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

---

*Saku Sugawara (The University of Tokyo)
Hikaru Yokono (Fujitsu Laboratories Ltd.)
Akiko Aizawa (National Institute of Informatics, Japan)

AAAI-17

February 8, 2017

Introduction - Reading Comprehension (RC)

What is reading comprehension?

ID: MCTest 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: A) Mountain *B) Forest C) Cave D) Castle

Introduction - Reading Comprehension (RC)

ID: MCTest 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: A) Mountain *B) Forest C) Cave D) Castle

Coreference resolution (*she* = *princess*)
Commonsense reasoning (*escaping* = *climbed down*)
Temporal relation (*climbed* $\rightarrow$ *wandered*)

Common System Analysis by Accuracy

Dataset A	System X
Q1	x
Q2	o
Q3	x
⋮	⋮
Q100	o
Accuracy	75.0%

$\rightarrow$ Only with accuracy, we cannot tell what the systems understand and what they don't.

Our Research Question

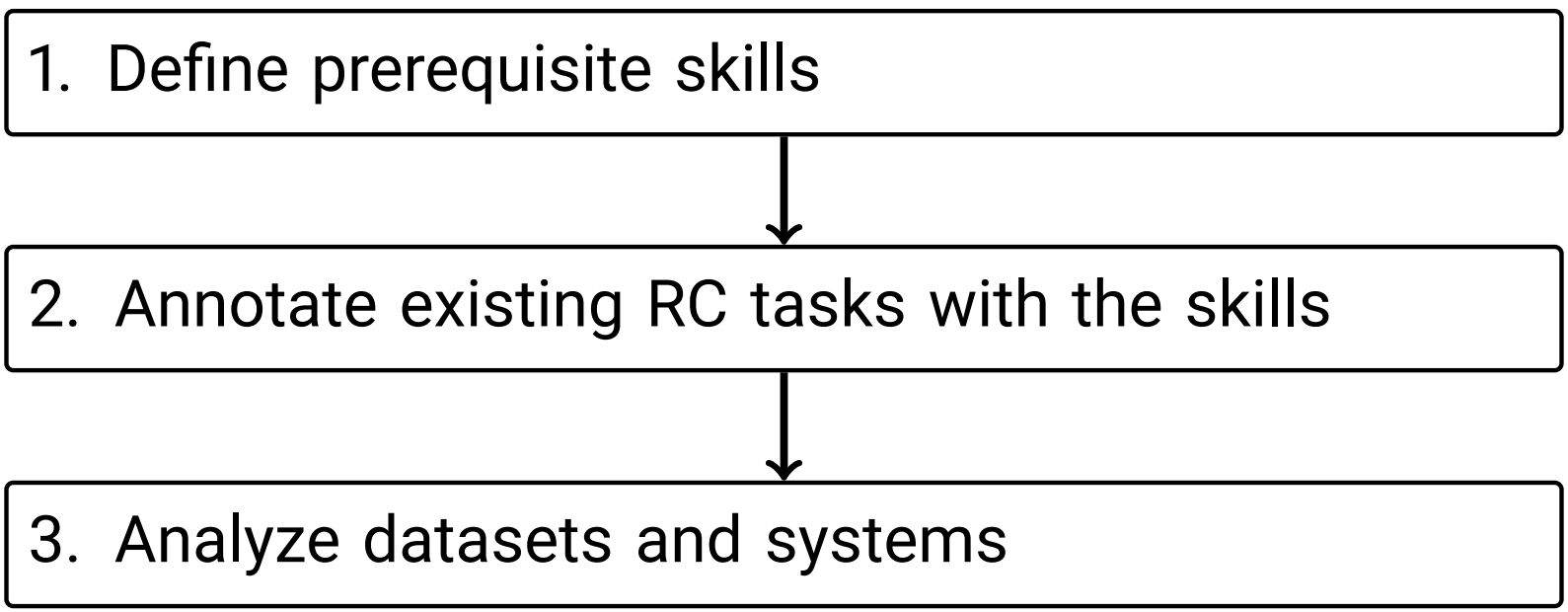
- How can we evaluate and analyze our RC systems?

