s.t. $f(x) = y$, if $f(\sigma_i(x)) = y$, then x is solvable without s_i.

issues with these approaches: (i) the human annotation does not necessarily reveal unintended biases that machines can make use of, and (ii) it requires costly annotation efforts. Therefore, we posit that a machine-based analysis is needed and that it should be performed in an automated manner.

5.3 Dataset Diagnosis by Input Ablation

5.3.1 Formulation

Our methodology uses a set of requisite skills and corresponding ablation methods. By checking the solvability of questions after applying the ablation methods, we can quantify to what degree the questions allow unintended solutions that do not require the requisite skills. Users can define an arbitrary set of skills to suit their purposes.

We develop a method σ_i that ablates features necessary for the corresponding skill s_i in a set of requisite skills S. For $(x, y) \in X \times Y$, whenever $f(x) = y$, if $f(\sigma_i(x)) = y$, we recognize that x is solvable without s_i. Here, X is the input, Y is the gold labels, (x, y) is a pair consisting of an input instance and its gold-standard answer, and f is a model. When the performance gap between the original and the modified dataset is small, we can infer that most of the questions already solved are solvable without s_i. On the other hand, if the gap is large, a sizable proportion of the solved questions may require s_i.

We note that we cannot draw general conclusions for instances given by conditions other than the abovementioned one. Consider the case where $f(x) = y$ and $f(\sigma_i(x)) \neq y$, for example. This only means that f cannot solve x without the features ablated by σ_i. We cannot conclude that x requires s_i in *every* model because there might exist a model that can solve x without s_i. However, if there is *at least one* model f that solves x without s_i, this may indicate an unintended way to solve x while ignoring s_i. Therefore our methodology only requires a single baseline model. Users can choose an arbitrary model for their purposes.

5.3.2 Example Set of Requisite Skills

In this section, we exemplify a skill set that consists of 12 skills along with two classes; reading and reasoning (Table 5.1). In psychology, there is a tradition of theoretical research on human text comprehension. The construction–integration model (Kintsch, 1988) is one of the most acknowledged theories. This model assumes that human text comprehension consists of two processes: (i) construction, in which a reader elaborates concepts and propositions in the text and (ii) integration, in which the reader associates the propositions to understand them consistently. We associate this two-step process with our two classes.

Reading skills. This class deals with six skills of observing and recognizing word appearances, which are performed before reasoning. In MRC, it has been shown that some existing questions can be solved by reading a limited number of words in the question and the context (e.g., by simply attending to context tokens that are similar to those of the questions in Chapter 4). Our goal of this class is, therefore, to ensure that the questions require the reading of the whole question and context uniformly.

Reasoning skills. This class comprises six skills of relational reasoning among described entities and events such as pronoun coreference resolution and logical reasoning. Although these skills are essential for sophisticated NLU, it is difficult to precisely determine whether these types of reasoning are genuinely required in answering a question. Therefore, in this class, we define reasoning-related skills that are performed using the *explicit* information contained in the context (e.g., s_9 explicit logical reasoning and s_{12} reasoning about explicit causality).

In the following, we highlight some of the defined skills. Skill s_1 is inspired by Feng et al. (2018) and Chapter 4. Although their studies proposed dropping question tokens based on their model-based importance or the question length, we instead drop tokens other than interrogatives as interpretable features. As s_2 and s_3, we propose limiting the information available in the context by dropping content and function words respectively, which is intended to ascertain the extent to which a question depends on the given word type (e.g., a preposition *in* before a time-related expression for a *when* question). Our vocabulary anonymization (s_4) is mainly inspired by Hermann et al. (2015) where they anonymized named entities to make their MRC task independent of prior knowledge. Skill s_5 provides a heuristic of the relative levels of *attention* between a question and the context. Skill s_6 is used to ensure that a model can extract the information conditioned on the word order. Our shuffle-based methods (s_6 to s_8) are inspired by existing analyses for other tasks (Khandelwal et al., 2018; Nie et al., 2019; Sankar et al., 2019). Among them, our purpose for s_7 is to analyze whether a question requires *precise* reasoning performed over syntactic and grammatical aspects in each sentence. Skill s_8 is for the understanding of discourse relations between adjacent sentences, which relies on information given by the sentence order in the context. When we shuffle the sentence order, various relations, such as causality and temporality, are expected to be broken. Skills s_9 to s_{12} are defined more specifically; we drop tokens that explicitly emphasize important roles in specific skills such as *if* and *not* in logical reasoning.

Although our proposed definitions can be extended, they are sufficient for the purpose of demonstrating and evaluating our approach. In Section 5.6, we discuss further directions to develop purpose-oriented skill sets.

5.4 Experiments and Further Analyses

5.4.1 Experimental Settings

Datasets. We use 10 datasets. For answer extraction datasets in which a reader chooses a text span in a given context, we use (1) CoQA (Reddy et al., 2019), (2) DuoRC (Saha et al., 2018), (3) HotpotQA (distractor) (Yang et al., 2018), (4) SQuAD v1.1 (Rajpurkar et al., 2016), and (5) SQuAD v2.0 (Rajpurkar et al., 2018). For multiple choice datasets in which a reader chooses a correct option from multiple options, we use (6) ARC (Challenge) (Clark et al., 2018), (7) MCTest (Richardson et al., 2013), (8) MultiRC (Khashabi et al., 2018a), (9) RACE (Lai et al., 2017), and (10) SWAG (Zellers et al., 2018). For the main analysis, we applied our ablation methods to development sets. We included SWAG because its formulation can be viewed as a multiple-choice MRC task and we would like to analyze the reasons for the high performance reported for the baseline model on this dataset (Devlin et al., 2019). For preprocessing the datasets, we use CoreNLP (Manning et al., 2014).

For CoQA, since this dataset allows for *yes/no/unknown* questions, we appended these words to the end of the context. These special words were not allowed to be dropped. Additionally, we appended the previous question-answer pair prior to the current question so that the model can consider the history of the QA conversation. To compute the performance on SQuAD v2.0, we used the best F1 value that was derived from the predictions with a no-answer threshold of 0.0. For DuoRC, we used the ParaRC dataset (the official preprocessed version provided by the authors). When training a model on DuoRC and HotpotQA, we used the first answer span; i.e., the document spans that have no answer span were not used in training. For MCTest and RACE, we computed accuracy by combining MC160 with MC500 and Middle with High, respectively. For MultiRC, which is allowed to have multiple correct options for a question, we cast a pair consisting of a question and one option as a two-option multiple choice (i.e., whether its option is true or false) and computed the micro-averaged accuracy for the evaluation. The SWAG dataset is a multiple-choice task of predicting which event is most likely to occur next to a given sentence and the subject (noun phrase) of a subsequent event. We cast the first sentence as the context and the subject of the second sentence as the question. To compute F1 scores for the answer

Dataset	d	b	lr	ep
CoQA	512	24	3×10^{-5}	2
DuoRC	512	24	3×10^{-5}	2
HotpotQA	512	24	3×10^{-5}	2
SQuAD v1.1	384	24	3×10^{-5}	2
SQuAD v2.0	384	24	3×10^{-5}	2
ARC	384	24	1×10^{-5}	4
MCTest	512	16	2×10^{-6}	4
MultiRC	512	24	2×10^{-5}	4
RACE	512	32	1×10^{-5}	4
SWAG	128	32	1×10^{-5}	4

Table 5.2: Hyperparameters used in the experiments, where d is the size of the token sequence fed into the model, b is the training batch size, lr is the learning rate, and ep is the number of training epochs. We set the learning rate warmup in RACE to 0.05 and 0.1 for the other datasets. We used stride = 128 for documents longer than d tokens.

extraction datasets, we used the official evaluation scripts provided for the answer extraction datasets.

Models. As the baseline model, we used BERT-large (Devlin et al., 2019).[1] We fine-tuned it on the original training set of each dataset and evaluated it on a modified development set. For σ_4 vocabulary anonymization, we train the model after the anonymization. For ARC, MCTest, and MultiRC, we fine-tuned a model that had already been trained on RACE to see the performance gained by transfer learning (Sun et al., 2019b). We report the hyperparameters of our models in Table 5.2. Although we trained the baseline model on the original training set, it is assumed that the upper-bound performance can be achieved by a model trained on the modified training set. Therefore, in Section 5.4.3, we also see the extent to which the performance improves when the model is trained on the modified training set.

Ablation methods. σ_2 and σ_3: we use a set of stopwords from NLTK (Loper and Bird, 2002) as function words. All other words are regarded as content words. We do not drop punctuation. When a token is dropped, it is replaced with an [UNK] token to preserve the correct answer span. σ_4: we use the same ID for the same word in a single given context but different IDs for different contexts. For inflectional words, we anonymize them using their lemma. For example, *are* would be replaced with *@verb2* (= *is*) if it appeared in Figure 5.1. In addition, to retain the information of the POS tags, we append its POS tag after each inflectional anonymized word (e.g., *is* is replaced with *@verb{ID} [VBZ]*). We used the tags as shown in Table 5.3 and *@other* tags for the other POS tags. σ_6: because it is necessary to maintain the correct answer span in the answer extraction datasets, we split the context into segments that have the same length as the gold answer span and shuffle them. σ_7: as with σ_6, we split each sentence into segments and shuffle them within each sentence. For σ_6 to

[1] Although our methodology only necessitates a single baseline model, note that we need to assume a reasonable level of performance as we mentioned in Section 5.1.

Anonymization tag	POS tag or tokens
@noun{ID}	NN, NNS, NNP, NNPS
@verb{ID}	VB, VBD, VBG, VBN, VBP, VBZ
@adj{ID}	JJ, JJR, JJS
@adv{ID}	RB, RBR, RBS
@number{ID}	CD
@wh{ID}	WDT, WP, WP$, WRB
@prep{ID}	IN, TO
@punct{ID}	(punctuation except for the period tokens below)
@period{ID}	. ! ?

Table 5.3: Examples of anonymization tags and corresponding POS tags (OntoNotes 5 version of Penn Treebank tag set). We use *@noun*, *@verb*, *@adj*, *@adv*, and *@number* for content words.

σ_8, we averaged the scores over five runs with different seeds. For σ_{10} logical words dropped, as logic-related terms, we used the following: *all, any, each, every, few, if, more, most, no, nor, not, other, same, some,* and *than*. For σ_{12} causal words dropped, as causality-related terms, we used the following: *as, because, cause, since, therefore,* and *why*. For σ'_3 training with content words only, we dropped function words as well as punctuation marks so that the model would see only content words.

5.4.2 Results of Reading and Reasoning Skills

We report the results for the skills in Tables 5.4 and 5.5. In Table 5.6, we report dataset statistics and the average number of tokens dropped in each drop-based method. In the following, % indicates a relative change from the original F1/accuracy unless specified otherwise. In this section, we describe the notable findings for several skills.

s_1: recognizing question words. For the first four answer-extraction datasets, the performance decreased by more than 70%. For the multiple-choice datasets, the performance decreased by an average of 23.9%.

s_2 and s_3: recognizing content words and function words. On all datasets, the relative changes for s_2 were greater than those for s_3. However, it is remarkable that even with function words alone, the model could achieve 53.0% and 17.4% F1 on CoQA and SQuAD v1.1, respectively.[2] On ARC, RACE, and SWAG, the model showed more than 40% accuracy (>25% of random choice). As for content words only, on all answer extraction datasets, the performance was greater than 78.7% that of the original. On all multiple-choice datasets, it was more than 90.2%. These results imply that most of the questions already solved do not necessarily require grammatical and syntactic reasoning, in which function words are used. We show examples of questions for this ablation method in Figure 5.2.

