Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability

Saku Sugawara[♠], **Yusuke Kido**[♠], **Hikaru Yokono**[♣], and **Akiko Aizawa**[◇♠]
[♠]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
[♣]Fujitsu Laboratories Ltd., 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Japan
[◇]Natural Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
sakus@is.s.u-tokyo.ac.jp mail@yusuk.eki.do
yokono.hikaru@jp.fujitsu.com aizawa@nii.ac.jp

Abstract

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

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 a reading comprehension (RC) task that requires the agent to read open-domain documents and answer questions about them. Constructing systems with RC competence is challenging because RC 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 RC 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 RC 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 RC systems. In such situations, it is difficult to obtain an accurate assessment of the RC system.

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

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 RC questions in Figure 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 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.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 806–817
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-1075

Furthermore, the accompanying categories of existing RC datasets cannot help with the analysis of this issue.

In this study, our goal is to investigate how these two types of difficulty, namely "answering questions" and "reading text," are correlated in RC. Corresponding to each type, we formalize two classes of evaluation metrics, *prerequisite skills* and *readability*, and analyze existing RC datasets. Our intention is to provide the basis of an evaluation methodology of RC 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. (2017) proposed for the fine-grained analysis of RC capability. Their study also presented an important observation of the relation between the difficulty of an RC task 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 work, 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 paper are as follows.

1. We adopt two classes of evaluation metrics to show the qualitative features of RC datasets. Through analyses of RC datasets, we demonstrate that there is only a weak correlation between the difficulty of questions and the readability of context texts in RC datasets.

2. We revise a previous classification of prerequisite skills for RC. 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 RC datasets, compared to the two datasets considered in Sugawara and Aizawa (2016), with our organized metrics being used in the comparison. We have made the results publicly available[1] and report on the characteristics of the datasets and the differences between them.

We should note that, in this study, RC 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 RC questions as *contextual entailments*.

The remainder of this paper is divided into the following sections. First, we discuss related work in Section 2. Next, we specify our two classes of metrics in Section 3. In Section 4, we annotate existing RC datasets with the prerequisite skills. Section 5 gives the results of our dataset analysis and Section 6 discusses their implications. Section 7 presents our conclusions.

2 Related Work

2.1 Reading Comprehension Datasets

In this section, we present a short history of RC datasets. To our knowledge, Hirschman et al. (1999) were the first to use NLP methods for RC. 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 et al., 1999) and

[1] http://www-al.nii.ac.jp/rc_dataset_analysis

textual entailment (Bentivogli et al., 2010; Sammons et al., 2010; Dagan et al., 2006). One of the RC tasks of the time was QA4MRE (Sutcliffe et al., 2013). The highest accuracy achieved for this task 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.

This means that an important issue for designing RC 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 RC 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 Figure 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.

2.2 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.[2]

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).

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 Evaluation Metrics for Datasets

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

For the prerequisite skills class (Section 3.1), we refined RC skills that were proposed by Sugawara et al. (2017) and Sugawara and Aizawa (2016). However, a problem in these studies is that their categorization of knowledge reasoning

[2] Note that this is a very simplified overview.

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 2.2, 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 3.2) 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).

3.1 Prerequisite Skills

Based on the 10 RC skills in Sugawara et al. (2017), 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.

In addition, we propose the following four categories by refining the “commonsense reasoning” category proposed originally in Sugawara et al. (2017).

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. (2017).

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 RC. We were motivated by the discussion in Smith et al. (2015).

Whereas the above 11 skills involve multiple items, the final pair of skills involve only a single sentence.

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 we did not construct this classification to be dependent on particular RC 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 RC systems.

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.[3]

We list the readability metrics in Table 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*). We applied these metrics only to sentences that needed to be read in answering questions.

- Ave. no. of characters per word (*NumChar*)
- Ave. no. of syllables per word (*NumSyll*)
- Ave. sentence length in words (*MLS*)
- Proportion of words in AWL (*AWL*)
- Modifier variation (*ModVar*)
- No. of coordinate phrases per sentence (*CoOrd*)
- Coleman–Liau index (*Coleman*)
- Dependent clause-to-clause ratio (*DC/C*)
- Complex nominals per clause (*CN/C*)
- Adverb variation (*AdvVar*)

Table 1: Readability metrics. *AWL* refers to the Academic Word List.[4]

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

4 Annotation of Reading Comprehension Datasets

We annotated six existing RC datasets with the prerequisite skills. We explain the annotation procedure in Section 4.1 and the annotated RC datasets in Section 4.2.