→ Simgle accuracy-based analysis is insufficient!!!

→ We need a methodology for more detailed analysis

Our study: A Methodology for Evaluation of Reading Comprehension

Consists of three steps:



Common System Analysis by Accuracy

Dataset A	System X
Q1	x
Q2	o
Q3	x
⋮	⋮
Q100	o
Accuracy	75.0%

System Analysis by Prerequisite Skills

1. Define prerequisite skills

Question	Dataset A				System X
	Skill 1	Skill 2	···	Skill 10	
Q1					x
Q2					o
Q3					x
⋮					⋮
Q100					o
Accuracy	-	-	···	-	75.0%

System Analysis by Prerequisite Skills

2. Annotation RC questions with defined skills

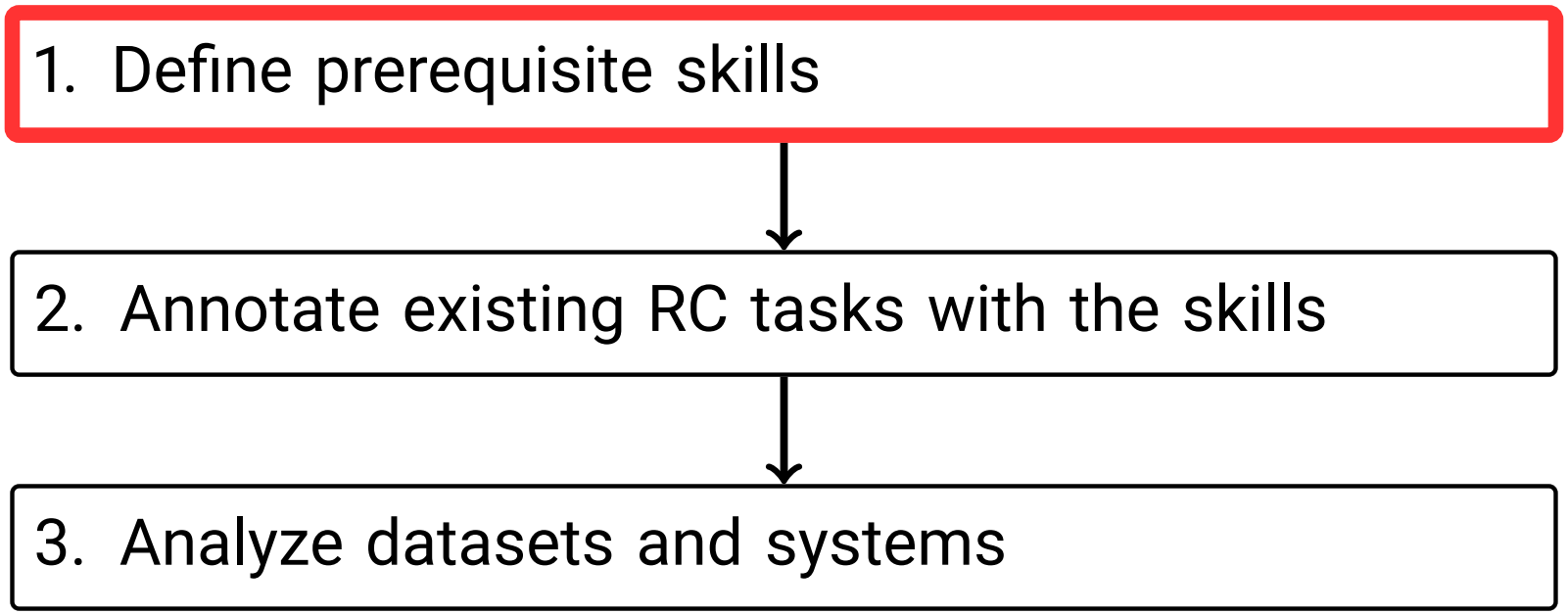
Question	Dataset A Skill 1	Skill 2	⋯	Skill 10	System X
Q1	yes	no	⋯	yes	x
Q2	no	yes	⋯	no	o
Q3	yes	yes	⋯	no	x
⋮	⋮	⋮	⋮	⋮	⋮
Q100	yes	yes	⋯	yes	o
Accuracy	-	-	⋯	-	75.0%

