Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

# Prerequisite Skills for Reading Comprehension: Multi-Perspective Analysis of MCTest Datasets and Systems

**Saku Sugawara**
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, Japan
sakus@is.s.u-tokyo.ac.jp

**Hikaru Yokono**
Fujitsu Laboratories Ltd.
4-1-1, Kamikodanaka, Nakahara-ku
Kawasaki, Kanagawa, Japan
yokono.hikaru@jp.fujitsu.com

**Akiko Aizawa**
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan
aizawa@nii.ac.jp

**Abstract**

One of the main goals of natural language processing (NLP) is synthetic understanding of natural language documents, especially reading comprehension (RC). An obstacle to the further development of RC systems is the absence of a synthetic methodology to analyze their performance. It is difficult to examine the performance of systems based solely on their results for tasks because the process of natural language understanding is complex. In order to tackle this problem, we propose in this paper a methodology inspired by unit testing in software engineering that enables the examination of RC systems from multiple aspects. Our methodology consists of three steps. First, we define a set of prerequisite skills for RC based on existing NLP tasks. We assume that RC capability can be divided into these skills. Second, we manually annotate a dataset for an RC task with information regarding the skills needed to answer each question. Finally, we analyze the performance of RC systems for each skill based on the annotation. The last two steps highlight two aspects: the characteristics of the dataset, and the weaknesses in and differences among RC systems. We tested the effectiveness of our methodology by annotating the Machine Comprehension Test (MCTest) dataset and analyzing four existing systems (including a neural system) on it. The results of the annotations showed that answering questions requires a combination of skills, and clarified the kinds of capabilities that systems need to understand natural language. We conclude that the set of prerequisite skills we define are promising for the decomposition and analysis of RC.

## 1 Introduction

Reading comprehension (RC) is "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (Snow 2002). RC tasks require understanding natural language texts and answering questions about them. In natural language processing (NLP), they are used as a method to evaluate natural language understanding systems. Such tasks are challenging because they involve a variety of activities, such as coreference resolution, discourse understanding, and commonsense reasoning.

In the development of RC systems, it is important to identify what the systems can and cannot understand. However, a critical problem in this regard is that the process of natural language understanding is so complicated that it is difficult to examine the performance of RC systems based on the results of tasks. Neural systems, in particular, are liable to strengthen this trend due to their black-box architectures. Therefore, an effective methodology is necessary to analyze the performance of RC systems for their development.



> Context:
> James the Turtle was always getting in trouble. One day, James went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries.
>
> Q1: What is the name of the trouble making turtle?
> A) Pudding, B) Jane, *C) James, D) Fries
> Q2: Where did James go after he went to the grocery store?
> A) His freezer, *B) A fast food restaurant, C) His room, D) His desk

Figure 1: An example of reading comprehension tasks excerpted from MCTest (Richardson et al., 2013) with the answers (*).

Consider the example of the questions shown in Figure 2. To solve Q1, an RC system must identify apposition (*James* is *the Turtle*) and recognize *James* as a name. Similarly, Q2 requires an understanding of temporal relations among the events (*went to the grocery store* → *walked to the fast food restaurant*), and the understanding that the verb (*walk to* means *go to*). These instances show that different types of skills are required in RC tasks.

In this study, our goal is to establish a general methodology to assess the performance of RC systems. Our methodology involves decomposing the capability of RC and analyzing its performance from multiple points of view. This is in stark contrast to simple accuracy-based analysis, the approach used to assess systems at present. Our methodology consists of the following three steps:

1. Define a set of basic prerequisite skills required to understand documents.
2. Annotate questions for an RC task using the defined skills.
3. Analyze the performance of RC systems on annotated questions to understand the differences among systems and the limitations of each.

We first define a set of basic skills in Section 3. Then, in order to exemplify our methodology, we annotate two datasets (MC160 and MC500) of the Machine Comprehension Test (MCTest) (Richardson, Burges, and Renshaw 2013). Section 4 describes the datasets and the annotation scheme. The results of the annotation are presented in Section 5. This section also contains an analysis of the performance of RC systems based on a set of the skills. We analyze four systems proposed by Richardson, Burges, and Renshaw (2013), Smith et al. (2015), and Yin, Ebert, and Schütze (2016). Finally, Sections 6 and 7 provide a discussion of the results and our conclusions for future research in this area.

Our contributions are threefold[1]:

1. We propose a set of prerequisite skills for RC. It provides an in-depth overview of RC and enables comprehensive analysis.
2. We annotate an RC task using these skills, which enables us to understand the characteristics of the task. This not only highlights the differences among RC tasks already proposed, but also helps us plan and design new tasks.
3. We analyze the performance of RC systems, including a neural system, and compare them from multiple aspects using the annotation. This helps us understand how well RC systems work for the task, and provides insight into directions of research to pursue further.