[2] 19.8% of the questions in CoQA are yes/no questions.

Ablation method \ Dataset	CoQA	DuoRC	Hotpot-QA	SQuAD v1.1	SQuAD v2.0
Answering style	answer extraction (F1)				
Original dataset	77.4 $_{0.0}$	58.4 $_{0.0}$	63.6 $_{0.0}$	91.5 $_{0.0}$	81.9 $_{0.0}$
1. Q interrogatives only	20.1$_{-74.0}$	14.2$_{-75.8}$	15.0$_{-76.4}$	15.2$_{-83.4}$	50.1$_{-38.9}$
2. Function words only	53.0$_{-31.5}$	5.8$_{-90.1}$	7.8$_{-87.8}$	17.4$_{-81.0}$	50.2$_{-38.7}$
3. Content words only	60.9$_{-21.3}$	47.9$_{-18.0}$	56.2$_{-11.6}$	80.7$_{-11.8}$	73.5$_{-10.3}$
4. Vocab. anonymization	39.0$_{-49.6}$	18.6$_{-68.2}$	16.8$_{-73.6}$	61.2$_{-33.1}$	59.4$_{-27.0}$
5. Most sim. sent. only	32.6$_{-57.9}$	35.8$_{-38.7}$	16.9$_{-73.4}$	68.5$_{-25.1}$	72.8$_{-11.2}$
6. Context words shuff.	29.8$_{-61.5}$	25.4$_{-56.6}$	23.6$_{-62.9}$	35.9$_{-60.7}$	52.4$_{-36.1}$
7. Sentence words shuff.	53.0$_{-31.6}$	35.9$_{-38.6}$	43.1$_{-32.2}$	62.1$_{-32.1}$	64.4$_{-21.4}$
8. Sentence order shuff.	72.2 $_{-6.8}$	56.1 $_{-4.0}$	53.7$_{-15.6}$	90.3 $_{-1.3}$	80.7 $_{-1.5}$
9. Dummy numerics	75.9 $_{-1.9}$	57.8 $_{-1.0}$	60.0 $_{-5.6}$	89.5 $_{-2.2}$	78.7 $_{-3.9}$
10. Logical words dropped	76.7 $_{-0.9}$	58.0 $_{-0.7}$	62.1 $_{-2.3}$	91.0 $_{-0.5}$	80.6 $_{-1.6}$
11. Pronoun words dropped	76.5 $_{-1.2}$	57.0 $_{-2.5}$	63.4 $_{-0.3}$	91.2 $_{-0.2}$	81.8 $_{-0.2}$
12. Causal words dropped	77.3 $_{-0.1}$	58.3 $_{-0.3}$	63.3 $_{-0.5}$	91.2 $_{-0.3}$	81.8 $_{-0.2}$

Table 5.4: The performances (%) of the baseline model with the ablation tests on the development set of the answer extraction datasets. Values in smaller font are changes (%) relative to the original baseline performance.

s_4: recognizing vocabulary beyond POS tags. Surprisingly, for SQuAD v1.1, the baseline model achieved 61.2% F1. It only uses 248 tokens as the vocabulary with the anonymization tags and no other actual tokens. For the other answer extraction datasets, the largest drop (73.6% relative) is by HotpotQA; it has longer context documents than the other datasets, which seemingly makes its questions more difficult. To verify the effect of its longer documents, we also evaluated the baseline model on HotpotQA without distracting paragraphs. We found that the model's performance was 56.4% F1 (the original performance was 76.3% F1 and its relative drop was 26.1%) which is much higher than that on the context with distracting paragraphs (16.8% F1). This indicates that adding longer distracting documents contributes to encouraging machines to understand a given context beyond matching word patterns.

On the other hand, the performance on the multiple choice datasets was significantly worse; if multiple choices do not have sufficient word overlap with the given context, there is no way to infer the correct answer option. Therefore, this result shows that multiple choice datasets might have a capacity for requiring more complex understanding beyond matching patterns between the question and the context than the answer extraction datasets.

s_5: attending to the whole context other than similar sentences. Even with only the most similar sentences, the baseline models achieved a performance level greater than half their original performances in 8 out of 10 datasets. In contrast, HotpotQA showed the largest decrease in performance. This result reflects the fact that this dataset contains questions requiring multi-hop reasoning across multiple sentences.

Ablation method \ Dataset	ARC	MCTest	Multi-RC	RACE	SWAG	Rel. avg.
Answering style			multiple choice (accuracy)			
Original dataset	52.7 $_{0.0}$	87.8 $_{0.0}$	78.0 $_{0.0}$	68.8 $_{0.0}$	85.4 $_{0.0}$	0.0
1. Q interrogatives only	35.6$_{-32.5}$	64.1$_{-27.0}$	52.6$_{-32.6}$	56.7$_{-17.5}$	77.1 $_{-9.7}$	-46.8
2. Function words only	44.0$_{-16.6}$	32.2$_{-63.3}$	61.9$_{-20.6}$	43.2$_{-37.3}$	68.9$_{-19.4}$	-48.6
3. Content words only	48.0 $_{-8.9}$	80.3 $_{-8.5}$	74.5 $_{-4.5}$	62.0 $_{-9.8}$	82.6 $_{-3.3}$	-10.8
4. Vocab. anonymization	29.2$_{-44.6}$	25.3$_{-71.2}$	57.2$_{-26.7}$	26.1$_{-62.1}$	25.5$_{-70.1}$	-52.6
5. Most sim. sent. only	43.6$_{-17.2}$	50.3$_{-42.7}$	67.9$_{-12.9}$	52.1$_{-24.3}$	85.4 $_{-0.1}$	-30.4
6. Context words shuff.	47.4 $_{-9.9}$	47.2$_{-46.3}$	64.3$_{-17.6}$	51.7$_{-24.9}$	78.6 $_{-8.0}$	-38.4
7. Sentence words shuff.	46.4$_{-11.8}$	70.6$_{-19.6}$	71.4 $_{-8.5}$	59.7$_{-13.3}$	80.3 $_{-6.0}$	-21.5
8. Sentence order shuff.	50.3 $_{-4.5}$	82.5 $_{-6.0}$	75.6 $_{-3.0}$	66.8 $_{-2.9}$	85.4 $_{-0.0}$	-4.6
9. Dummy numerics	49.7 $_{-5.7}$	85.0 $_{-3.2}$	76.2 $_{-2.3}$	67.8 $_{-1.5}$	85.3 $_{-0.1}$	-2.8
10. Logical words dropped	52.0 $_{-1.3}$	85.3 $_{-2.8}$	77.3 $_{-1.0}$	67.7 $_{-1.5}$	85.4 $_{0.0}$	-1.3
11. Pronoun words dropped	52.0 $_{-1.3}$	86.6 $_{-1.4}$	77.4 $_{-0.8}$	68.3 $_{-0.7}$	84.8 $_{-0.8}$	-0.9
12. Causal words dropped	52.0 $_{-1.3}$	87.5 $_{-0.4}$	77.6 $_{-0.6}$	68.2 $_{-0.8}$	85.5 $_{0.0}$	-0.4

Table 5.5: The performances (%) of the baseline model with the ablation tests on the development set of the multiple choice datasets. Values in smaller font are changes (%) relative to the original baseline performance, and the rightmost column ("Rel. avg.") shows the averages over all datasets.

s_6: recognizing the context word order (context words shuffle). We found that for the answer extraction datasets, the relative performance decreased by 55.6% on average. A moderate number of questions are solvable even with the context words shuffled. We also found that, surprisingly, the average decrease was 21.3% for the multiple choice datasets. The drop on MCTest is more prominent than that on the others. We posit that this is because its limited vocabulary makes questions more context dependent. ARC, in contrast, uses factoid texts, and appears less context dependent. We report the variance for shuffling methods $s_{6,7,8}$ in Table 5.7.

s_7: grasping sentence-level compositionality (sentence words shuffle). The performance with sentence words shuffled was greater than 60% and 80% those of the original dataset on the answer extraction and multiple-choice datasets, respectively. This result means that most of the solved questions are solvable even with the sentence words shuffled. However, we should not say that all questions must require this skill; a question can require the performance of some complex reasoning (e.g., logical and multi-hop reasoning) and merely need to identify the sentence that gives the correct answer without precisely understanding that sentence. Nevertheless, if the question is not intended to require such reasoning, we should care whether it can be solved with only a (sentence-level) bag of words. In order to ensure that a model can understand the precise meaning of a described event, we may need to include questions to evaluate the grammatical and syntactic understanding into a dataset.

Dataset	CoQA	DuoRC	HotpotQA	SQuAD1.1	SQuAD2.0
Text genre	various	movie	Wikipedia		
Avg. # Q tokens	6.6	8.7	18.0	11.7	11.4
Avg. # C tokens	344.0	691.3	1206.5	147.6	151.6
Avg. # sentences in C	18.8	25.3	47.8	5.7	6.1
Avg. # dropped tokens					
1. Q interrogatives only	5.8 100.0	7.7 100.0	16.8 100.0	10.7 100.0	10.4 100.0
2. Function words only	151.0 100.0	357.1 100.0	606.4 100.0	76.6 100.0	78.5 100.0
3. Content words only	131.6 100.0	305.6 100.0	366.3 100.0	50.8 100.0	52.7 100.0
5. Most sim. sent. only	300.0 99.8	623.6 97.8	1139.0 100.0	105.0 97.8	109.8 98.5
9. Dummy numerics	6.3 93.2	5.4 85.9	60.7 100.0	5.7 86.3	5.3 83.4
10. Logical words drop.	6.7 100.0	8.0 91.1	8.6 100.0	2.1 75.7	2.5 78.8
11. Pronoun words drop.	19.7 98.6	49.4 99.4	22.3 99.9	2.3 72.6	1.9 70.8
12. Causal words drop.	2.4 84.4	5.5 88.4	9.8 99.0	1.4 66.5	1.5 69.5

Dataset	ARC	MCTest	MultiRC	RACE	SWAG
Text genre	science	story	story	various	video
Avg. # Q tokens	25.5	9.2	17.6	11.1	3.0
Avg. # C tokens	131.4	247.7	339.9	326.8	13.3
Avg. # sentences in C	8.6	20.1	15.9	19.8	1.0
Avg. # dropped tokens					
1. Q interrogatives only	24.4 100.0	8.1 100.0	16.5 100.0	10.6 100.0	2.9 100.0
2. Function words only	67.8 100.0	106.5 100.0	168.7 100.0	146.5 100.0	6.4 100.0
3. Content words only	46.5 100.0	106.7 100.0	113.6 100.0	132.6 100.0	5.4 99.8
5. Most sim. sent. only	89.3 98.3	217.6 99.7	299.4 100.0	288.0 99.8	0.0 0.1
9. Dummy numerics	2.2 53.0	1.5 67.5	20.1 100.0	6.2 90.0	0.1 8.0
10. Logical words drop.	2.8 73.2	4.6 97.5	4.7 95.7	7.6 97.2	0.1 7.3
11. Pronoun words drop.	1.8 65.8	22.0 100.0	13.5 99.2	19.4 99.1	0.8 54.1
12. Causal words drop.	1.3 56.7	1.2 51.2	2.2 87.3	2.3 78.2	0.1 9.2

Table 5.6: Statistics of the datasets examined and average numbers of tokens dropped by our ablation methods σ_i ($i = 1, 2, 3, 5, 9, ..., 12$). The tokens are counted after tokenization of the punctuation. Values in smaller font denote the proportion (%) of questions that contain dropped tokens.

s_8: discourse relation understanding (sentence order shuffle). The smallest drop, excluding SWAG, which has one context sentence, was -1.3%, on SQuAD v1.1.[3] Except for HotpotQA, the datasets show small drops (less than 10%), which indicates that most solved questions do not require understanding of adjacent discourse relations and are solvable even if the sentences appear in an unnatural order.

s_9–s_{12}: various types of reasoning. For these skills, we can see that the performance drops were small; given that the drop for s_3 recognizing content words alone was under 20%, we can infer that specific types of reasoning might not be critical for answering the questions. Some types of reasoning, however, might play an essential role for some datasets: s_9 numerical reasoning in HotpotQA (whose questions

[3] Min et al. (2018) also reported that more than 90% of questions on SQuAD v1.1 necessitate only a single sentence to answer them.