4.1 Annotation Procedure

We prepared annotation guidelines according to Sugawara et al. (2017). 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 RC 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 an-

[3] 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.

[4] http://en.wikipedia.org/wiki/Academic_Word_List

RC dataset	Genre	Query sourcing	Task formulation
QA4MRE (2013)	Technical documents	Handcrafted by experts	Multiple choice
MCTest (2013)	Narratives by crowd workers	Crowdsourced	Multiple choice
SQuAD (2016)	Wikipedia articles	Crowdsourced	Text span selection
Who-did-What (2016)	News articles	Automated	Cloze
MS MARCO (2016)	Segmented web pages	Search engine queries	Description
NewsQA (2016)	News articles	Crowdsourced	Text span selection

Table 2: Analyzed RC datasets, their genres, query sourcing methods, and task formulations.

notators 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.

4.2 Datasets

As summarized in Table 2, the annotation was performed on six existing RC 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., 2016). 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 RC 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 RC datasets. (We explain the method of choosing questions for the annotation in Appendix A.)

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 RC-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: Frequencies (%) of prerequisite skills needed for the RC datasets.

#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	1.19	1.99

Table 4: Frequencies (%) of the number of required prerequisite skills for the RC datasets.

5 Results of the Dataset Analysis

We now present the results of evaluating the RC 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 RC dataset; (ii) number of prerequisite skills required per question; (iii) readability metrics for each RC dataset; and (iv) correlation between readability metrics and the number of required prerequisite skills.

(i) Frequencies of prerequisite skills (see Table 3): QA4MRE had the highest scores for frequencies among the datasets. This seems to reflect

Metrics	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
NumChar	5.026	3.892	**5.378**	4.988	5.016	5.017
NumSyll	1.663	1.250	**1.791**	1.657	1.698	1.635
MLS	28.488	11.858	23.479	**29.146**	19.634	22.933
AWL	0.067	0.003	**0.071**	0.033	0.047	0.038
ModVar	0.174	0.114	**0.188**	0.150	0.186	0.138
CoOrd	**0.922**	0.309	0.722	0.467	0.651	0.507
Coleman	12.553	4.333	**14.095**	12.398	11.836	12.138
DC/C	**0.343**	0.223	0.243	0.254	0.220	0.264
CN/C	1.948	0.614	1.887	**2.310**	1.935	1.702
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 5: Results of readability metrics for the RC 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.

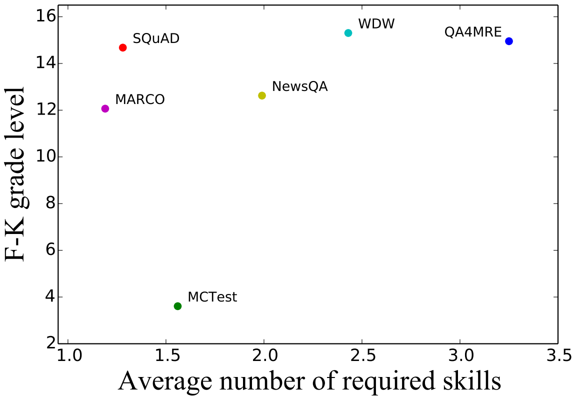

Figure 2: Flesch–Kincaid grade levels and average number of required prerequisite skills for the RC datasets.

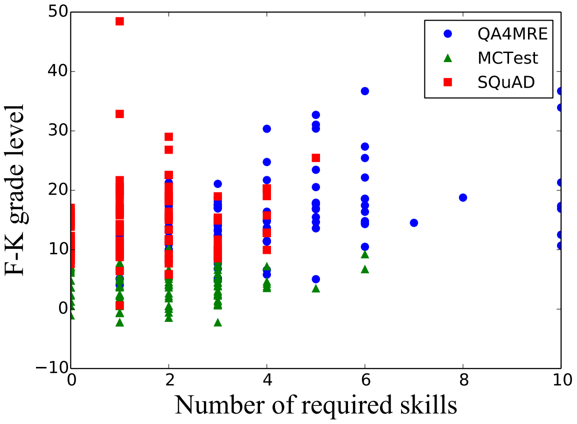

Figure 3: Flesch–Kincaid grade levels and number of required prerequisite skills for all questions in the selected RC datasets.

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 (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 filtering (described in Appendix A), 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 (see Table 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 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 (see Table 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 (see Figures 2 and 3, and Table 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.