## 2 Related Work

### Existing Tasks for Natural Language Understanding

Existing tasks on natural language understanding are categorized into two groups. The first group focuses on specific abilities; examples include Recognizing Textual Entailment (RTE) (Dagan, Glickman, and Magnini 2006) for textual entailment, the CoNLL 2012 shared task (CoNLL2012st) (Pradhan et al. 2012) for coreference resolution, Choice of Plausible Alternatives (COPA) (Roemmele, Bejan, and Gordon 2011), the Winograd Schema Challenge (WSC) (Levesque, Davis, and Morgenstern 2011) for commonsense reasoning, the Aristo Challenge (Clark 2015) for elementary-level science and mathematics, and Shallow Discourse Parsing (SDP) (CoNLL 2015 shared task) (Xue et al. 2015) for discourse relations. For some of these tasks, high-performance models have been already proposed. However, they are mostly designed as both domain-dependent and task-specific methods, which prevents their extension to other tasks.

The second group focuses on synthetic understanding of documents, reading comprehension (RC), which was initiated by Hirschman et al. (1999). One type of RC tasks is the Cloze test. For instance, the DeepMind Q&A Dataset (DMQA) (Hermann et al. 2015) and the Children's Book Test (CBT) (Hill et al. 2016) involve filling in the blanks in a summary or explanatory sentences, where the correct answer appears in the contextual document. For these tasks, neural network-based methods (e.g., Chen, Bolton, and Manning (2016) in the DMQA task) have been reported to achieve good results. More recently, Paperno et al. (2016) proposed the LAMBADA dataset that requires contextual understanding, but excludes questions that can be solved using n-gram language models.

Another type of RC tasks involves answering multiple-choice questions or extracting a word sequence (a text span) in passages. For example, the Machine Comprehension Test (MCTest) (Richardson, Burges, and Renshaw 2013), ProcessBank Berant et al. (2014) and Question Answering for Machine Reading Evaluation (QA4MRE) (Sutcliffe et al. 2013) are multiple-choice question-based tasks. Moreover, Rajpurkar et al. (2016) proposed SQuAD, an RC dataset consisting of more than 100,000 questions, where the task was to extract the correct word sequence of documents curated from Wikipedia.

Several methods have been proposed for these tasks. Their performance, however, is still suboptimal: e.g., Trischler et al. (2016) yielded an accuracy of 77.5% on MCTest, which was poor in comparison with human performance (>95%). This is because in this style, systems need to first understand the content of questions, and gather clues to choose the correct option (or the correct word sequence). This operation is more difficult than the Cloze test, which simply requires filling in blanks.

### Analytic Approaches to Reading Comprehension

One of the most important aspects of the RC tasks listed above is the difficulty in identifying the reason why models cannot generate correct answers to questions. RC involves various operations, such as coreference resolution, understanding discourse relations, commonsense reasoning, and so on. However, existing tasks are not sufficiently simplified to separately assess the performance of the process.

Here, we review existing research contributing to the construction of an analytic evaluation methodology for RC.

Smith et al. (2015) proposed a system that can deal with *wh-* question types as well as their contents: negation, temporal relation, numbers, narratives of stories, and quantifiers. They found questions with rules that matched certain words pertaining to their respective types, and analyzed the results based on these rules. This approach seems to be effective for upgrading RC systems. We generalize the concept of question types based on skills required in natural language understanding: from a syntactic view to a semantic view.

With regard to other RC tasks, Rajpurkar et al. (2016), for instance, analyzed their dataset using several types of reasoning, e.g., lexical variation, syntactic variation, and multiple sentence reasoning. Although the purpose of their analysis was similar to that of ours here, the reasoning types they listed were too coarse to analyze reasoning performance for multiple sentences. In order to assess the contextual understanding required for RC, the method of evaluation should be able to analyze the performance of such reasoning from multiple points of view.

[1]The preliminary results of this study were presented in the extended abstract at Uphill Battles in Language Processing Workshop (Sugawara and Aizawa 2016).

| RC skills | Descriptions or examples | Major tasks |
|---|---|---|
| List/Enumeration | Tracking, retaining, and list/enumeration of entities or states | bAbI |
| Mathematical operations | Four arithmetic operations and geometric comprehension | Aristo |
| Coreference resolution | Detection and resolution of coreference | CoNLL2012st |
| Logical reasoning | Induction, deduction, conditional statement, and quantifier | Aristo, FraCaS |
| Analogy | Trope in figures of speech, e.g., metaphor | - |
| Spatiotemporal relations* | Spatial and/or temporal relations of events | SDP, bAbI |
| Causal relations* | Relations of events expressed by why, because, the reason for, and so on | COPA, SDP |
| Commonsense reasoning | Taxonomic knowledge, qualitative knowledge, action, and event changes | COPA, WSC |
| Schematic/Rhetorical clause relations* | Coordination or subordination of clauses in a sentence | SDP |
| Special sentence structure* | Scheme in figures of speech, constructions, and punctuation marks in a sentence | - |

Table 1: Reading comprehension skills, their descriptions, and major tasks to which each skill pertains (from Section 2 and 3). The asterisks (*) with items represent "understanding of."