Ablation method	CoQA	DuoRC	HotpotQA	SQuAD1.1	SQuAD2.0
6. Context w. shuff.	29.8 (0.3)	25.4 (0.4)	23.6 (0.3)	35.9 (0.3)	52.4 (0.2)
7. Sent. w. shuff.	53.0 (0.2)	35.9 (0.3)	43.1 (0.3)	62.1 (0.3)	64.4 (0.3)
8. Sent. ord. shuff.	72.2 (0.2)	56.1 (0.4)	53.7 (0.3)	90.3 (0.1)	80.7 (0.1)

Ablation method	ARC	MCTest	MultiRC	RACE	SWAG
6. Context w. shuff.	47.4 (1.9)	47.2 (1.3)	64.3 (0.2)	51.7 (0.4)	78.6 (0.2)
7. Sent. w. shuff.	46.4 (2.0)	70.6 (1.6)	71.4 (0.3)	59.7 (0.1)	80.3 (0.1)
8. Sent. ord. shuff.	50.3 (0.9)	82.5 (1.4)	75.6 (0.4)	66.8 (0.3)	85.4 (0.0)

Table 5.7: Ablation results with variances in parentheses for shuffle-related skills (s_6, s_7, and s_8) for five different runs.

Ablation method \Subset	#Has-ans 5928	#No-ans 5945	#Total 11873
Original dataset	$82.6_{0.0}$	$79.9_{0.0}$	$81.9_{0.0}$
1. Interrogatives in Q	$8.6_{-89.6}$	$47.3_{-40.8}$	$50.1_{-38.9}$
2. Function words only	$0.4_{-99.5}$	$99.6_{24.7}$	$50.1_{-38.8}$
3. Content words only	$65.6_{-20.5}$	$81.2_{1.6}$	$73.5_{-10.3}$
4. Vocab. anonymization	$41.9_{-49.3}$	$76.9_{-3.8}$	$59.4_{-27.5}$
5. Most sim. sent. only	$69.2_{-16.2}$	$83.2_{4.1}$	$72.8_{-11.1}$
6. Context words shuff.	$9.1_{-89.0}$	$95.5_{19.5}$	$52.4_{-36.1}$
7. Sentence words shuff.	$38.8_{-53.0}$	$90.2_{12.9}$	$64.6_{-21.2}$
8. Sentence order shuff.	$78.4_{-5.1}$	$81.9_{2.5}$	$80.3_{-2.0}$
9. Dummy numerics	$74.7_{-9.6}$	$82.0_{2.6}$	$78.7_{-3.9}$
10. Logical words dropped	$80.4_{-2.6}$	$80.0_{0.1}$	$80.6_{-1.6}$
11. Dummy pronoun res.	$82.0_{-0.7}$	$80.6_{0.8}$	$81.8_{-0.2}$
12. Causal words dropped	$82.1_{-0.5}$	$79.9_{0.0}$	$81.8_{-0.2}$
All Q words dropped	$10.8_{-86.9}$	$17.7_{-77.9}$	$50.1_{-38.9}$
Trained & evaluated on			
4′. Content words only	$75.6_{-8.5}$	$72.9_{-8.8}$	$74.8_{-8.7}$
6′. Context words shuff.	$63.6_{-22.9}$	$97.1_{21.5}$	$80.6_{-1.7}$
7′. Sentence words shuff.	$75.5_{-8.5}$	$79.3_{-0.8}$	$80.3_{-2.0}$

Table 5.8: Results on the development set of SQuAD v2.0 (Rajpurkar et al., 2018) for the subsets with normal (Has-ans) and no-answer (No-ans) questions.

sometimes require answers with numbers) and s_{11} pronoun coreference resolution in DuoRC (consisting of movie scripts).

For SQuAD v2.0, we observed that the model recall increases for the no-answer questions. Because F1 score is computed between the has- and no-answer question subsets, the scores tend to be higher than those for SQuAD v1.1. See Table 5.8 for detailed numbers.

Ablation method \ Dataset	CoQA	DuoRC	Hotpot-QA	SQuAD v1.1	SQuAD v2.0
Original dataset	$77.4_{0.0}$	$58.4_{0.0}$	$63.6_{0.0}$	$91.5_{0.0}$	$81.9_{0.0}$
Drop all Q words	$6.7_{-91.3}$	$10.8_{-81.6}$	$10.0_{-84.2}$	$12.0_{-86.9}$	$50.1_{-38.9}$
Trained & evaluated on					
3′. Content words only	$71.0_{-8.3}$	$51.1_{-12.6}$	$61.7_{-3.0}$	$85.4_{-6.6}$	$74.8_{-8.7}$
6′. Context word shuff.	$52.9_{-31.7}$	$40.2_{-31.2}$	$46.1_{-27.4}$	$68.0_{-25.7}$	$80.6_{-1.7}$
7′. Sentence word shuff.	$68.3_{-11.8}$	$47.7_{-18.4}$	$66.8_{5.0}$	$82.4_{-9.9}$	$80.3_{-2.0}$

Table 5.9: Results of further analyses on the answer extraction datasets: the performance (%) after dropping all question (Q) and/or context (C) words, and that of the baseline model both trained and evaluated on the modified inputs.

Ablation method \ Dataset	ARC	MCTest	MultiRC	RACE	SWAG	Rel. avg.
Original dataset	$52.7_{0.0}$	$87.8_{0.0}$	$78.0_{0.0}$	$68.8_{0.0}$	$85.4_{0.0}$	0.0
Drop all Q words	$36.6_{-30.6}$	$61.6_{-29.9}$	$53.2_{-31.8}$	$55.4_{-19.5}$	$76.9_{-10.0}$	-50.5
Drop all C words	$40.3_{-23.6}$	$32.5_{-63.0}$	$61.7_{-20.9}$	$41.0_{-40.4}$	$71.7_{-16.0}$	-32.8
Drop all C&Q words	$29.9_{-43.3}$	$35.3_{-59.8}$	$57.2_{-26.7}$	$34.9_{-49.3}$	$62.1_{-27.3}$	-41.3
Trained & evaluated on						
3′. Content words only	$49.0_{-7.0}$	$80.6_{-8.2}$	$74.5_{-4.4}$	$58.4_{-15.2}$	$84.3_{-1.4}$	-7.5
6′. Context word shuff.	$46.6_{-11.5}$	$55.3_{-37.0}$	$70.1_{-10.2}$	$54.7_{-20.5}$	$83.6_{-2.1}$	-19.9
7′. Sentence word shuff.	$47.7_{-9.6}$	$75.0_{-14.6}$	$73.6_{-5.6}$	$59.2_{-14.0}$	$84.0_{-1.6}$	-8.2

Table 5.10: Results of further analyses on the multiple choice datasets: the performance (%) after dropping all question (Q) and/or context (C) words, and that of the baseline model both trained and evaluated on the modified inputs.

5.4.3 Further Analyses

To complement the observations in Section 5.4.2, we performed further experiments as follows.

The whole question and/or context ablation. To correctly interpret the result for s_1, we should know the performance on the *empty questions*. Likewise, for multiple-choice questions, the performance on the *empty context* should be investigated to reveal biases contained in the answer options. Therefore, we report the baseline results on the whole question and/or context ablations.[4]

Our results are reported in Tables 5.9 and 5.10. Although the performance on SQuAD v2.0 was relatively high, we found that the model predicted *no answer* for all of the questions (in this dataset, almost half of the questions are *no answer*). The other answer extraction datasets showed a relative drop of 80–90%. This result is not surprising since this setting forces the model to choose an answer span arbitrarily. On the multiple-choice datasets, on the other hand, the accuracies were higher than those

[4] This approach was already investigated by Kaushik and Lipton (2018). However, there is no overlap in datasets between ours and those they analyzed other than SQuAD v1.1.

of random choice (50% for MultiRC and 25% for the others), which implies that some bias exists in the context and/or the options.

Training and evaluating on the modified context. A question that was raised during the main analysis is what would happen if the model was trained on the modified input. For example, given that the performance with the content words only is high, we would like to know the upper bound performance when the model is forced to ignore function words also during training. Hence we trained the model with the ablations for the following skills: s_3 content words only; s_6 context word shuffle; and s_7 sentence word shuffle.

The results are reported in the bottom rows of Tables 5.9 and 5.10. On almost all datasets, the baseline model trained on the ablation training set (s'_3, s'_6, and s'_7) displayed higher scores than that on the original training set (s_3, s_6, and s_7). On CoQA, for instance, the relative change from the original score was only -8.3% when the model was trained on s_3 content words only. Although s'_3 and s'_7 with RACE were exceptions, their learning did not converge within the specified number of epochs. We observed that for all datasets the relative upper bounds of performance were on average 92.5%, 80.1%, and 91.8% for s_3, s_6, and s_7, respectively. These results support our observations in Section 5.4.2, that is, the questions allow solutions that do not necessarily require these skills, and thus fall short of testing precise NLU. Even without tuning on the ablation training set, however, our methods can make an optimistic estimation of questions that are possibly dubious for evaluating intended skills. In Figures 5.3 and 5.4, we compare the results of s_3, s_4, s_6, s_7, and s_8 when the model is trained on the ablated training set.

Data leakage in BERT for SWAG. BERT's performance on SWAG is close to the performance by humans (88.0%). However, the questions and corresponding options for SWAG are generated by a language model trained on the BookCorpus (Zhu et al., 2015), on which BERT's language model is also pretrained. We therefore suspect that there is severe data leakage in BERT's language model as reported in Zellers et al. (2019b). To confirm this issue, we trained a model without the context (i.e., the first given sentence). The accuracy on the development set, which was also without the context, was 74.9% (a relative decrease of 12.2%). This result suggests that we need to pay more attention to the relations of corpora on which a model is trained and evaluated, but leave further analysis for future work.