System Analysis by Prerequisite Skills

3. Analyze datasets and systems

	Dataset A				System X
Question	Skill 1	Skill 2	···	Skill 10	
Q1	x	-	···	x	x
Q2	-	o	···	-	o
Q3	x	x	···	-	x
⋮	⋮	⋮	⋮	⋮	⋮
Q100	o	o	···	o	o
Accuracy	40.0%	90.0%	···	70.0%	75.0%
	[not good]	[good]			

Our study: A Methodology for Evaluation of Reading Comprehension



1. Prerequisite Skills and Their Descriptions

Prerequisite Skills	Descriptions	Major tasks
1. **List/Enumeration**	Tracking, retaining, and list/enumeration of entities or states	bAbI
2. **Mathematical operations**	Four arithmetic operations and geometric comprehension	Aristo
3. **Coreference resolution**	Detection and resolution of coreference	CoNLL2012st
4. **Logical reasoning**	Induction, deduction, conditional statement, and quantifier	Aristo, FraCaS
5. **Analogy**	Metaphor etc.	-
6. **Spatiotemporal relations***	Spatial and/or temporal relations	CoNLLst2015, bAbI
7. **Causal relations***	Relations of events expressed by why, because, the reason...	COPA, CoNLLst2015
8. **Commonsense reasoning**	Taxonomic/qualitative knowledge, action, and event changes	COPA, WSC
9. **Schematic clause relations***	{Co/sub}ordination of clauses	CoNLLst2015
10. **Special sentence structure***	Constructions and punctuation marks in a sentence	-

The asterisks (*) with items represent "understanding of." / CoNLLst2015 = Shallow Discourse Parsing for PDTB

Prerequisite Skills and Their Descriptions

Prerequisite Skills	Descriptions	Major tasks
1. **List/Enumeration**	Tracking, retaining, and list/enumeration of entities or states	bAbI
2. **Mathematical operations**	Four arithmetic operations and geometric comprehension	Aristo
3. **Coreference resolution**	Detection and resolution of coreference	CoNLL2012st
4. **Logical reasoning**	Induction, deduction, conditional statement, and quantifier	Aristo, FraCaS
5. **Analogy**	Metaphor etc.	-
6. **Spatiotemporal relations***	Spatial and/or temporal relations	CoNLLst2015, bAbI
7. **Causal relations***	Relations of events expressed by why, because, the reason...	COPA, CoNLLst2015
8. **Commonsense reasoning**	Taxonomic/qualitative knowledge, action, and event changes	COPA, WSC
9. **Schematic clause relations***	{Co/sub}ordination of clauses	CoNLLst2015
10. **Special sentence structure***	Constructions and punctuation marks in a sentence	-

The asterisks (*) with items represent “understanding of.” / CoNLLst2015 = Shallow Discourse Parsing for PDTB

Toy Examples for Prerequisite Skills

Context:

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

Q1: *Why was John annoyed?*

A1: *Because his sister ate his cake.*

Toy Examples for Prerequisite Skills

Context:

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

Q1: *Why was John annoyed?*

A1: *Because his sister ate his cake.*

Required skill:

- **Causal relation**
  Why $\rightarrow$ *Because ...*

Toy Examples of Prerequisite Skills

Context:

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

Q2: *Why was John irritated?*

A2: *Because Sylvia ate his cake.*

Toy Examples for Prerequisite Skills

Context:

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

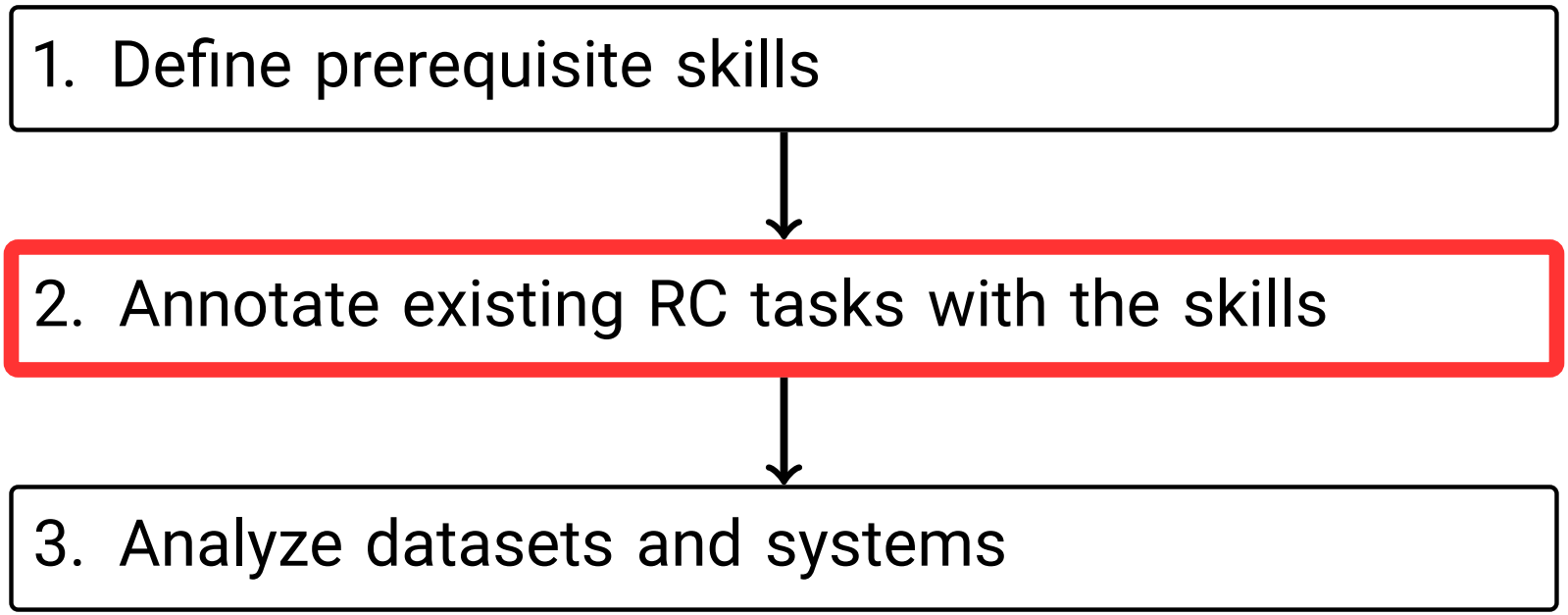
Q2: *Why was John irritated?*

A2: *Because Sylvia ate his cake.*

Required skills:

- **Causal relation**
  Why $\rightarrow$ *Because ...*
- **Commonsense reasoning**
  annoyed = irritated
  The name of John's sister is Sylvia. = Sylvia is John's sister.
- **Coreference resolution**
  his sister = John's sister

Our study: A Methodology for Evaluation of Reading Comprehension



2. Annotation Datasets

- MCTest development sets [Richardson+ 2013]
  Corpus: narratives written for young children
  Task formulation: multiple choice
  320 questions from the development sets from MC160+MC500
  Agreement: 85% (two annotators; for sampled questions)

Additional:

- SQuAD development set [Rajpurkar+ 2016]
- Corpus: expository texts from Wikipedia
- Task: selecting text span in context
  80 questions from the development set (v1.1)
  Details are reported in Sugawara+ (2016) EMNLP workshop