Sammons, Vydiswaran, and Roth (2010) proposed a methodology for robust development to recognize textual entailment. LoBue and Yates (2011) followed this and classified types of commonsense knowledge. These analytic methodologies are also available for RC, and not only for textual entailment.

Furthermore, our approach is partly inspired by the bAbI tasks (Weston et al. 2015). These tasks consist of simplified questions that test natural language understanding according to 20 skills, such as recognizing relations among sentences, counting objects, and temporal reasoning. Unfortunately, the dataset of bAbI has poor vocabulary and minor grammatical elements; hence, there is room for improvement. Nonetheless, we think that Weston et al. (2015)'s idea of using prerequisite skills is promising, that is, useful for analyzing the capability of natural language understanding. In this study, we adopt their idea, not in terms of the task formulation, but annotations of tasks and analyses of systems.

## 3 Reading Comprehension Skills

We investigate prevalent tasks for natural language understanding and define a set of prerequisite basic skills. Hereafter, we refer to these skills as **reading comprehension skills (RC skills)**, as shown in the first column of Table 1.

In this study, we define RC skills as abilities to understand the relations among multiple clauses[2]. Here, we assume that when RC systems use an RC skill, they already have the capability to recognize the facts described in the clauses that are related to the skill. This assumption is derived from the observation that reading comprehension is composed of two steps: understanding the content of each clause in the target passage, and combining content. We also assume that systems already understand semantic roles in a clause.

For example, the skill of understanding temporal relations between events implicitly requires the recognition of expressions such as conjunctions (*when, as, since*, ...), time indexicals (*morning*, *evening*, ...), and tense and aspects (*went*, *is going*, *will go*, ...). When these expressions have a relation with another clause, this skill is required.

[2]More precisely, we assume that a clause has a subject-verb pair, and represents a certain event or state.

We define the following 10 RC skills:

**List/Enumeration** target tracking, retention, and list/enumeration of multiple entities, states, and facts at the same time. This skill implicitly requires memorizing or storing objects and recalling them. For another task concerning with tracing entities, see *Lists/sets* tasks in the bAbI tasks (Weston et al. 2015).

**Mathematical operations** consist of four arithmetic operations, more advanced mathematics, and geometric comprehension as in Clark, Harrison, and Balasubramanian (2013) and Clark (2015)[3].

**Coreference resolution** is the detection and resolution of all possible demonstratives, i.e., reference terms that have a relation with another word/phrase in the given context. For task formulation and configuration of coreference resolution, see Pradhan et al. (2011).

**Logical reasoning** is defined for the derivation of new facts from relations between statements in the given context (including defeasible reasoning). We do not doubt that other relational skills include some kind of logical reasoning. For example, the transitive rule is necessary to understand the relations among three timepoints (if both *X occurred after Y* and *Y occurred after Z*, *X occurred after Z*). For clarification, however, we exclude such simple cases, and we focus on reasoning involving induction, deduction, conditional statements, and quantifiers as in Cooper et al. (1994).

**Analogy** focuses on the recognition of linguistic expressions corresponding to a cognitive process of transferring information or meaning from a particular fact to another, e.g., simile and metaphor. Considering that skills for understanding figures of speech are necessary for reading, we take two factors into account: "trope" and "scheme," the latter of which is the last skill in our list.

**Spatiotemporal relations** involve the skill to understand spatial or temporal relations among facts in the given context, e.g., a time series of events, locations, and any expres-

[3]In our definition, a list/enumeration without mathematical operations is possible. An enumeration requires incremental "calls" of natural numbers ("one," "two," "three," ...). This is easier than calculation that uses operations on multiple numbers. Therefore, we define such operations as an independent skill.

sion related to them (*before*, *after*, *in front of*, ...). To see simplified tasks in this context, refer to *Time reasoning* and *Positional reasoning* in the bAbI tasks.

**Causal relations** involve the skill to understand causal relations. We suppose that this skill is needed only if systems need to recognize expressions corresponding to causality, e.g., *why, because*, and *the reason for*.