5.5 Qualitative Evaluation

In this section, we qualitatively investigate our ablation methods in terms of the human solvability of questions and the reconstructability of ablated features.

We analyze questions of SQuAD v1.1 and RACE which cover both answering styles and are influential in the community. We randomly sampled 20 questions from each dataset that are correctly solved (100% F1 and accuracy) by the baseline model on the original datasets. Our analysis covers four ablation methods (σ_3 content words only (involving $\sigma_{10,11,12}$), σ_4 vocabulary anonymization, σ_6 context word shuffle, and σ_7 sentence word shuffle) which provided specific insights in Section 5.4.

Method \ Dataset	SQuAD v1.1		RACE	
	Human	Baseline	Human	Baseline
3. Content words only	**100.0**	86.7	**95.0**	90.0
4. Vocab. anonymization	70.0	**77.6**	10.0	**25.0**
6. Context words shuff.	40.0	**53.3**	30.0	**75.0**
7. Sentence words shuff.	70.0	**70.5**	75.0	**85.0**

Table 5.11: Comparison of the human solvability and the baseline model's performance (%) on questions that are sampled from the ablation tests.

5.5.1 Human Solvability after the Ablation

Motivation. In Section 5.4, we observed that the baseline model exhibits remarkably high performance on some ablation tests. To interpret this result, we investigate if a question is solvable by humans and the model. Concretely, the question after the ablation can be (A) solvable by both humans and the model, (B) solvable by humans but unsolvable by the model, (C) unsolvable by humans but solvable by the model, or (D) unsolvable by both humans and the model. For Case A, the question is easy and does not require complex language understanding. For Cases B and C, the model may use unintended solutions because (B) it does not use the same solution as humans or (C) it *cleverly* uses biases that humans cannot recognize. For Case D, the question may require the skill intended by the ablation method. Although Cases A to C are undesirable for evaluating the systems' skills, it seems to be useful to distinguish them for further improvement of the dataset creation. We therefore perform the annotation of questions with human solvability; We define that a question is solvable if a reasonable rationale for answering the question can be found in the context.

Results. Table 5.11 shows the human solvability along with the baseline model's performance on the sampled questions. The model's performance is taken from the model trained on the original datasets except for the vocabulary anonymization method. For the content words only on both datasets, the human solvability is higher than the baseline performance. Although these gaps are not significant, we might be able to infer that the baseline model relies on content words more than humans (Case B). Given that the high performance of both humans and the baseline model, most of the questions fall into Case A, i.e., they are easy and do not necessarily require complex reasoning involving the understanding of function words.

For the other three methods, the human solvability is lower than the baseline performance. This result indicates that the questions correctly solved only by the baseline model may contain unintended biases (Case C). For example, the gap in the context word shuffle of RACE is significant (30.0% vs. 75.0%). Figure 5.5 shows a question that is unsolvable for humans but can be solved by the baseline model. We conjecture that while humans cannot detect biases easily, the model can exploit biases contained in the answer options and their relations to the given context.

5.5.2 Reconstructability of Ablated Features

Motivation. We also seek to investigate the reconstructability of ablated features. Even if a question falls under Case A in the previous section, it might require the

Method \ Dataset	SQuAD v1.1				RACE			
	α	β	γ	δ	α	β	γ	δ
3. Content words only	.45	.00	.55	.00	.80	.05	.15	.00
4. Vocab. anonymization	.70	.30	.00	.00	.10	.90	.00	.00
6. Context words shuff.	.40	.60	.00	.00	.30	.70	.00	.00
7. Sentence words shuff.	.70	.30	.00	.00	.70	.25	.05	.00

Table 5.12: Frequency of questions for Cases α to δ for SQuAD v1.1 (Rajpurkar et al., 2016) and RACE (Lai et al., 2017). Ablated features are (α) unreconstructable and unnecessary, (β) unreconstructable and necessary, (γ) reconstructable and unnecessary, and (δ) reconstructable and necessary. Questions for Case δ are problematic for interpreting our main observations.

skill intended by the ablation; If a reader is able to *guess* the dropped information and uses it to solve the question, we cannot say that the question does not require the corresponding skill. For example, even after dropping function words (σ_3), we might be able to guess which function word to fill in a cloze based on grammaticality and lexical knowledge. Such *reconstructable* features possibly exist for some ablation methods. However, they are not critical if they are unnecessary for answering questions. We can list the following cases: ablated features are (α) unreconstructable and unnecessary, (β) unreconstructable and necessary, (γ) reconstructable and unnecessary, and (δ) reconstructable and necessary. To verify that ablation methods work, we need to confirm that there are few questions of Case δ. The other cases are not critical to our observations in the main experiment. We therefore perform the annotation with the following queries: (i) *are ablated features reconstructable?* and (ii) *are reconstructable features really necessary for answering?* When the answers for both queries are yes, a question is in Case δ. In the annotation, we define that features in a question are reconstructable if the features existing around the rationale for answering the question are guessable. We also require that these features are necessary to decide the answer if the correct answer becomes undecidable without them.

Results. For both datasets, the annotation shows that, not surprisingly, almost all features are unreconstructable in the shuffled sentence/context words and the vocabulary anonymization (except for one example in RACE). When these questions are solvable / unsolvable by humans, we can say that features are unnecessary (Case α) / necessary (Case β) for answering the questions. In contrast, the annotators could guess function words for some questions even if these words are dropped (SQuAD: 55.0% and RACE: 15.0%). The annotation of the necessity also shows that, however, reconstructable features (function words in this case) for all the questions are not necessary to answer them (i.e., Case γ). Therefore, we could not find any question in Case δ.

We report the annotation results in Table 5.12. It is not easy for the annotator to completely ignore the information of reconstructed features. We leave designing a solid, scalable annotation scheme for future work.

In summary, we found that almost all ablated features are unreconstructable. Although for some questions ablated features are reconstructable for the content words only, these words are not necessarily required for answering the questions. Overall, this result supports our observations in Section 5.4, i.e., questions already solved in existing datasets do not necessarily require complex language understanding.

5.6 Discussion

In this section, we discuss two requirements for developing the MRC task as an NLU benchmark.

The control of question solvability. Not to allow the model to focus on unintended objectives, we need to ensure that each question is unsolvable without its intended requisite skill. Therefore, when benchmarking, we first need to identify necessary features whose presence determines the question's solvability. To identify them, we might need to perform ablation testing with humans. Further, we need to evaluate a model in both regular and ablation settings. This is because a model may detect some biases that enable it to solve the question; such biases can actually be false for humans and may be acquired by the model through overfitting to datasets. Nonetheless, there is a case in which, even if we can identify necessary features, the model can have prior, true knowledge (e.g., world knowledge) of the correct answer. In this case, the model can answer the question without the context. To avoid this circumvention, we may need to evaluate the model on fictional texts.

Comprehensiveness of requisite skills. Another aspect of NLU benchmarking is the comprehensiveness of skills. Our proposed approach can be expanded in two further directions: (i) inner-sentence and (ii) multiple-sentence levels. For (i), we can focus on understanding of specific linguistic phenomena. This includes logical and semantic understanding such as in FraCaS (Cooper et al., 1994) and SuperGLUE (Wang et al., 2019). To investigate particular syntactic phenomena, we might be able to use existing analysis methods (Marvin and Linzen, 2018). For (ii), our skills can include complex/implicit reasoning, e.g., spatial reasoning (Weston et al., 2015) and lexically dependent causal reasoning (Sap et al., 2019). Although we do not need to include all of these skills in a single dataset, we need to consider the generalization of models across them.

5.7 Conclusion

Existing analysis work in MRC was largely concerned with evaluating the capabilities of *systems*. By contrast, in this chapter, we proposed an analysis methodology for the benchmarking capacity of *datasets*. Our methodology consisted of input-ablation tests, in which each ablation method was associated with a skill requisite for MRC. We exemplified our ablation-based methods along with 12 requisite skills and analyzed 10 existing datasets. The experimental results suggested that for benchmarking sophisticated NLU, datasets should be more carefully designed to ensure that questions correctly evaluate the intended skills.

▷ CoQA / source: mctest, id: 3tdxmtx3cbu3qs5x4zz64vf5jpxi67, turn-id: 6

Context

On a snowy winter morning , the brown-haired lady saw a squirrel that was hurt . It only had three legs , and it looked hungry . She put some corn out for the squirrel to eat , but other bully squirrels came , too . The brown-haired lady started giving the little squirrel peanuts to eat . She gave some to the bully squirrels , too , so they would leave the three-legged squirrel alone . The winter snow melted and then it was spring . the grass turned green and the air was warm . Now , when the little squirrel with three legs would come to see the brown-haired lady with the peanuts , it would take the peanuts and dig a little hole and hide the peanuts for later . The squirrel would hold the peanut in its mouth and dig and dig and dig , and then it would put the peanut in the hole and pat it down with its little front paw . Then it would run back over to the brown-haired lady and get some more peanuts to eat . unknown yes no

Was he hungry ? yes

Context with dropped function words

[redacted] [redacted] snowy winter morning , [redacted] brown-haired lady saw [redacted] squirrel [redacted] [redacted] hurt . [redacted] [redacted] [redacted] three legs , [redacted] [redacted] looked hungry . [redacted] put [redacted] corn [redacted] [redacted] [redacted] squirrel [redacted] eat , [redacted] [redacted] bully squirrels came , [redacted] . [redacted] brown-haired lady started giving [redacted] little squirrel peanuts [redacted] eat . [redacted] gave [redacted] [redacted] [redacted] bully squirrels , [redacted] , [redacted] [redacted] would leave [redacted] three-legged squirrel alone . [redacted] winter snow melted [redacted] [redacted] [redacted] [redacted] spring . [redacted] grass turned green [redacted] [redacted] air [redacted] warm . [redacted] , [redacted] [redacted] little squirrel [redacted] three legs would come [redacted] see [redacted] brown-haired lady [redacted] [redacted] peanuts , [redacted] would take [redacted] peanuts [redacted] dig [redacted] little hole [redacted] hide [redacted] peanuts [redacted] later . [redacted] squirrel would hold [redacted] peanut [redacted] [redacted] mouth [redacted] dig [redacted] dig [redacted] dig , [redacted] [redacted] [redacted] would put [redacted] peanut [redacted] [redacted] hole [redacted] pat [redacted] [redacted] [redacted] [redacted] little front paw . [redacted] [redacted] would run back [redacted] [redacted] [redacted] brown-haired lady [redacted] get [redacted] [redacted] peanuts [redacted] eat . unknown yes no

[redacted] [redacted] hungry ? yes

Question

What did the lady put out for the squirrel ?

Answer

corn

Prediction before and after dropping function words

corn → peanuts

Figure 5.2: Examples of questions after the application of our ablation method σ_3 (content words only). Blacked-out text denotes words that were dropped by the ablation function.

