An Analysis of Prerequisite Skills for Reading Comprehension

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

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

Abstract

In this paper, we focus on the synthetic understanding of documents, specifically reading comprehension (RC). A current problem with RC is the need for a method of analyzing the RC system performance to realize further development. We propose a methodology for examining RC systems from multiple viewpoints. Our methodology consists of three steps: define a set of basic skills used for RC, manually annotate questions of an existing RC task, and show the performances for each skill of existing systems that have been proposed for the task. We demonstrated the proposed methodology by annotating MCTest, a freely available dataset for testing RC. The results of the annotation showed that answering RC questions requires combinations of multiple skills. In addition, our defined RC skills were found to be useful and promising for decomposing and analyzing the RC process. Finally, we discuss ways to improve our approach based on the results of two extra annotations.

1 Introduction

Reading comprehension (RC) tasks require machines to understand passages and respond to questions about them. For the development of RC systems, precisely identifying what systems can and cannot understand is important. However, a critical problem is that the RC process is so complicated that it is not easy to examine the performances of RC systems.

Our present goal is to construct a general evaluation methodology that decomposes the RC process and elucidates the fine-grained performance from multiple points of view rather than based only on accuracy, which is the approach used to date. Our methodology has three steps:

1. Define a set of prerequisite skills that are required for understanding documents (Section 2.1)
2. Annotate questions of an RC task with the skills (Section 2.2)
3. Analyze the performances of existing RC systems for the annotated questions to grasp the differences and limitations of their individual performances (Section 2.3)

In Section 2, we present an example of our methodology, where we annotated MCTest (MC160 development set) (Richardson et al., 2013)[1] for Step 2 and analyzed systems by Smith et al. (2015) for Step 3. In Section 3, we present two additional annotations in order to show the outlook for the development of our methodology in terms of the classification of skills and finer categories for each skill. In Section 4, we discuss our conclusions.

2 Approach

2.1 Reading Comprehension Skills

We investigated existing tasks for RC and defined a set of basic prerequisite skills, which we refer to as *RC skills*. These are presented in Table 1.

The RC skills were defined to understand the relations between multiple clauses. Here, we assumed

[1] http://research.microsoft.com/en-us/um/redmond/projects/mctest/

RC skills	Freq.	Descriptions or examples	Smith no RTE	Smith RTE
List/Enumeration	11.7%	Tracking, retaining, and list/enumeration of entities or states	78.6%	71.4%
Mathematical operations	4.2%	Four basic operations and geometric comprehension	20.0%	20.0%
Coreference resolution	57.5%	Detection and resolution of coreferences	65.2%	69.6%
Logical reasoning	0.0%	Induction, deduction, conditional statement, and quantifier	-	-
Analogy	0.0%	Trope in figures of speech, e.g., metaphor	-	-
Spatiotemporal relations*	28.3%	Spatial and/or temporal relations of events	70.6%	76.5%
Causal relations*	18.3%	Why, because, the reason, etc.	63.6%	68.2%
Commonsense reasoning	49.2%	Taxonomic/qualitative knowledge, action and event change	59.3%	64.4%
Complex sentences*	15.8%	Coordination or subordination of clauses	52.6%	68.4%
Special sentence structure*	10.0%	Scheme in figures of speech, constructions, and punctuation marks	50.0%	50.0%
-	-	(Accuracy in all 120 questions)	67.5%	70.0%

Table 1: Reading comprehension skills, their frequencies (in percentage) in MCTest (MC160 development set, 120 questions), their descriptions or examples, and the accuracies of the two systems (Smith et al., 2015) for each skill. The asterisks (*) with items represent "understanding of."

ID: MC160.dev.29 (1) multiple:
C1: The princess climbed out the window of the high tower and climbed down the south wall when her mother was sleeping.
C2: She wandered out a good ways.
C3: Finally she went into the forest where there are no electric poles but where there are some caves.
Q: Where did the princess wander to after escaping?
A: Forest

Coreference resolution:
· *She* in C2 = *the princess* in C1
· *She* in C3 = *the princess* in C1
Temporal relations:
· the actions in C1 → *wandered out ...* in C2
→ *went into...* in C3
Complex sentence and special sentence structure:
· C1 = *the princess climbed out...*
and [*the princess*] *climbed down...* (ellipsis)
Commonsense reasoning:
· *escaping* in Q ⇒ the actions in C1
· *wandered out* in C2 and *went into the forest*
in C3 ⇒ *wander to the forest* in Q and A

Figure 1: Example of task sentences in MCTest and annotations with comments for verification (itemized).

that, when an RC system uses an RC skill, it must already recognize individual facts described in those clauses to which the skill relates.