**Commonsense reasoning** is performed using taxonomic knowledge, qualitative knowledge, action, and event change, which we adopt from Davis and Marcus (2015). Moreover, we add arbitrary knowledge to the scope of this skill, except grammatical knowledge and basic vocabulary pertaining to other skills.

**Schematic/Rhetorical clause relations** target the understanding of relations among schematic or rhetorical clauses in a sentence, e.g., coordinating or subordinating clauses introduced by conjunctions such as *and*, *or*, *that*, and *although*. As listed above, we handle the understanding of conditional, spatiotemporal, and causal clauses as independent skills that can be regarded as holding the content themselves.

**Special sentence structure** is required to recognize linguistic symbols or structures in sentences and introduce their interpretation as a new fact. As explained above, this skill is intended for the understanding of "scheme" in figures of speech, which changes the normal arrangement of words in the structure of a sentence, e.g., apposition, ellipsis, and transposition. This skill also targets linguistic constructions ("the more ..., the more", ...) and punctuation marks (";", "—", the quotation itself, ...) in reference to Huddleston and Pullum (2002).

Note that the last two skills are exceptional: they are required in a single sentence, while the first eight skills mainly target relations among sentences. Moreover, the skill of recognizing textual entailment (TE) is not listed because we assume that TE involves a broad range of knowledge and inferences, and is therefore a generic task itself (Dagan, Glickman, and Magnini 2006).

ID: MC160.dev.1 (3) one:
C1: Sally had a very exciting summer vacation.
C2: She went to summer camp for the first time.
C3: Sally's favorite activity was walking in the woods because she enjoyed nature.
Q: Why does Sally like walking in the woods?
A: She likes nature.

---

Coreference resolution:
· *she* in C3 = *Sally* in C3
Causal relation:
· *she enjoyed nature* in C3 → *Sally's favorite activity was walking in the woods* in C3
Commonsense reasoning:
· *Sally's favorite activity was walking ...* in C3 ⇒ *Sally likes walking in the woods ...* in Q
· *enjoyed nature* in C3 ⇒ *likes nature* in A

ID: MC500.dev.36 (1) one:
C1: Shelly is in second grade.
C2: She is a new student at her school.
C3: Her new teacher, Mrs. Borden, makes her stand in front of the class and say something about herself.
Q: Who is Shelly's second grade teacher?
A: Mrs. Borden

---

Coreference resolution:
· *Shelly* in C1 = *she* in C2 = *her* in C3
Temporal relation:
· *Shelly is in second grade* in C1 → *Shelly's second grade ...* in Q
Schematic/Rhetorical clause relations:
· C3 = *... makes her stand ...* and *... [makes her] say ...*
Special sentence structure (apposition):
· *Her new teacher = Mrs. Borden* C3

Figure 2: Examples of task sentences and annotations with comments for verification (itemized).

## 4 Annotation of Reading Comprehension Task

### Dataset Description

MCTest (Richardson, Burges, and Renshaw 2013) is a reading comprehension task that requires open-domain understanding of stories with small vocabulary limited to what young children would understand (approximately 8,000 words). The task is multiple choice, and candidates for answers do not necessarily appear in the context. It consists of fictional stories (approximately 200 words long on average) with four questions per story. Each question relates to the content of the story. MCTest has two datasets: MC160 and MC500, with 160 and 500 stories, and 640 and 2,000 questions, respectively. The datasets of MCTest are smaller than those of other existing tasks; this is a problem for machine learning. On the contrary, MCTest is a high-quality task from the viewpoint of testing natural language understanding.

We chose MCTest for annotation for three reasons. The first is that MCTest is a multiple-choice task: expressions of answers have no limitation, that is, any words, phrases, and sentences are permitted; they need not actually appear in the context. This means that MCTest may require a wide range of skills. The second reason is that datasets of MCTest consist of elementary-level passages and questions: they seem to be easy for comprehension, and not difficult to annotate. The third reason for using MCTest is that every question on MCTest is annotated with one label, *multiple* or *one*, that reveals whether the question requires understanding multiple sentences or a single sentence. We can compare the results of our annotation with those labels for verification[4].

### Annotation Specification

We asked two annotators to annotate questions in the development sets of both MC160 and MC500 (30 and 50 stories, 120 and 200 questions, respectively) with RC skills required to answer each question by allowing multiple labeling. The

[4]We have made our annotation results publicly available (http://www-al.nii.ac.jp/mctest-rcskills-annot/).

inter-annotator agreement was 85.0% for eight stories (32 questions) that we randomly sampled.