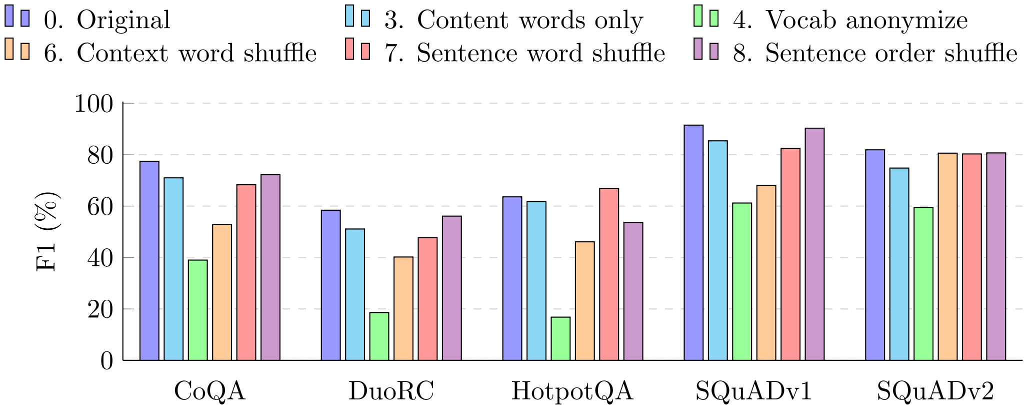

Figure 5.3: Excerpted results of s_3, s_4, s_6, s_7, and s_8 for the answer extraction datasets where the model is trained on the ablated training set.

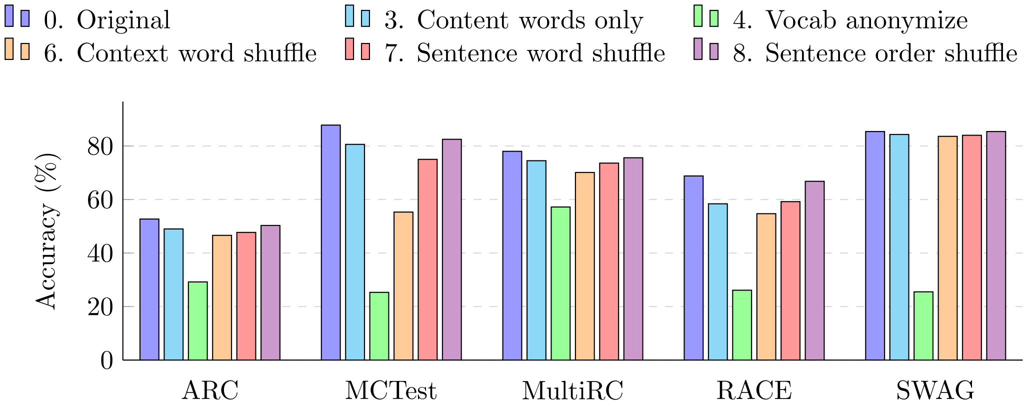

Figure 5.4: Excerpted results of s_3, s_4, s_6, s_7, and s_8 for the multiple choice datasets where the model is trained on the ablated training set.

Original context

[...] By now you have probably heard about Chris Ulmer, the 26-year-old teacher in Jacksonville, Florida, who starts his special education class by calling up each student individually to give them much admiration and a high-five. I couldn’t help but be reminded of Syona’s teacher and how she supports each kid in a very similar way. Ulmer recently shared a video of his teaching experience. All I could think was: how lucky these students are to have such inspirational teachers. [...]

Context with shuffled context words

[...] their with and to kids combined , t always of (has) mean problems the palsy five cerebral that communication , her standard ” assess (. teacher a a now gesture Florida admiration and , much calling Ulmer to individually (of class his heard Jacksonville year special you up Chris greeting five) congratulation by give education who , them or about probably the in by each - student high , old - - have starts 26 . I s she similar reminded be ’ each t and in help ’ kid teacher [...]

Question

What can we learn about Chris Ulmer?

Options (the answer is in bold)

(A) He praises his students one by one. (B) He is Syona’s favorite teacher. (C) He use videos to teach his students. (D) He asks his students to help each other.

Figure 5.5: Example of questions with shuffled context words from RACE (Lai et al., 2017). Although the question appears unsolvable for humans, the baseline model predicts the correct answer.

Chapter 6

Discussion: Requirements for the Explainability

Machine reading comprehension (MRC) is receiving lots of attention in natural language processing in the past few years, and many datasets have been published. However, the conventional task design of MRC lacks explainability beyond the model interpretation, i.e., the internal mechanics of the model cannot be explained in human terms. To this end, this chapter provides a theoretical basis for the design of MRC tasks 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.

6.1 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 proposed, and machine learning 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. In Chapter 5, we 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 context is not necessary. These studies proved that we cannot *explain* what kind of understanding is actually 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 model and to test hypotheses for reading comprehension, its theoretical foundation is lacking in the existing literature.

In this chapter, 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

Question	What is reading comprehension?	How can we evaluate reading comprehension?
Foundation	Representation levels in human reading comprehension: (A) surface structure, (B) textbase, and (C) situation model.	Construct validity in psychometrics: (1) content, (2) substantive, (3) structural, (4) generalizability, (5) external, and (6) consequential aspects.
Requirements	(A) Linguistic-level understanding, (B) making the comprehensiveness of skills for inter-sentence understanding, and (C) evaluation of coherent and grounded representation.	(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.
Direction	(C) Context dependency with the defeasibility and novelty, and non-textual grounding with a long passage.	(2) Improving the question quality by filtering and ablation, and designing a task formulation for visualizing the internal process.

Table 6.1: An overview of theoretical foundations, requirements, and future directions of MRC discussed in this chapter.

those who create MRC datasets and analyze the behavior of model. In the context of explainability, Gilpin et al. (2018) indicated that interpreting the internals of a system are closely related to its architecture only and are insufficient for explaining how the task is accomplished. This is because even if the internals of models can be interpreted, we cannot explain what is measured by the datasets. Therefore, our study focuses on the explainability of the task and datasets rather than on the interpretability of models.

The remainder of this chapter is organized as follows. An overview of explanation issues in MRC is presented in Section 6.2. We also review the analytical studies, which indicated existing datasets might fail to correctly evaluate their intended behavior. Subsequently, we present the psychological study of human reading comprehension in Section 6.3 for answering the *what* question (i). We argue that the concept of *representation levels* can serve as a conceptual hierarchy for organizing the existing technologies in MRC. Section 6.4 focuses on answering the *how* question (ii). Here, we implement psychometrics to analyze the prerequisites for the task design of MRC. Further, we introduce the concept of *construct validity*, which emphasizes on validating the performance of model interpretation in the task. Finally in Section 6.5, we discuss the future directions toward the advancement of MRC. Regarding the *what* question, we indicate that datasets should evaluate the capability of the *situation model*, which refers to the construction of a coherent and grounded representation of text based on human understanding. Regarding the *how* question, we argue that among the important aspects of the construct validity, *substantive validity* must be ensured, which necessitates the verification of the internal mechanism of comprehension.

Table 6.1 gives an overview of theoretical foundations, requirements, and future directions of MRC discussed in this chapter. Our conclusions for the further development of MRC are as follows:

- MRC may be the most suitable task for evaluating NLU. Situation model can be the next focal point for evaluating and achieving human-level understanding of natural language.

- The substantive validity for the explainability of the internal mechanism of NLU must be ensured by improving the question quality and by designing a white-box task.

6.2 Explanation Issues

In this section, we describe the analytical studies that revealed a major drawback of these datasets, i.e., the lack of explainability for reading comprehension. In some datasets, the performance of machines has already reached human-level performance. However, Jia and Liang (2017) indicated that models can be easily fooled by manual injection of distracting sentences. They highlighted that existing models may not understand given passages precisely. Although this does not imply that machine learning models cannot solve such adversarial questions even when these questions are given in their training (Liu et al., 2019b), they revealed that the questions simply gathered by crowdsourcing without careful guidelines or constraints are insufficient to precisely evaluate NLU.

This argument has been supported by further studies on existing datasets. For example, Min et al. (2018) found that more than 90% of the questions in SQuAD (Rajpurkar et al., 2016) required obtaining an answer from a single sentence despite being provided with a passage. In Chapter 4, we showed that large parts of 12 datasets were easily solved only by observing the few tokens of the first question and by analyzing the similarity between the given questions and context. Similarly, Feng et al. (2018) and Mudrakarta et al. (2018) demonstrated that models did not change their predictions even when the question tokens were partly dropped in SQuAD. Kaushik and Lipton (2018) also observed that question- and passage-only models often perform well. In Chapter 5, we observed that the already solved questions in existing datasets could be solved even after shuffling words in the sentence or dropping content words, which indicated that these questions did not necessarily require complex understanding of given texts. Min et al. (2019a) and Chen and Durrett (2019) concurrently indicated that for 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 contained stylistic biases in the answer options, were embedded by a language-based model. These biases were detrimental to the performance of dataset in requiring commonsense reasoning.

Overall, these investigations highlight a grave issue of the task design, i.e., even if the models show human-level scores, we cannot prove that they successfully perform reading comprehension. This issue may be attributed 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 can interpret the internals of models, we cannot explain what is measured by the datasets. We speculate that the explainability issue in MRC can be attributed to the following two points:

- We do not have a comprehensive theoretical basis of reading comprehension for specifying what we should ask. (Section 6.3)

- We do not have a well-established methodology for creating a dataset and for analyzing the performance of a model based on it. (Section 6.4)

In the remainder of this chapter, we argue that these issues can be addressed by using insights from the psychological study of reading comprehension and by implementing psychometric means of validation.

6.3 Reading Comprehension from Psychology to MRC

6.3.1 Computational Model in Psychology

Human text comprehension has been studied in psychology for a very long time (McNamara and Magliano, 2009; Kintsch and Rawson, 2005; Zwaan and Radvansky, 1998; Graesser et al., 1994; Kintsch, 1988). Connectionist and computational architectures have been proposed for such comprehension including a mechanism pertinent to knowledge activation and memory storing. Among the computational models, the construction–integration (CI) model is the most influential model and provides a strong foundation of 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 an event, which is often elaborated by a reader's background knowledge. (ii) The *integration* step includes the process of associating the propositions and create their network. These steps are not exclusive, i.e., propositions are iteratively updated in accordance with the surrounding propositions with which they are linked.

Besides, the CI model assumes that these steps 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 and coherent mental representation in which the propositions are globally connected (*macrostructure*), and it is often grounded to not only texts but also to sounds, images, and personal information.

To recapitulate, the CI model first decodes textual information (i.e., the surface structure) from the raw textual input, then create the propositions (i.e., textbase) and their local connections occasionally using the reader's knowledge, and finally construct a coherent representation (i.e., situation model) that is organized according to five dimensions including space, causation, intentionality, objects, and time (Zwaan and Radvansky, 1998)), which provides a global description of the events. Although the definition of successful reading comprehension can vary, Hernández-Orallo (2017b) stated that the objective of text comprehension is to create a situation model that best explains the given text and the reader's background knowledge. This definition effectively proves that the situation model plays a vital role in human reading comprehension.