There are two exceptional RC skills:

Complex sentences target the understanding of relations between clauses in one sentence (except those having spatiotemporal or causal meanings). Such relations have schematic or rhetorical meanings. For example, the words "and" and "or" introduce coordinating clauses (we regard them as having schematic relations). In addition, the word "although" introduces a subordinate clause that represents concession (i.e., it modifies the rhetorical meaning).

Special sentence structure is defined as recognizing linguistic symbols or structures in a sentence and introducing their interpretations as new facts. For example, if we take "scheme" in figures of speech, this skill deals with apposition, ellipsis, transposition, and so on. This skill also targets linguistic constructions and punctuation marks.

These two skills target a single sentence, while the other skills target multiple clauses and sentences. We did not list the skill of recognizing textual entailment (TE) because we assumed that TE involves a broad range of knowledge and inferences and is therefore a generic task itself (Dagan et al., 2006).

2.2 Annotation of RC Questions

We manually annotated the questions of the MC160 development set (120 questions) with the RC skills that are required to answer each question. In the annotation, we allow multiple labeling.

Because the RC skills are intended for understanding relations between multiple clauses, we excluded sentences that had no relations with others and required only simple rules for answering (e.g., mc160.dev.2 (3) Context: Todd lived in a town outside the city. Q: Where does Todd live in? A: in a town). These questions were considered to require no skills.

An example of the annotations is shown in Figure 1. The percentages of the questions in which RC skills appear are in the second column of Table 1. Some of the questions are annotated with multiple labels. The number of skills required in each question is 0 for 9.2% of the questions, 1 for 27.5%, 2 for 30.0%, 3 for 26.7%, 4 for 5.8%, and 5 for 0.8%.

2.3 Analysis of Existing Systems

The accuracies of the system by Smith et al. (2015) and its extension with RTE (Stern and Dagan, 2011) are represented in the last two columns of Table 1.

The results showed that adding RTE to the Smith et al. (2015)'s original system provided the most effective contribution to the skill of *complex sentences*; however, it did not affect the skills of *math operations* and *special sentence structure*. Adding RTE had a relatively small contribution to the skill of *causal relations*. This did not exactly meet our expectation because we still do not have sufficient number of annotations to determine the differences between combinations of skills.

3 Additional Annotations

In order to improve our methodology, we considered two questions: (i) What is the difference between distributions of RC skills in two RC tasks? (ii) Can RC skills be broken up into finer categories?

To answer these questions, here we present two additional annotations. The first treated SQuAD (Rajpurkar et al., 2016). We counted the frequencies of RC skills required in that task and compared their distribution with that of MCTest. This gave clues for establishing the ideal categorization of RC skills.

For the second, we divided the skill of *commonsense reasoning* into three subcategories and used them to annotate MCTest. This should help for a sharper definition of common sense.

3.1 SQuAD with RC Skills

SQuAD[2] is an RC task based on a set of Wikipedia articles. The questions are made by crowdworkers, and their answers are sure to appear in the context as a word sequence. We chose 80 questions over seven articles from the development set (v1.1) and annotated them with RC skills. Figure 2 shows an example of the annotations.

RC skills	Frequency SQuAD	Frequency MCTest
List/Enumeration	5.0%	11.7%
Mathematical operations	0.0%	4.2%
Coreference resolution	6.2%	57.5%
Logical reasoning	1.2%	0.0%
Analogy	0.0%	0.0%
Spatiotemporal relations	2.5%	28.3%
Causal relations	6.2%	18.3%
Commonsense reasoning	86.2%	49.2%
Complex sentences	20.0%	15.8%
Special sentence structure	25.0%	10.0%

Table 2: Reading comprehension skills and their frequencies (in percentage) in SQuAD and MCTest (MC160 development set).

The annotation results are presented in Table 2. Most questions require *commonsense reasoning*. This is because the crowdworkers were asked to avoid copying words from their context as much as possible. That is, most questions require understanding of paraphrases. Compared with MCTest, the frequencies were generally low except for a few skills. This was due to the task formulation of SQuAD. For example, because SQuAD does not involve multiple choice (a candidate answer can contain multiple entities), the skill of *list/enumeration* is not required. Additionally, except for articles on a particular person or historical event, there are fewer descriptions that require *spatiotemporal relations* than in MCTest, whose datasets mainly describe tales about characters and events for young children. On the other hand, *complex sentences* and *spacial sentence structure* appear more frequently in SQuAD than in MCTest because the documents of SQuAD are written for adults. In this way, by annotating RC tasks and comparing the results, we can see the difference in characteristics among those tasks.