Since RC skills are intended for understanding relations among clauses, the annotators were asked to exclude sentences with no relations with others and required only simple rules to answer, e.g., mc160.dev.2 (3) Context: *Todd lived in a town outside the city.* Q: *Where does Todd live?* A: *In a town*. These questions were considered to require no skill.

For combinations of multiple skills, consider the next toy example sharing the same context (C):

C: *The name of John's sister is Sylvia. John was annoyed because his sister ate his cake.*

1. Q: *Why was John annoyed?* A: *Because his sister ate his cake.* Required skill: causal relation
2. Q: *Why was John irritated?* A: *Because his sister ate his cake.* Required skills: causal relation, commonsense reasoning (*annoyed = irritated*)
3. Q: *Why was John irritated?* A: *Because Sylvia ate his cake.* Required skills: causal relation, commonsense reasoning (*annoyed = irritated*, *The name of John's sister is Sylvia. = Sylvia is John's sister.*), coreference resolution (*his sister = John's sister*)

In the first example, the question is why John was annoyed. The reason is described in the context. Thus, one skill is required to the question: *causal relation*. In the second example, the word *annoyed* is paraphrased to *irritated*. To bridge this change, a system must be able to exercise common sense about the meanings of words. Thus, commonsense reasoning is listed as a required skill. In the last example, *his sister* in the answer is changed to *Sylvia*. To determine whether these two expressions have identical referents, a system must be able to exercise common sense about proper names and resolve coreference; three skills are finally listed. More concrete examples appearing in the annotation are shown in Figure 2.

## 5 Annotation Results and System Analysis

### Annotation Results

We first considered compatibility between the original MCTest annotation and our annotation. Note that the definition of RC skills targets relations among multiple clauses, whereas the original annotation caters to multiple sentences. In order to compare our annotation with the original one, we temporarily regarded only questions that required *schematic/rhetorical clause relations* and *special sentence structure*[5] as ones that required no skill. Considering also that, in our definition, questions labeled with *one* often required some RC skill (e.g., *commonsense reasoning*), we estimated consistency by checking the percentage of the questions with *multiple* and no RC skills, which were not desirable by definition. Therefore, the smaller this ratio, the more consistent the two annotations. The result was that the percentage of such questions was only 3.4% (11/320). This shows that our annotation agreed well with the original one.

[5]More precisely, we checked whether there was a question that required only one of the last five skills in Table 1, and we regard those questions as ones requiring single sentence understanding (*one*). Nonetheless, we found two of such questions (mc500.dev.9 (2), mc500.dev.48 (1)) with *multiple* in the original annotation.

| #RC skills | Freq. | Baseline SW+D | Smith No RTE | Smith RTE | Yin ABCNN |
|---|---|---|---|---|---|
| 0 | 10.3 | 57.6 | 72.7 | 75.8 | 54.5 |
| 1 | 28.4 | 52.7 | 67.6 | 67.9 | 47.3 |
| 2 | 28.4 | 51.6 | 66.5 | 64.8 | 50.5 |
| 3 | 23.8 | 47.4 | 67.1 | 69.1 | 46.1 |
| 4 | 8.1 | 46.2 | 52.2 | 44.6 | 42.3 |
| 5 | 0.9 | 33.3 | 41.7 | 33.3 | 33.3 |

Table 2: Frequencies and accuracies (%) for required numbers of RC skills for each question in the development sets of MC160 and MC500 (320 questions).

| RC skills | Freq. | Baseline SW+D | Smith No RTE | Smith RTE | Yin ABCNN |
|---|---|---|---|---|---|
| List/Enumeration | 14.7 | 51.1 | 65.1 | 61.9 | 40.4 |
| Mathematical ops. | 1.6 | 20.0 | 30.0 | 35.0 | 60.0 |
| Coreference res. | 63.8 | 52.5 | 63.6 | 62.1 | 48.0 |
| Logical rsng. | 0.9 | 100.0 | 75.0 | 66.7 | 33.3 |
| Analogy | 0.3 | 0.0 | 100.0 | 100.0 | 0.0 |
| Spatiotemporal rels. | 27.5 | 48.9 | 66.9 | 67.1 | 45.5 |
| Causal rels. | 14.4 | 45.7 | 62.0 | 60.9 | 52.2 |
| Commonsense rsng. | 41.9 | 44.0 | 61.3 | 59.6 | 44.8 |
| S/R clause rels. | 20.6 | 50.0 | 65.9 | 64.0 | 48.5 |
| Special sentence stru. | 8.1 | 46.2 | 69.2 | 73.1 | 46.2 |
| Accuracy | - | 50.9 | 66.2 | 65.9 | 48.1 |

Table 3: Frequencies and accuracies (%) for RC skills in the development sets of MC160 and MC500 (320 questions). "S/R" is an abbreviation for "Schematic/Rhetorical."