Metrics	r	p	Metrics	r	p
NumChar	0.068	0.095	CoOrd	0.166	0.000
NumSyll	0.057	0.161	Coleman	0.140	0.001
MLS	0.416	0.000	DC/C	0.188	0.000
AWL	0.114	0.005	CN/C	0.131	0.001
ModVar	0.025	0.545	AdvVar	0.026	0.515
F–K	0.343	0.000	Words	0.355	0.000

Table 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 RC datasets.

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 2 and 3 show the relation between two values for each dataset and for each question, respectively. Figure 2 shows the trends of the datasets. QA4MRE 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. Three separate domains can be seen.

Table 6 presents Pearson's correlation coefficients between the number of required prerequisite skills and each readability metric for all questions in the RC 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 RC datasets does not directly affect the difficulty of their questions. That is, RC 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 RC 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 RC 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.

6 Discussion

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

How to utilize the two classes of metrics for system development: one possible scenario for developing an RC 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 RC 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, RC 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 RC. In addition, annotators 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 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.[4] The first row of Table 7 gives the average number of required sentences per question for each RC 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.

Sentence	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
Number	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 7: Average number and distance apart of sentences that need to be read to answer a question in the RC datasets.

Metrics of RC 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 RC task design, and one that has never been explicitly considered in conventional NLP tasks. To date, the difference between human and machine RC 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 RC datasets and improved RC systems.

7 Conclusion

In this study, we adopted evaluation metrics that comprise two classes, namely refined *prerequisite skills* and *readability*, for analyzing the quality of RC datasets. We applied these classes to six existing datasets and highlighted their characteristics according to each metric. Our dataset analysis suggests that the readability of RC datasets does not directly affect the difficulty of the questions and that it is possible to create an RC dataset that is easy to read but difficult to answer. In future work, we plan to use the analysis from the present study in constructing a system that can be applied to multiple datasets.

Acknowledgments

We would like to thank anonymous reviewers for their insightful comments. This work was supported by JSPS KAKENHI Grant Numbers 15H02754 and 16K16120.

[4] 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.

References

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

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

Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. 2010. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Citeseer.

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

Herbert H Clark. 1975. Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*. Association for Computational Linguistics, pages 169–174. https://doi.org/10.3115/980190.980237.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, Springer, pages 177–190. https://doi.org/10.1007/11736790_9.

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

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

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*. pages 1693–1701.

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

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

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Chief of Naval Technical Training, Research Branch Report 8-75.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review* 95(2):163. https://doi.org/10.1037/0033-295X.95.2.163.

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

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 329–334. https://aclweb.org/anthology/P11-2057.

Danielle S McNamara and Joe Magliano. 2009. Toward a comprehensive model of comprehension. *Psychology of learning and motivation* 51:297–384. https://doi.org/10.1016/S0079-7421(09)51009-2.

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

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

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did What: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2230–2235. https://aclweb.org/anthology/D16-1241.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22(10):1345–1359. 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 Fernandez. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1525–1534. https://aclweb.org/anthology/P16-1144.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2383–2392. https://aclweb.org/anthology/D16-1264.

Abigail Razon and John Barnden. 2015. 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. https://aclweb.org/anthology/R15-1068.

Matthew Richardson, J.C. Christopher Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 193–203. http://aclweb.org/anthology/D13-1020.

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

Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. 2010. “ask not what textual entailment can do for you...”. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1199–1208. https://aclweb.org/anthology/P10-1122.

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

Saku Sugawara and Akiko Aizawa. 2016. 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*. Association for Computational Linguistics, pages 1–5. https://aclweb.org/anthology/W16-6001.

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

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

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. NewsQA: A machine comprehension dataset. *CoRR* abs/1611.09830.

Sowmya Vajjala and Detmar Meurers. 2012. 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*. Association for Computational Linguistics, pages 163–173. https://aclweb.org/anthology/W12-2019.

Teun Adrianus van Dijk and Walter Kintsch. 1983. *Strategies of discourse comprehension*. Citeseer.

Ellen M Voorhees et al. 1999. The TREC-8 question answering track report. In *TREC*. volume 99, pages 77–82.

Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. 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*. Association for Computational Linguistics, pages 22–32. https://aclweb.org/anthology/D/D07/D07-1003.

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

A Sampling Methods for Questions

In this appendix, we explain the method of choosing questions for annotation.

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.

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., 2016): 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*.