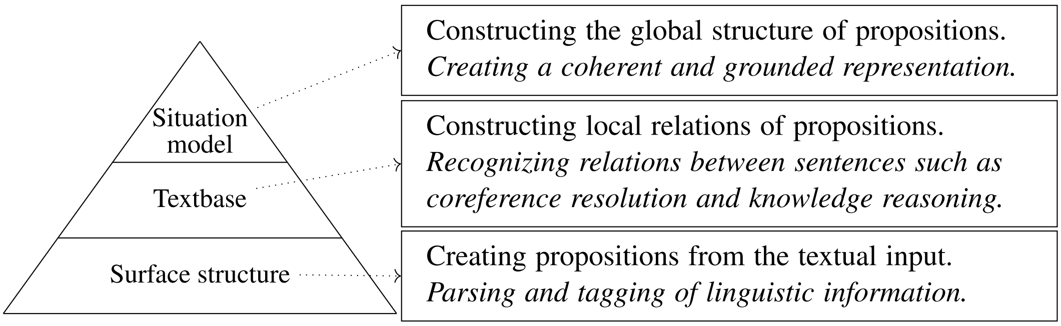
Situation model
Textbase
Surface structure
Constructing the global structure of propositions.
Creating a coherent and grounded representation.
Constructing local relations of propositions.
Recognizing relations between sentences such as coreference resolution and knowledge reasoning.
Creating propositions from the textual input.
Parsing and tagging of linguistic information.

Figure 6.1: Representation levels and corresponding natural language understanding skills.

What can we learn from these psychological theories? A major difference between human and machine reading comprehension is their distinct architecture. Therefore, we do not need to adhere to the human brain's architecture for developing MRC models. Our aim in this section is to provide a basis for defining reading comprehension, which requires *units* of explanation (Doshi-Velez and Kim, 2018). In the computational model above, the representation levels appear to be useful for organizing such units. Therefore, our goal in Section 6.3.2 is to ground existing NLP technologies and tasks to different representation levels.

6.3.2 Skill Hierarchy for MRC

Here, we associate the existing NLP tasks with the three representation levels introduced above. The biggest advantage of MRC is its general formulation, which makes it the most general task for evaluating NLU. This emphasizes the importance of the requirement of various *skills* in MRC, which can serve as the units for the explanation of reading comprehension. Therefore, our motivation is twofold: (i) to provide an overview of the skills as a hierarchical taxonomy and (ii) to highlight the missing aspects in existing MRC datasets that are required for comprehensively covering the representation levels.

Existing taxonomies. To understand the existing tasks and technologies, we first provide a brief overview of the existing taxonomies of *skills* in the context of NLU tasks. For recognizing textual entailment (Dagan et al., 2006), several studies presented a classification of reasoning and commonsense knowledge (Bentivogli et al., 2010a; Sammons et al., 2010; LoBue and Yates, 2011). For scientific question answering, Jansen et al. (2016) categorized knowledge and inference for an elementary-level dataset. Similarly, Boratko et al. (2018) proposed types of knowledge and reasoning for scientific questions in MRC (Clark et al., 2018). A limitation of both these studies is that the proposed sets of knowledge and inference are limited to the domain of elementary-level science. Although some existing datasets for MRC have their own classifications of skills, they are coarse and only cover a limited extent of typical NLP tasks (e.g., word matching and paraphrasing). Among them, multiple-sentence reasoning is too simplified in which several types of relations can exist between sentences (Khashabi et al., 2018a). In contract, for a more generalized definition, we proposed a set of 13 skills for MRC in Chapter 3. However, these skills were defined at a single level, and multiple representation levels were not considered.

In what follows, we describe the three representation levels that basically follow the three representations of the CI model but are modified for MRC. The three levels are shown in Figure 6.1. 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, which can serve as a foundation to highlight the missing aspects in existing MRC.

Surface structure. This level broadly covers the linguistic information and its semantic meaning, which can be based on the raw textual input. Although these features form a proposition according to psychology, it 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 these basic tasks can be accomplished by some recent pretraining-based neural language models (Liu et al., 2019a), they are hardly required in NLU tasks including MRC. McCoy et al. (2019) indicated that the natural language inference (NLI) task (Bowman et al., 2015) fails to ask the syntactic understanding of given sentences. White et al. (2017), McCoy et al. (2019), and Kim et al. (2019) suggested that local-level tasks including probing tasks require sentence-level semantics and syntax. For MRC, we also indicated that questions are solvable even after dropping function words in Chapter 5. Although it is not obvious that these basic tasks should be included in MRC, we should always care about the capabilities of basic tasks when the performance of a model is being assessed (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, this implies several types of relations between sentences. These relations not only include the typical relations between sentences (discourse relations) but also the links between entities such as coreference resolution and knowledge reasoning. In addition, we can include mathematical reasoning and logical reasoning. Consequently, 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, and mathematical 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 provide comprehensive definition of knowledge and commonsense reasoning, non-textual reasoning and knowledge are not included in this level. For example, Davis and Marcus (2015) indicated 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. Talmor and Berant (2019) also found that training on multiple datasets leads to robust generalization. However, because requisite skills are not identified, their attempts still lacks explainability. Therefore, we should create a dataset in which the skills at this level are comprehensively identified instead of a dataset focusing on a single skill.

Situation model. This level targets the global structure of propositions in human reading comprehension. In It includes a coherent and situational representation of a given context and its grounding to non-textual information. A coherent representation has well-organized sentence-to-sentence transitions (Barzilay and Lapata, 2008), which are vital for using procedural and script knowledge. However, most existing MRC datasets fail to target the situation model for a coherent understanding of given texts and grounding to non-textual information. The future directions for this level are presented in Section 6.5.1.

In summary, we indicate that the following features are missing in existing datasets:

- Considering the capability to acquire basic understanding of the linguistic-level information.
- Ensuring that the 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.

6.4 Machine Reading Comprehension on Psychometrics

In this section, we provide a theoretical foundation for the explainable evaluation of MRC models. In this framework, a key concept is validity. To elaborate, as MRC measures the capability of reading comprehension, validation of the measurement is crucial to obtain reliable and useful explanation. Therefore, we focus on psychometrics—a field of study concerned with the assessment of the quality of psychological measurement (Furr, 2018). We assume that the insights obtained from psychometrics can facilitate a better task design. In Section 6.4.1, we first review 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. Subsequently, in Section 6.4.2, we examine the aspects that correspond to construct validity in MRC and then indicate the prerequisites for verifying the intended explanation of MRC in its task design.

6.4.1 Construct Validity in Psychometrics

According to psychometrics, *construct validity* is necessary to validate the interpretation of outcomes of psychological experiments.[1] Messick (1995) reported that construct validity consists of the following six aspects.

1. The *content aspect* refers to the match between the actual content of a test and the content to be revealed in the test.
2. The *substantive aspect* refers to theoretical rationales for the observed consistencies in test responses including process models for task performance.

[1] In psychology, a construct is an abstract concept, which facilitates the understanding of human behavior such as vocabulary, skills, and comprehension. A construct can be seen as a unit for the explanation in this chapter.

Validity aspects	Definition in psychometrics	Correspondence in MRC
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 6.2: Aspects of the construct validity in psychometrics and corresponding features in reading comprehension.

3. The *structural aspect* appraises the fidelity of the scoring structure to the structure of the construct domain at issue.

4. The *generalizability aspect* refers to the extent to which score interpretations generalize to and across population groups, settings, and tasks (i.e., reliability).

5. The *external aspect* refers to the extent to which the assessment scores' relationship with other measures and non-assessment behavior reflect the expected relations.

6. The *consequential aspect* appraises the value implication of score interpretation as a basis for test use.

In the design of educational and psychological measurement, these aspects collectively provide verification questions that need to be answered for justifying the interpretation and use of test scores. In this sense, the construct validation can be viewed as an empirical evaluation of the meaning and consequence of psychological measurement.

Given that MRC is intended to capture the capability of reading comprehension, the task designers must consider these validity aspects as much as possible. Otherwise, users of the task cannot justify the score interpretation, i.e., it cannot be confirmed that successful systems actually perform intended reading comprehension.

6.4.2 Construct Validity in MRC

The six aspects of construct validity and the corresponding MRC features are summarized in Table 6.2. In this section, we associate these aspects with MRC and discuss the requisites for validating the score interpretation in MRC.

In what follows, we discuss the missing aspects that are needed to achieve the construct validity of the current MRC.

Content aspect. As discussed in Section 6.3, sufficiently covering the skills across all the representation levels is an important requirement of MRC. Therefore, 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 on different domains of corpus. There are two important points that should be considered for the content aspect of the construct validity in MRC: *coverage* and *representativeness*.

Substantive aspect. This aspect appraises the evidence for the consistency of model behavior. We consider that this is the most important aspect for 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, it is necessary to ensure that the questions correctly assess the internal steps in the process of reading comprehension. However, as stated in Section 6.2, most existing datasets fail to verify that a question is solved by using an intended skill, which implies that it cannot be proved that a successful system can actually perform intended reading comprehension. The substantive aspect is further discussed in Section 6.5.2.

Structural aspect. Another issue in most existing datasets is that they only provide simple accuracy as a metric. Given that the substantive aspect necessitates the evaluation of the internal process of reading comprehension, the structure of metrics needs to reflect it. However, only few studies have attempted to provide a dataset with multiple metrics. For example, QuAC (Choi et al., 2018), a dialogue-based dataset, introduced a metric for the percentage of dialogues, whose entire questions were correctly answered by the system. If consecutive questions in a dialogue were mutually dependent, this metric was able to evaluate the understanding of a given dialogue within accompanying questions. Another example is HotpotQA (Yang et al., 2018), which not only asks for the answers to questions but also provides sentence-level *supporting facts*. This metric can also evaluate the process of multi-hop reasoning whenever the supporting sentences need to be understood for answering a question. Therefore, we need to consider both substantive and structural aspects simultaneously.

Generalizability aspect. The generalizability of MRC can be understood from two perspectives: (i) the reliability of metrics and (ii) the reproducibility of findings.

For (i), reliability may be an issue in the context of given correct answers and model predictions. Regarding the accuracy of answers, the performance and interpretation of the model become unreliable when the answers are unintentionally ambiguous or impractical. During sourcing of a dataset, the questions could be unintentionally ambiguous or unanswerable. Because in most datasets, the correct answers are just decided by the majority vote of crowd workers, the ambiguity of the answers is not considered. It may be useful if such ambiguity can be reflected in the evaluation metrics (e.g., using the item response theory for recognizing textual entailment (Lalor et al., 2016)). Regarding model predictions, the reproducibility of results (Bouthillier et al., 2019), which implies that the reimplementation of system generates statistically similar predictions, may be an issue. As Dror et al. (2018) pointed out, it is rarely confirmed that the generated results are statistically significant in NLP. For the reproducibility of models, we should use statistical testing methods to evaluate MRC models.

For (ii), Bouthillier et al. (2019) stressed the reproducibility of findings, i.e., the transferability of findings in a dataset to another dataset. In other words, a unit for the explanation must be established, which should be common in both datasets. Such units were called *cognitive chunks* by Doshi-Velez and Kim (2018) in the context of the explainability of machine learning models. Therefore, the generalizability aspect highlights the importance of content aspect.

External aspect. Although this aspect is important in psychometrics, it might be relatively insignificant in MRC due to the difference in their objectives (psychological measurement versus the development of systems). Nonetheless, to develop a general NLU system, it is necessary to evaluate it on various NLU tasks such as MRC, NLI, dialogue, and visual question answering. In addition, it is 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 the use of a successful model in practical situations other than tasks. Wallace et al. (2019) showed that existing NLP models are vulnerable to adversarial examples and thereby generate egregious outputs. Therefore, we should focus on the robustness of a model to adversarial attacks and the accountability for unintended model behaviors.