In Table 2, we report a more concrete distribution for the number of skills required for each question. From this, we can observe two important points. First, 89.7% of the MCTest questions required at least one skill. This simply indicated that questions of MCTest usually involve multiple clauses. This feature is useful for evaluating the quality of existing RC tasks or constructing a new RC task in terms of contextual understanding, rather than question answering through simple sentences. Second, 61.3% of the questions required multiple RC skills (avg. 1.94). This explains the complexity of RC systems often need to take several factors into account for comprehension.

Table 3 shows the frequency of each RC skill. *Coreference resolution* and *commonsense reasoning* were more often required than others; we can conclude that in the first place systems proposed for MCTest should be readied for these skills. The third most frequent RC skill was *spatiotemporal relations*, which might have been because MCTest contains stories of vacations, animals, school, and so forth. We found few questions that required *mathematical operations*, *logical reasoning*, and *analogy* because MCTest gauges reading comprehension for children.

In this way, the characteristics of the dataset were clarified through the proposed set of skills. Based on such analysis

of the dataset, we can plan the direction of development of RC systems, e.g., developers can make it more efficient to improve their systems if they concentrate on skills that the dataset especially requires, and at which their system is not adept.

## System Analysis

We analyzed the performance of four systems based on the annotation results. The first (Baseline SW+D) is Richardson, Burges, and Renshaw (2013)'s baseline system that uses a sliding window and a word distance algorithm. The second (Smith No RTE) was proposed by Smith et al. (2015). This system is provided with a lexical matching method that uses stemming and takes into account both the type of question (*wh*-words) and coreference resolution. The third (Smith RTE) is an extension of the second that recognizes textual entailment (RTE) (Stern and Dagan 2011). We chose Smith's systems because they inherit the baseline algorithm. Thus, we expected that the differences among the systems would be directly reflected in performance. The last (Yin ABCNN) is a hierarchical attention-based convolutional neural network (HABCNN-TE) that answers a question as textual entailment without any linguistic feature (Yin, Ebert, and Schütze 2016). It should be noted that the performance of the system we report in this paper is incompatible with the one of the original paper (Yin, Ebert, and Schütze 2016): this might be contributed to the difference of the data used for the evaluation (test vs. development). Also, we used the following hyperparameters: the learning rate ($lr$) was 0.03, the constant in loss function ($\alpha$) was 0.19, and the other variables were the same as in the original system. It was intended for an analysis of its performance rather than measurement. In the following, we analyze the performance of these systems.

**Four systems:** The accuracy values of the systems for each skill are shown in Table 2. The first three systems achieved higher accuracies (than the average) for each question that required no skill and one skill. By contrast, for four and five skills, all systems yielded lower accuracy values. This indicates that systems are not good at reasoning with multiple sentences that have complex relations. Moreover, in all systems, the accuracy values for *coreference resolution* and *commonsense reasoning* were generally lower than total accuracy (Table 3). This result also indicates that systems need to be able to accommodate these two skills, as in the analysis of the datasets above.

**Baseline SW+D vs. Smith No RTE:** As shown in Tables 2 and 3, almost all accuracy values improved from the baseline. As mentioned above, however, the accuracy values for *coreference resolution* and *commonsense reasoning* were lower than the average. For further analysis, we investigated the accuracy values for these two skills when each appeared alone or together with other skills (Table 4). This result had two important implications. The first was that the skill of *coreference resolution* improved when it appeared by itself. This indicates that a method for coreference resolution worked effectively in Smith No RTE. The second was that the skill of *commonsense reasoning* did not improve when it appeared by itself, though the total accuracy for this skill improved (from 44.0% to 61.3%). This might be obvious, given that the latter system did not accommodate commonsense reasoning.

| | Coreference resolution | | Commonsense reasoning | |
|---|---|---|---|---|
| System | single (47) | multiple (157) | single (20) | multiple (114) |
| Baseline SW+D | 59.6 | 50.3 | 60.0 | 41.2 |
| Smith No RTE | 66.0 | 63.0 | 60.0 | 61.5 |
| Yin ABCNN | 46.8 | 48.4 | 45.0 | 44.7 |

Table 4: Accuracies (%) for coreference resolution and commonsense reasoning when each skill appeared alone ("single") or together with other skills ("multiple"). Numbers with "single" and "multiple" show occurrences in the development sets of MC160 and MC500 (320 questions).

**Smith No RTE vs. Smith RTE:** Unfortunately we could not observe that, at least in the development sets, adding RTE significantly increased accuracy (Tables 2 and 3). Nonetheless, we noticed that accuracy for questions requiring no skill improved compared to other questions. We think this result is because RTE is suitable for reasoning regarding a relation between two sentences, a hypothesis made from both a question and its candidate answer, and context sentence (which requires no skill according to our definitions), but is not suitable for questions involving multiple context sentences (which required one or more skills).