Annotation Example in MCTest

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

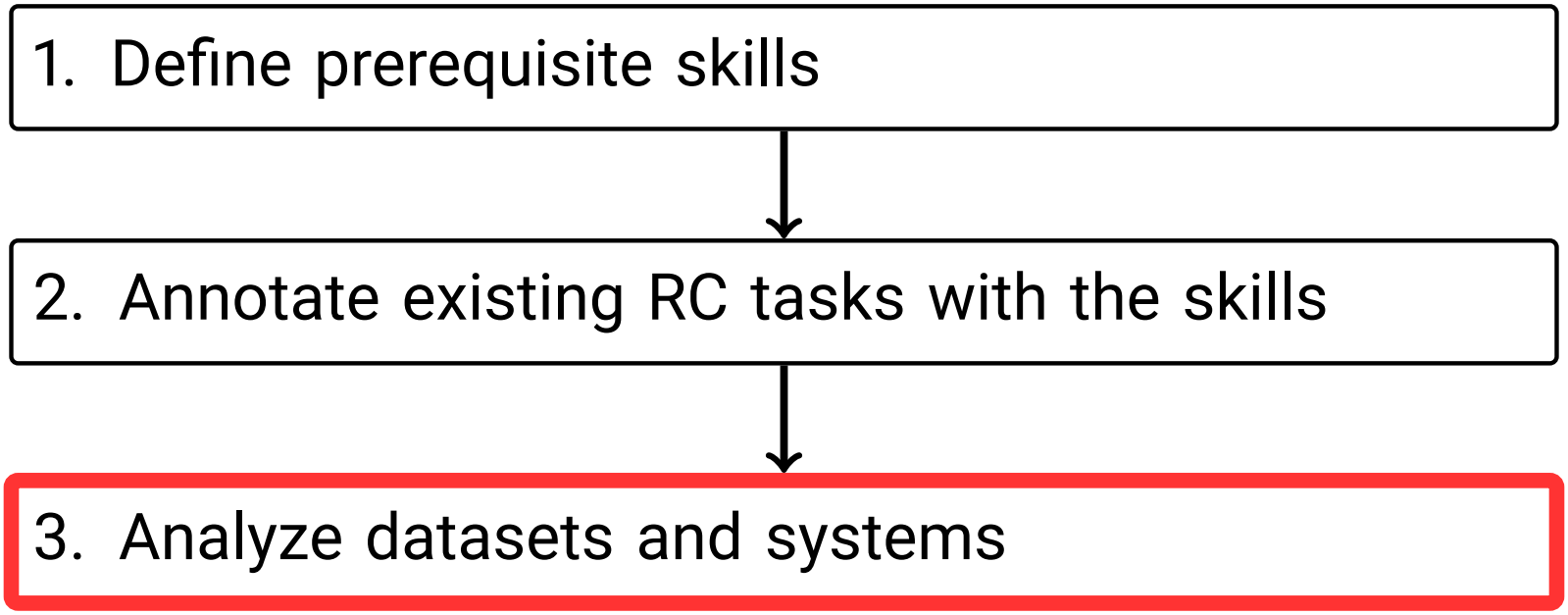
- **Coreference resolution**:
  - *She* in C2 = *the princess* in C1
  - *She* in C3 = *the princess* in C1
- **Temporal relation**:
  - the actions in C1 $\rightarrow$ *wandered out ...* in C2 $\rightarrow$ *went into ...* in C3

Annotation Example in MCTest

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

- **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
- **S/R clause (=complex) sentence** and **Special sentence structure**:
  · C1 = *the princess climbed out ...*
  and [*the princess*] *climbed down ...* (ellipsis)

Our study: A Methodology for Evaluation of Reading Comprehension



Three Analyzed Systems on MCTest

1. Baseline SW+D [Richardson$^+$ 2013]
   - Sliding window + word distance algorithm
2. Smith LexMatch [Smith$^+$ 2015]
   - Baseline + lexical matching method
   - (stemming + question type + coreference)
3. Yin ABCNN [Yin$^+$ 2016]
   - Attention-based CNN without any linguistic features
   - Answers a question as textual entailment

Annotation Results - each skills

Prereq. skills	MCTest Freq.	Accuracy Baseline SW+D	Accuracy Smith LexMatch	Accuracy Yin ABCNN
List/Enumeration	14.7	51.1	65.1	40.4
Mathematical ops.	1.6	20.0	30.0	60.0
Coreference resol.	63.8	52.5	63.6	48.0
Logical reasoning	0.9	100.0	75.0	33.3
Analogy	0.3	0.0	100.0	0.0
Spatiotemporal rel.	27.5	48.9	66.9	45.5
Causal rel.	14.4	45.7	62.0	52.2
Commonsense rsng.	41.9	44.0	61.3	44.8
S/R clause rel.	20.6	50.0	65.9	48.5
Special sentence stru.	8.1	46.2	69.2	46.2
Ave. accuracy	-	50.9	66.2	48.1

Table: Frequencies and accuracies (%) for each prereq. skill.