6.5 Future Directions

This section discusses the future directions of MRC toward answering the *what* and *how* questions introduced in Sections 6.3 and 6.4. In particular, we infer that the *situation model* and *substantive validity* are critical for developing human-level explainable MRC.

6.5.1 *What* question: Evaluating Situation Models

As mentioned in Section 6.3, existing datasets fail to assess the situation model in reading comprehension. As a future direction, we suggest that the task should deal with two features of the situation model: context dependency and grounding to non-textual information.

Context-dependent Situations

A vital feature of the situation model is that it is conditioned on a given text, i.e., a representation is constructed distinctively depending on the given context. In this chapter, this property is called *context dependency*. We elaborating it by discussing the following two key features: defeasibility and novelty.

Defeasibility. The defeasibility of a constructed representation implies that a reader can modify and revise it according to the newly acquired information (Davis and Marcus, 2015; Schubert, 2015). Although the defeasibility of NLU is tackled in the task of if-then reasoning (Sap et al., 2019), abductive reasoning (Bhagavatula et al., 2019), and counterfactual reasoning (Qin et al., 2019), only few reports have addressed it in MRC. However, the defeasibility of reasoning is governed by the context, which may be a critical advantage 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 solver needs to combine multiple commonsense knowledge to derive the correct answer. Such a situation appears more easily in a longer MRC document rather than in a short sentence of NLI. If non-fiction documents such as newspaper and Wikipedia articles are only used, some questions just require the reasoning of facts already known in web-based corpus and do not require novel reasoning. Therefore, fictional narratives may be a better source for creating a dataset of novel questions.

Grounding to Other Media

As Burges (2013) raised as an issue, only a few MRC datasets can be used for grounding texts to non-textual information. For example, Kembhavi et al. (2017) proposed a dataset based on science textbooks, which contained multiple-choice questions with passages, diagrams, and images. Kahou et al. (2018) also proposed a figure-based question answering dataset that required the 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 have the following issues for accurately evaluating NLU: (i) skills required for answering questions are not identified; (ii) proposed models are likely to be domain- and task- specific, which lack generalizability to other domains and tasks; and (iii) most datasets do not contain long descriptions but short questions about images, which may cause flaws in obtaining a precise understanding of given texts. Therefore, it is 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).

6.5.2 How side: Assuring Substantive Validity

As we viewed in Section 6.4.2, the substantive validity requires to ensure that the questions correctly assess the internal steps of reading comprehension. Therefore, an obvious exigency is to assure the substantive validity of MRC datasets and to provide an explanation for it. We discuss two approaches for this challenge: creating *high-quality questions* and designing a *white-box task*.

Collecting High-quality Questions

Gururangan et al. (2018) revealed that NLU datasets can contain unintended biases embedded by annotators (*annotation artifacts*). If machine learning models exploit such biases for answering questions, we cannot evaluate the precise NLU of models. Therefore, we need to alleviate such biases by filtering out undesirable questions. Besides, for the explainability of MRC, we need to identify the skills 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 trained an ensemble of stylistic classifiers and used them to filter out the questions. Sakaguchi et al. (2020) also proposed filtering methods based on both machines and humans to alleviate *dataset-specific* and *word-association* biases. However, a major issue is the inability to discern 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 includes everyday experience, patterns are called *commonsense*. However, as mentioned in Section 6.5.1, commonsense knowledge can be *defeasible*, which implies that the knowledge can be false in unusual situations. Another type of commonsense is called the law of nature, which can be false in other possible worlds. Besides, when the domain is our real world, indefeasible patterns are called *factual knowledge*.

Therefore, the distinction of bias and knowledge depends on where the pattern is recognized. This means that a dataset should be created so that it can evaluate reasoning on the intended knowledge. For example, to test defeasible reasoning, we must filter out questions that are solvable by usual commonsense only. If we want to determine the ability of reading comprehension independently from factual knowledge, we can consider counterfactual or fictional situations. This also supports the importance of testing the situation model as discussed in Section 6.5.1.

Identifying requisite skills by ablating input features. Another approach is to verify the quality of questions by analyzing the human answerability of questions after ablating their key features. We speculate that if a question is still answerable by humans even after removing the features, the question does not require understanding of ablated features. We pointed out a similar hypothesis for using machines in Chapter 5. This method can be used for verifying that the 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, explainability needs to indicate textual features associated with certain skills as units for the explanation. In addition, answering a question is equivalent to choosing the correct answer from the candidate answers. Therefore, necessary features are critical for discriminating between different but semantically similar candidate answers (Khashabi, 2019). In summary, the task design should consider the collection of such similar candidates while identifying critical features. Winograd Schema Challenge (Levesque, 2011) is a successful task in this sense.

Formulating a White-box Task

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

Generating the introspective explanation. Inoue et al. (2019) classified two types of explanation in 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 the derivation of the answer for making the decision. Inoue et al. (2019) annotated introspective explanation with multi-hop reasoning questions and proposed a task that required the derivation of the correct answer of a given question to improve the explainability. Similarly, Rajani et al. (2019) collected human explanations for commonsense reasoning and improved the system's performance by modeling the generation of the explanation. Although gathering human explanations is costly, this approach can facilitate the explicit verification of a model's understanding.

Creating dependency between questions. 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 required a procedural understanding of scientific facts. In this dataset, a set of questions corresponds to the steps of the entire process of a scientific phenomenon. Therefore, this set can be viewed as a single question that requires a complete understanding of the scientific phenomenon. Yagcioglu et al. (2018) also proposed a dataset in the recipe domain in which the questions required an understanding of cooking procedure by choosing the correct order of the images to make a complete recipe. Dialogue-based datasets also contain mutually dependent questions. However, one issue with such questions is that the relationships between questions are not identified. Overall, these approaches can facilitate the explicit validation of a model's understanding.

6.6 Conclusion

In this chapter, we outlined the issues and future directions of the machine reading comprehension task. Specifically, we focused on the situation model in psychology to analyze *what* we should ask of reading comprehension and on the substantive validity in psychometrics to analyze *how* we should correctly evaluate it. We deduced 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.

Chapter 7

Conclusions

In this thesis, we discussed how we can evaluate the capability of natural language understanding in the task of machine reading comprehension.

In Chapter 2, we gave an overview of machine reading comprehension (MRC). We first provided a formal definition of the task, and then describe task components and representative existing datasets. Next, we introduced a short history of datasets and systems along with some trends in the field. We finally discussed the difference between MRC and other related natural language understanding tasks.

In Chapter 3, we adopted evaluation metrics that comprise two classes, namely refined prerequisite skills and readability, for analyzing the quality of MRC datasets. We applied these classes to six existing datasets and highlighted their characteristics according to each metric. Our dataset analysis suggested that difficult-to-read texts do not necessarily create difficult-to-answer questions, and that we could create an MRC dataset that is easy to read but difficult to answer.

In Chapter 4, we addressed how to investigate the quality of questions so that they can correctly evaluate intended language understanding. Proposing analysis methods to look into question difficulty and requisite skills, we examined MRC questions from 12 datasets to determine what makes such questions easier to answer. Using our defined heuristics, the datasets were split into easy and hard subsets. Our experiments revealed that the baseline performance degraded with the hard questions, which required knowledge inference and multiple-sentence reasoning compared to easy questions. These results suggested that one might overestimate the ability of the baseline systems.

In Chapter 5, we combined automatic evaluation and skill-focused metrics into a single evaluation methodology. Existing analysis work in this field mainly focused on evaluating the capabilities of systems. By contrast, in this chapter, we proposed an analysis methodology for the benchmarking capacity of datasets. Our methodology consisted of semi-automated input-ablation tests, in which each ablation method was associated with a skill requisite for MRC. We defined 12 requisite skills with corresponding ablation-based methods and analyzed 10 existing datasets. The experimental results suggested that for benchmarking sophisticated natural language understanding, datasets should be more carefully designed to ensure that questions correctly evaluate the intended skills.

In Chapter 6, we overviewed issues and future directions of the machine reading comprehension task. Particularly, we focused on the situation model in psychology for what we should ask in MRC and the substantive validity in psychometrics for how we should evaluate in MRC. We concluded that future datasets should evaluate the capability of the situation model for understanding context-dependent situations and for grounding to non-textual information and should ensure the substantive validity by improving the question quality and designing a white-box task formulation.

Altogether, our focus was around what language understanding is. To get close to this question, MRC could be the most suitable testbed, which enables us to evaluate our hypotheses on reading comprehension. We consider that providing verifiable explanations (falsifiable hypotheses) is the most important for the scientific study of reading comprehension. A detailed explanation is also essential to facilitate industry applications. Although significant progress has been made in this field, it is just the first step for constructing human-level language understanding systems. Because we still do not exactly know what language understanding is, the content of possible questions is limited to what we can imagine. In the future, we have to design a task that *more scientifically* evaluates an agent's capability of language understanding *with sufficient explanation* using the heritage from linguistics, philosophy, psychology, and cognitive science.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas, November 2016. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D16-1203`.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics, June 2010. URL `https://aclweb.org/anthology/W10-1001`.

Daniel Andor, Luheng He, Kenton Lee, and Emily Pitler. Giving BERT a calculator: Finding operations and arguments with reading comprehension. 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 5946–5951, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1609. URL `https://aclweb.org/anthology/D19-1609`.

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

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `https://aclweb.org/anthology/W05-0909`.

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008. doi: 10.1162/coli.2008.34.1.1. URL `https://aclweb.org/anthology/J08-1001`.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009. URL `https://doi.org/10.1145/1553374.1553380`.

Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010a. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The sixth pascal recognizing textual entailment challenge. In *Proceedings of the Third Text Analysis Conference (TAC 2010)*, 2010b.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The seventh pascal recognizing textual entailment challenge. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*, 2011.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1510. Association for Computational Linguistics, 2014. URL `https://aclweb.org/anthology/D14-1159`.

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

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. 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, July 2018.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 725–734, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/bouthillier19a.html`.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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, 2015. doi: 10.18653/v1/D15-1075. URL `https://aclweb.org/anthology/D15-1075`.

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

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

Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2367. Association for Computational Linguistics, August 2016. URL `https://aclweb.org/anthology/P16-1223`.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 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, 2017. doi: 10.18653/v1/P17-1171. URL `https://aclweb.org/anthology/P17-1171`.

Jifan Chen and Greg Durrett. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1405. URL `https://aclweb.org/anthology/N19-1405`.

Jianpeng Cheng, Siva Reddy, Vijay Saraswat, and Mirella Lapata. Learning structured natural language representations for semantic parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44–55. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1005. URL `https://aclweb.org/anothology/P17-1005`.

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. Coarse-to-fine question answering for long documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–220. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1020. URL `https://aclweb.org/anthology/P17-1020`.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL `https://aclweb.org/anthology/D18-1241`.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL `https://aclweb.org/anthology/N19-1300`.

Herbert H. Clark. Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174. Association for Computational Linguistics, 1975.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018.

Robin Cooper, Richard Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspers, Hans Kamp, Manfred Pinkal, Massimo Poesio, Stephen Pulman, et al. FraCaS: A framework for computational semantics. *Deliverable*, 8:62–051, 1994.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. Using the framework. Technical report, 1996.