**Baseline SW+D vs. Yin ABCNN:** The attention-based CNN could not achieve superior accuracy results to the other systems. Nevertheless, it should be emphasized that the result of this neural system was comparable with the baseline, despite the fact that this system did not leverage any linguistic feature. Specifically, the accuracies of two skills (*coreference resolution* and *commonsense reasoning*) appeared with other skills ("multiple" in Table 4) were competitive toward the results of the baseline. Moreover, "single" and "multiple" indicated very similar scores; we can conjecture that this reflected the architecture of ABCNN: that is, jointly computing attentions for multiple sentences in the context.

# 6 Discussion

In this section, we discuss issues concerning our annotation and explore directions for future research.

## Analogy

At least in MCTest, we found few questions that required the skill of analogy. We think this was the case because there was a certain vagueness in the distinction of commonsense reasoning and analogy in our annotation: for example, as in the definition of commonsense reasoning, we simply regarded the relation between *enjoy nature* and *like nature* (Figure 2) as paraphrased in commonsense knowledge. However, this paraphrase seems to contain the understanding of metonymy. It follows that analogy requires more consideration of its definition (metonymy, synecdoche,

metaphor, and so forth) and concrete examples in the annotation guideline.

As for analogy, LoBue and Yates (2011) defined *synecdoche* as an individual category of commonsense knowledge. Considering this classification, we will refine the definition of analogy.

### Topical Understanding

Some questions require understanding a topic of each story as background knowledge. See mc500.dev.44 (3):

C: *Jake [...] was playing baseball every day. The more he played, the better he got.*

Q: *How did Jake get so good at baseball?*

A: *He played a lot.*

This task requires understanding the topic, playing baseball, because the second sentence does not directly describe that Jake improved at baseball. This problem is not simply solved by coreference resolution. Systems must know previous sentences as background topics and interpret a new sentence using them. Such an inference is similar to *bridging* (Asher and Lascarides 2003), which is implicitly performed using some commonsense knowledge.

### Commonsense Reasoning

In the annotation, we recognized that the skill of commonsense reasoning covered a wide range of areas of knowledge; thus, it was difficult to analyze sentences that required commonsense reasoning.

Commonsense reasoning in our classification included all processes of reasoning involving a certain kind of knowledge. Therefore, we also had to consider difference between commonsense and domain-specific knowledge, such as is required in elementary science tests, say Clark, Harrison, and Balasubramanian (2013).

We will therefore refine the categorization of commonsense according to discussions by LoBue and Yates (2011), Schubert (2015) and Davis and Marcus (2015), among others.

### Story Narratives

As Smith et al. (2015) indicated, some questions in MCTest require variation in perspective at the level of narrative. Certain questions require understanding meta-level information (e.g., *Who are the principal characters of the story?*) concerning the story's narrative, whereas ordinary questions relate to the information concerning individual events or situations in the story. For such narrative questions, the systems must understand the contents of the story and recognize that they are actually "reading" that story, from meta-viewpoints.

### Use and Development of Our Methodology

We have made our annotated skill labels and an annotation guideline publicly available so that they can be easily used by other researchers. If RC tasks are provided with our RC skills, they will help other researchers understand the characteristics of datasets, develop systems to tackle those datasets, and analyze errors. Specifically, we recognize that the major problem with neural systems is their reproducibility. When researchers of neural systems make use of annotated datasets, it encourages the understanding of the internal architectures of systems.

A problem with our current methodology is that an annotation incurs cost and requires annotators to understand the definitions of skills. Therefore, we should build a method that realizes smooth and easy annotations. One such methods involves providing yes/no questions that help annotators decide one among mutually involved skills: for example, between analogy and commonsense knowledge. For this method, we may have to establish exclusive subclasses for commonsense knowledge and other related areas.

## 7 Conclusion and Future Work

In this paper, we proposed a methodology for evaluating and analyzing RC datasets and systems. We defined a set of prerequisite skills for RC, and annotated an existing task with these skills. Based on the annotation, we were able to analyze the characteristics of an RC task and the difference among existing systems in terms of performance from multiple aspects. We concluded that the defined skills are promising for the decomposition and analysis of RC.

We believe that our methodology is general, and hence can be applied to other formulations of RC tasks. Although existing tasks are mostly Cloze or multiple choice, Clark and Etzioni (2016) discussed the importance of the ability to "explain" a fact. They also mentioned the ability to understand a diagram and a figure as context. Such task formulation is more difficult than existing ones. However, analyzing systems with a steady methodology like ours can help promote improvement.