Annotation Results - each skills

Prereq. skills	MCTest Freq.	Accuracy Baseline SW+D	Smith LexMatch	Yin ABCNN
List/Enumeration	14.7	51.1	65.1	40.4
Mathematical ops.	1.6	20.0	30.0	60.0
Coreference resol.	**63.8**	52.5	63.6	48.0
Logical reasoning	0.9	100.0	75.0	33.3
Analogy	0.3	0.0	100.0	0.0
Spatiotemporal rel.	27.5	48.9	66.9	45.5
Causal rel.	14.4	45.7	62.0	52.2
Commonsense rsng.	**41.9**	44.0	61.3	44.8
S/R clause rel.	20.6	50.0	65.9	48.5
Special sentence stru.	8.1	46.2	69.2	46.2
Avg. accuracy	-	50.9	66.2	48.1

Annotation Results - each skills

Prereq. skills	MCTest Freq.	Accuracy Baseline SW+D	Smith LexMatch	Yin ABCNN
List/Enumeration	14.7	51.1	65.1	40.4
Mathematical ops.	1.6			
Coreference resol.	**63.8**			
Logical reasoning	0.9			
Analogy	0.3			
Spatiotemporal rel.	27.5			
Causal rel.	14.4			
Commonsense rsng.	**41.9**			
S/R clause rel.	20.6			
Special sentence stru.	8.1			
Avg. accuracy	-			

Solving MCTest requires:

- **Coreference resolution**
  → characters in narratives
- **Commonsense reasoning**
  → general knowledge in our social and physical environment

Annotation Results - each skills

Prereq. skills	MCTest Freq.	Accuracy Baseline SW+D	Accuracy Smith LexMatch	Accuracy Yin ABCNN
List/Enumeration	14.7	51.1	65.1	40.4
Mathematical ops.	1.6	20.0	30.0	60.0
Coreference resol.	63.8	52.5	**63.6**	**48.0**
Logical reasoning	0.9	100.0	75.0	33.3
Analogy	0.3	0.0	100.0	0.0
Spatiotemporal rel.	27.5	48.9	66.9	45.5
Causal rel.	14.4	45.7	62.0	52.2
Commonsense rsng.	41.9	**44.0**	**61.3**	**44.8**
S/R clause rel.	20.6	50.0	65.9	48.5
Special sentence stru.	8.1	46.2	69.2	46.2
Avg. accuracy	-	**50.9**	**66.2**	**48.1**

- The three systems are not good at these two skills?

→ At least we can say that “need more development for them”

Annotation Results - Numbers of required skills

		Accuracy		
#Skills	MCTest Freq.	Baseline SW+D	Smith LexMatch	Yin ABCNN
0	10.3	57.6	72.7	54.5
1	28.4	52.7	67.6	47.3
2	28.4	51.6	66.5	50.5
3	23.8	47.4	67.1	46.1
4	8.1	46.2	52.2	42.3
5	0.9	33.3	41.7	33.3

Table: Frequencies and accuracies (%) for required numbers of prereq. skills for each question.

Annotation Results - Numbers of required skills

#Skills	MCTest Freq.	Accuracy Baseline SW+D	Smith LexMatch	Yin ABCNN
0	10.3	57.6	72.7	54.5
1	28.4	52.7	67.6	47.3
2	28.4	51.6	66.5	50.5
3	23.8	47.4	67.1	46.1
4	8.1	46.2	52.2	42.3
5	0.9	33.3	41.7	33.3

- Observation: the more required skills, the more difficult?