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

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220, 2013.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL `https://aclweb.org/anthology/P19-1285`.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 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, 2018. URL `https://aclweb.org/anthology/N18-1144`.

Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, and Andrew McCallum. Building dynamic knowledge graphs from text using machine reading comprehension. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=S1lhbnRqF7`.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 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, November 2019. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D19-1606`.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclweb.org/anthology/N19-1423`.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846. Association for Computational Linguistics, 2017a. doi: 10.18653/v1/P17-1168. URL `https://aclweb.org/anthology/P17-1168`.

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

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. Cognitive graph for multi-hop reading comprehension at scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2694–2703, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1259. URL `https://aclweb.org/anthology/P19-1259`.

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

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 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, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL `https://aclweb.org/anthology/P18-1128`.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL `https://aclweb.org/anthology/N19-1246`.

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

Oren Etzioni, Michele Banko, and Michael J Cafarella. Machine reading. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, volume 6, pages 1517–1519, 2006.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1156–1165. ACM, 2014.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 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, 2018. URL `https://aclweb.org/anthology/D18-1407`.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5801. URL `https://aclweb.org/anthology/D19-5801`.

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

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*, pages 340–348. Association for Computational Linguistics, 2010.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 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, 2018.

Arthur C. Graesser, Murray Singer, and Tom Trabasso. Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371, 1994. URL `https://doi.org/10.1037/0033-295X.101.3.371`.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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, 2018. URL `https://aclweb.org/anthology/N18-2017`.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1175. URL `https://aclweb.org/anthology/N18-1175`.

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. 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, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2605. URL https://aclweb.org/anthology/W18-2605.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 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., 2015. URL http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf.

José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48(3):397–447, 2017a.

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

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 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, August 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/P16-1145.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 928–939, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://aclweb.org/anthology/C14-1088.

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

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 325–332. Association for Computational Linguistics, 1999. URL https://doi.org/10.3115/1034678.1034731.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Pro-*

cessing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclweb.org/anthology/D19-1243.

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

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 633–644, Doha, Qatar, October 2014. Association for Computational Linguistics. URL https://aclweb.org/anthology/D14-1070.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 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, December 2016. The COLING 2016 Organizing Committee. URL https://aclweb.org/anthology/C16-1278.

Robin Jia and Percy Liang. 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, September 2017.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclweb.org/anthology/D19-1259.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 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, 2017. doi: 10.18653/v1/P17-1147. URL https://aclweb.org/anthology/P17-1147.

Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning, 2018. URL https://openreview.net/forum?id=SyunbfbAb.

Divyansh Kaushik and Zachary C. Lipton. 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, 2018. URL `https://aclweb.org/anthology/D18-1546`.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294. Association for Computational Linguistics, July 2018. URL `https://aclweb.org/anthology/P18-1027`.

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

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 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, 2018a.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. Question answering as global reasoning over semantic abstractions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1905–1914, 2018b.

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. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-1026. URL `https://aclweb.org/anthology/S19-1026`.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

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

Walter Kintsch. Information accretion and reduction in text processing: Inferences. *Discourse processes*, 16(1-2):193–202, 1993.

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

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. ISSN 2307-387X. URL `https://transacl.org/ojs/index.php/tacl/article/view/1197`.

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. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, March 2019. doi: 10.1162/tacl_a_00276. URL `https://aclweb.org/anthology/Q19-1026`.

Igor Labutov, Bishan Yang, Anusha Prakash, and Amos Azaria. 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, 2018. URL `https://aclweb.org/anthology/P18-1077`.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 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, September 2017.

John Lalor, Hao Wu, and Hong Yu. 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, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1062. URL `https://aclweb.org/anthology/D16-1062`.

Hector J. Levesque. 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*, 2011.

Hector J. Levesque. On our best behaviour. *Artificial Intelligence*, 212:27 – 35, 2014. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/j.artint.2014.03.007. URL `http://www.sciencedirect.com/science/article/pii/S0004370214000356`.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 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, June 2019a.

Association for Computational Linguistics. doi: 10.18653/v1/N19-1112. URL https://aclweb.org/anthology/N19-1112.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 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, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/N19-1225. URL https://aclweb.org/anthology/N19-1225.

Peter LoBue and Alexander Yates. 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, June 2011. Association for Computational Linguistics. URL https://aclweb.org/anthology/P11-2057.

Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics, 2002.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1103. URL https://aclweb.org/anthology/P17-1103.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1185. URL https://aclweb.org/anthology/N18-1185.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. Association for Computational Linguistics, 2014. doi: 10.3115/v1/P14-5010. URL https://aclweb.org/anthology/P14-5010.

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

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics, October-November 2018. URL https://aclweb.org/anthology/D18-1151.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclweb.org/anthology/P19-1334.

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

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

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclweb.org/anthology/D18-1260.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 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, 2018. URL https://aclweb.org/anthology/P18-1160.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 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, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1416. URL https://aclweb.org/anthology/P19-1416.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1613. URL https://aclweb.org/anthology/P19-1613.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/N16-1098.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 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, 2017. URL https://aclweb.org/anthology/W17-0906.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 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, July 2018. URL https://aclweb.org/anthology/P18-1176.

Erik T. Mueller. Story understanding through multi-representation model construction. In *Proceedings of the HLT-NAACL 2003 workshop on Text meaning*, pages 46–53. Association for Computational Linguistics, 2003.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclweb.org/anthology/C18-1198.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268, 2016.

Yixin Nie, Yicheng Wang, and Mohit Bansal. Analyzing compositionality-sensitivity of NLI models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 6867–6874, 2019.

Peter Norvig. Marker passing as a weak method for text inferencing. *Cognitive Science*, 13(4):569–620, 1989. URL https://doi.org/10.1207/s15516709cog1304_4.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 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, November 2016. URL https://aclweb.org/anthology/D16-1241.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 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), may 2018.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-1012. URL https://aclweb.org/anthology/S19-1012.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1258. URL `https://aclweb.org/anthology/D18-1258`.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. URL `https://doi.org/10.1109/TKDE.2009.191`.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 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, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1144. URL `https://aclweb.org/anthology/P16-1144`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclweb.org/anthology/P02-1040`.

Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics, October 2008. URL `https://aclweb.org/anthology/D08-1020`.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1487. URL `https://aclweb.org/anthology/P19-1487`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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, November 2016. URL `https://aclweb.org/anthology/D16-1264`.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 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, July 2018. URL `https://aclweb.org/anthology/P18-2124`.

Abigail Razon and John Barnden. A new approach to automated text readability classification based on concept indexing with integrated part-of-speech n-gram features. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 521–528, September 2015. URL `https://aclweb.org/anthology/R15-1068`.

François Recanati. *Literal meaning*. Cambridge University Press, 2004.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, March 2019. doi: 10.1162/tacl_a_00266. URL `https://aclweb.org/anthology/Q19-1016`.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 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, 2013. URL `https://aclweb.org/anthology/D13-1020`.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics, September 2015. URL `https://aclweb.org/anthology/D15-1044`.

Mrinmaya Sachan, Kumar Dubey, Eric Xing, and Matthew Richardson. Learning answer-entailing structures for machine comprehension. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 239–249. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1024. URL `https://aclweb.org/anthology/P15-1024`.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1233. URL `https://aclweb.org/anthology/D18-1233`.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 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, July 2018. URL `https://aclweb.org/anthology/P18-1156`.

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

Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. "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, July 2010. URL `https://aclweb.org/anthology/P10-1122`.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1004. URL `https://aclweb.org/anthology/P19-1004`.

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

Roger C. Schank and Robert P. Abelson. *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum, 1977.

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

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. Story cloze task: UW NLP system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 52–55, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://aclweb.org/anthology/W17-0907`.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*, 2017.

Ellery Smith, Nicola Greco, Matko Bosnjak, and Andreas Vlachos. A strong lexical matching method for the machine comprehension test. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1693–1698. Association for Computational Linguistics, 2015.

Saku Sugawara and Akiko Aizawa. An analysis of prerequisite skills for reading comprehension. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 1–5. Association for Computational Linguistics, November 2016. URL `https://aclweb.org/anthology/W16-6001`.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 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, 2017a. doi: 10.18653/v1/P17-1075. URL `https://aclweb.org/anthology/P17-1075`.

Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. Prerequisite skills for reading comprehension: Multi-perspective analysis of mctest datasets and systems. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3089–3096, 2017b.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 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, 2018. URL `https://aclweb.org/anthology/D18-1453`.

Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. Requirements for explainable machine reading comprehension: A position paper. Manuscript, 2020a.

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

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-2034. URL `https://aclweb.org/anthology/P17-2034`.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, March 2019a. doi: 10.1162/tacl_a_00264. URL `https://aclweb.org/anthology/Q19-1014`.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Improving machine reading comprehension with general reading strategies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2019b.

Simon Suster and Walter Daelemans. 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, 2018. URL `https://aclweb.org/anthology/N18-1140`.

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

Alon Talmor and Jonathan Berant. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1485. URL https://aclweb.org/anthology/P19-1485.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclweb.org/anthology/N19-1421.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding stories in movies through question-answering. In *the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-2623. URL https://aclweb.org/anthology/W17-2623.

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

Sowmya Vajjala and Detmar Meurers. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. Association for Computational Linguistics, June 2012. URL https://aclweb.org/anthology/W12-2019.

Teun Adrianus van Dijk and Walter Kintsch. *Strategies of discourse comprehension*. Academic Press New York, 1983.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Ellen M. Voorhees and Dawn M. Tice. The TREC-8 question answering track evaluation. In *the Eighth Text Retrieval Conference (TREC-8)*, 1999.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 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, November 2019.

Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL https://aclweb.org/anthology/D19-1221.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32*, pages 3261–3275. Curran Associates, Inc., 2019.

Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. Machine comprehension with syntax, frames, and semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 700–706, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2115. URL https://aclweb.org/anthology/P15-2115.

Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. What is the Jeopardy model? a quasi-synchronous grammar for QA. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 22–32. Association for Computational Linguistics, June 2007. URL https://aclweb.org/anthology/D/D07/D07-1003.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL https://aclweb.org/anthology/K17-1028.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018. ISSN 2307-387X. URL https://transacl.org/ojs/index.php/tacl/article/view/1325.

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

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 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, November 2017. Asian Federation of Natural Language Processing. URL https://aclweb.org/anthology/I17-1100.

Shimon Whiteson, Brian Tanner, Matthew E Taylor, and Peter Stone. Protecting against evaluation overfitting in empirical reinforcement learning. In *Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2011 IEEE Symposium on*, pages 120–127. IEEE, 2011.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1257. URL https://aclweb.org/anthology/D18-1257.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1166. URL https://aclweb.org/anthology/D18-1166.

Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclweb.org/anthology/D15-1237.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 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, 2018. URL https://aclweb.org/anthology/D18-1259.

Wenpeng Yin, Sebastian Ebert, and Hinrich Schütze. Attention-based convolutional neural network for machine comprehension. In *Proceedings of the Workshop on Human-Computer Question Answering*, pages 15–21. Association for Computational Linguistics, June 2016. URL https://aclweb.org/anthology/W16-0103.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. QANet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B14TlG-RW.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 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, 2018. URL https://aclweb.org/anthology/D18-1009.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 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, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL `https://aclweb.org/anthology/P19-1472`.

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

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

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