3.2 MCTest with Commonsense Types

By referring to Davis and Marcus (2015), we defined the following three types of common sense, as given in Table 3, and annotated the MC160 development set while allowing multiple labeling. We found three questions that required multiple types.

Lexical knowledge focuses on relations of words or phrases, e.g., synonyms and antonyms, as in WordNet. This includes hierarchical relations of

[2] http://stanford-qa.com

ID: Civil_disobedience, paragraph 1, question 1
C1: One of its earliest massive implementations was brought about by Egyptians against the British occupation in the 1919 Revolution.
C2: Civil disobedience is one of the many ways people have rebelled against what they deem to be unfair laws.
Q: What is it called when people in society rebel against laws they think are unfair?
A: Civil disobedience

Coreference resolution:
· *they* in C2 = *people* in C2 (different clauses)
· *they* in Q = *people* in Q (different clauses)
Temporal relation:
· *people have rebelled...* in C2
→ *when people in society rebel...* in Q
Complex sentences:
· C2 = *one of the many ways people have* (relative clause)
· C2 = *Civil disobedience is... against* [the object]
and [it is] *what they deem to...* (relative clause)
· Q = *What is it called... laws*
and *they think* [the laws] *unfair?*
Commonsense reasoning:
· *What is it called* in Q ⇒ *Civil disobedience is*
· *laws they think...* in C2 = *what they deem to* in Q

Figure 2: Example of task sentences (excerpted) in the development set of SQuAD and their annotations with comments for verification (itemized).

content words. Therefore, this knowledge is taxonomic and categorical.

Qualitative knowledge targets various relations of events, including "about the direction of change in interrelated quantities" (Davis and Marcus, 2015). In addition, this knowledge deals with implicit causal relations such as physical law and theory of mind. Note that these relations are semantic, so this type of knowledge ignores the understanding of syntactic relations, i.e., the skills of *spatiotemporal relations* and *causal relations*.

The skill of *known facts* targets named entities such as proper nouns, locations, and dates. Davis and Marcus (2015) did not mention this type of knowledge. However, we added this just in case because we considered the first two types as unable to treat facts such as a proper noun indicating the name of a character in a story.

Table 3 presents the frequencies of these types and accuracies of Smith et al. (2015)'s RTE system. Because MCTest was designed to test the capability of young children's reading, known facts were hardly required. Although not reported here, we found that understanding them was more required in MC500 (e.g., days of a week). While the frequencies of the first two types were relatively high, their accuracies were comparable. Unfortunately, this meant that they were inadequate for revealing the weakness of the system on this matter. We concluded that finer classification is needed. However, the distribution of the frequencies showed that even these commonsense types can characterize a dataset in terms of the knowledge types required in that task.

Commonsense type	Frequency	Accuracy Smith RTE
Lexical knowledge	19.2%	67.2%
Qualitative knowledge	30.8%	67.6%
Known facts	2.5%	33.3%

Table 3: Commonsense types, their frequencies (in percentage) in MCTest (MC160 development set), and accuracies by Smith et al. (2015)'s RTE system.

4 Discussion and Conclusion

As discussed in Section 2.3, our methodology has the potential to reveal differences in system performances in terms of multiple aspects. We believe that it is necessary to separately test and analyze new and existing RC systems on each RC skill in order to make each system more robust. We will continue to annotate other datasets of MCTest and RC tasks and analyze the performances of other existing systems.

From the observations presented in this paper, we may be able to make a stronger claim that researchers of RC tasks (more generally, natural language understanding) should also provide the frequencies of RC skills. This will help in developing a standard approach to error analysis so that systems can be investigated for their strengths and weaknesses in specific skill categories. We can determine the importance of each skill by weighting them according to their frequencies in the test set.

Acknowledgments

We are grateful to the anonymous reviewers for their thoughtful comments and the workshop organizers. This work was supported by JSPS KAKENHI Grant Number15H0275.

References

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*, pages 177–190. Springer.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

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 (EMNLP)*, pages 193–203.

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

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 (EMNLP)*, pages 1693–1698. Association for Computational Linguistics.

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

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