Additional Annotation: MCTest vs. SQuAD

Prerequisite skills	MCTest	SQuAD
List/Enumeration	14.7	5.0
Mathematical operations	1.6	0.0
Coreference resolution	63.8	6.2
Logical reasoning	0.9	0.0
Analogy	0.0	0.0
Spatiotemporal relations	27.5	2.5
Causal relations	14.4	6.2
Commonsense reasoning	41.9	86.2
S/R clause sentences	20.6	20.0
Special sentence structure	8.1	25.0

Table: Frequencies and accuracies (%) for each prereq. skill.

Additional Annotation: MCTest vs. SQuAD

Prerequisite skills	MCTest	SQuAD
List/Enumeration	14.7	5.0
Mathematical operations	1.6	0.0
Coreference resolution	63.8	6.2
Logical reasoning	0.9	0.0
Analogy	0.0	0.0
Spatiotemporal relations	27.5	2.5
Causal relations	14.4	6.2
Commonsense reasoning	41.9	86.2
S/R clause sentences	20.6	20.0
Special sentence structure	8.1	25.0

MCTest: narratives with characters and events
SQuAD: expository articles

Additional Annotation: MCTest vs. SQuAD

Prerequisite skills	MCTest	SQuAD
List/Enumeration	14.7	5.0
Mathematical operations	1.6	0.0
Coreference resolution	63.8	6.2
Logical reasoning	0.9	0.0
Analogy	0.0	0.0
Spatiotemporal relations	27.5	2.5
Causal relations	14.4	6.2
Commonsense reasoning	41.9	86.2
S/R clause sentences	20.6	20.0
Special sentence structure	8.1	25.0

SQuAD: written for adults (Wikipedia) and questions are paraphrased by crowdworkers who wrote them

Additional Annotation: MCTest vs. SQuAD

#Skills	MCTest Freq.	SQuAD Freq.
0	10.3	5.0
1	28.4	48.8
2	28.4	37.5
3	23.8	6.2
4	8.1	2.5
5	0.9	0.0
Avg.	1.94	1.52

Table: Frequencies and accuracies (%) for required numbers of prereq. skills for each question.

Additional Annotation: MCTest vs. SQuAD

#Skills	MCTest Freq.	SQuAD Freq.
0	10.3	5.0
1	28.4	48.8
2	28.4	37.5
3	23.8	6.2
4	8.1	2.5
5	0.9	0.0
Avg.	1.94	1.52

Issues & Discussions

- **Mathematical ops.** and **Logical rsng.** were few required
  - → We will try to annotate other (specialized) datasets. (e.g., Aristo and FraCaS) to check if the skill definitions are valid.
- **Analogy** was also few required
  - → Was the definition ambiguous?
- **Commonsense reasoning** needs fine-grained subclasses
  - → We will take *entailment phenomena* into account.
  - → There are many classifications of knowledge and inference in AI/NLP and psychology.

Summary

We proposed an evaluation methodology for RC:

1. Defined ten prerequisite skills in terms of NLU tasks
2. Annotated existing RC tasks with the skills
3. Analyzed datasets and systems

Observation:

- There may be the coreelation of number of required skills in each question and the difficulty of the question.

Thank you!!

- Annotation data is now available!!
  http://www-al.nii.ac.jp/mctest-rcskills-annot/
- Questions, comments:
  sakus(at)is.s.u-tokyo.ac.jp

Textual Entailment and RC

- Textual entailment
  - Recognizing and testing:
    premise $\rightarrow$ hypothesis
- Reading comprehension as textual entailment
  - Recognizing and testing:
    Multiple premises $\rightarrow$ hypothesis

Issue:

Our methodology cannot evaluate the following processes:

- Multiple premises $\leftarrow$ gathered from context sentences
- hypothesis $\leftarrow$ generated from answer candidates

$\Rightarrow$ depending on context length and question formulation

Combinations of prerequisite skills

We can see the weakness of a system from the difference of combinations of skills:

Example

:

- Accuracy for (skill 1, skill 4, skill 5): 60%

  Accuracy for (skill 1, skill 4, skill 7): 80%

  Accuracy for (skill 1, skill 4): 80%

→ not good at skill 5?