For future work, we plan to build a framework to construct simple tasks that can generally be used for unit testing for reading comprehension. The idea of this framework is inspired by Weston et al. (2015). We will modify the formulation of bAbI tasks, create new tasks based on the RC skills, and apply these tasks to existing systems.

## Acknowledgments

We thank anonymous reviewers for their insightful comments. This work was supported by JSPS KAKENHI Grants Number 15H02754 and 16K16120.

## References

Asher, N., and Lascarides, A. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

Berant, J.; Srikumar, V.; Chen, P.-C.; Vander Linden, A.; Harding, B.; Huang, B.; Clark, P.; and Manning, C. D. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1499–1510. Association for Computational Linguistics.

Chen, D.; Bolton, J.; and Manning, D. C. 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*, 2358–2367. Association for Computational Linguistics.

Clark, P., and Etzioni, O. 2016. My computer is an honor student-but how intelligent is it? standardized tests as a measure of ai. *AI Magazine* 37(1):5–12.

Clark, P.; Harrison, P.; and Balasubramanian, N. 2013. A study of the knowledge base requirements for passing an elementary science test. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 37–42. ACM.

Clark, P. 2015. Elementary school science and math tests as a driver for AI: Take the Aristo challenge! In *AAAI Conference on Artificial Intelligence*, 4019–4021.

Cooper, R.; Crouch, R.; van Eijck, J.; Fox, C.; van Genabith, J.; Jaspers, J.; Kamp, H.; Pinkal, M.; Poesio, M.; Pulman, S.; et al. 1994. Fracas: A framework for computational semantics. *Deliverable* 8:62–051.

Dagan, I.; Glickman, O.; and Magnini, B. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*. Springer. 177–190.

Davis, E., and Marcus, G. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM* 58(9):92–103.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, 1693–1701.

Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hirschman, L.; Light, M.; Breck, E.; and Burger, J. D. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 325–332. Association for Computational Linguistics.

Huddleston, R., and Pullum, G. K. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

LoBue, P., and Yates, A. 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*, 329–334. Portland, Oregon, USA: Association for Computational Linguistics.

Paperno, D.; Kruszewski, G.; Lazaridou, A.; Pham, Q. N.; Bernardi, R.; Pezzelle, S.; Baroni, M.; Boleda, G.; and Fernandez, R. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1525–1534. Association for Computational Linguistics.

Pradhan, S.; Ramshaw, L.; Marcus, M.; Palmer, M.; Weischedel, R.; and Xue, N. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 1–27. Portland, Oregon, USA: Association for Computational Linguistics.

Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; and Zhang, Y. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR* abs/1606.05250.

Richardson, M.; Burges, J. C.; and Renshaw, E. 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 (EMNLP)*, 193–203.

Roemmele, M.; Bejan, C. A.; and Gordon, A. S. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Sammons, M.; Vydiswaran, V.; and Roth, D. 2010. "ask not what textual entailment can do for you...". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1199–1208. Uppsala, Sweden: Association for Computational Linguistics.

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

Smith, E.; Greco, N.; Bosnjak, M.; and Vlachos, A. 2015. A strong lexical matching method for the machine comprehension test. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1693–1698. Association for Computational Linguistics.

Snow, C. 2002. *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.

Stern, A., and Dagan, I. 2011. A confidence model for syntactically-motivated entailment proofs. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 455–462. Hissar, Bulgaria: RANLP 2011 Organising Committee.

Sugawara, S., and Aizawa, A. 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*, 1–5. Austin, TX: Association for Computational Linguistics.

Sutcliffe, R.; Peñas, A.; Hovy, E.; Forner, P.; Rodrigo, Á.; Forascu, C.; Benajiba, Y.; and Osenova, P. 2013. Overview of QA4MRE main task at CLEF 2013. *Working Notes, CLEF*.

Trischler, A.; Ye, Z.; Yuan, X.; He, J.; Bachman, P.; and Suleman, K. 2016. A parallel-hierarchical model for machine comprehension on sparse data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Weston, J.; Bordes, A.; Chopra, S.; and Mikolov, T. 2015. Towards AI-complete question answering: a set of prerequisite toy tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xue, N.; Ng, H. T.; Pradhan, S.; Bryant, R. P. C.; and Rutherford, A. T. 2015. The CoNLL-2015 shared task on shallow discourse parsing. *CoNLL 2015*.

Yin, W.; Ebert, S.; and Schütze, H. 2016. Attention-based convolutional neural network for machine comprehension. In *Proceedings of the Workshop on Human-Computer Question Answering*, 15–21. San Diego, California: Association for Computational Linguistics.