



Evaluation Metrics for Machine Reading Comprehension (RC): Prerequisite Skills and Readability

Sugawara et al.

The University of Tokyo, Fujitsu Laboratories Ltd., Natural
Institute of Informatics

Presented by: Shaima AbdulMajeed

RC Task

To give the agent the ability to:

- 1.** Read open-domain documents
- 2.** Answer questions about them

Goal

Knowing Quality of Reading
Comprehension (RC) datasets

Why

To know Which dataset to use that
best evaluates the developed RC
system

RC dataset Example

Nikola Tesla (Serbian Cyrillic: Никола Тесла; 10 July 1856 – 7 January 1943) was a Serbian American inventor, electrical engineer, mechanical engineer, physicist, and futurist best known for his contributions to the design of the modern alternating current (AC) electricity supply system.

In what year was Nikola Tesla born?

Ground Truth Answers: 1856 1856 1856

What was Nikola Tesla's ethnicity?

Ground Truth Answers: Serbian Serbian Serbian

In what year did Tesla die?

Ground Truth Answers: 1943 1943 1943

When was Nikola Tesla born?

Ground Truth Answers: 1856 10 July 1856 1856

Datasets evaluated

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

Current dataset metrics

- Question types
- Answer types
- Categories

Is that enough ?

Does readability of text **correlates**
with difficulty of answering
questions about it?

Evaluation Metrics Proposed

1. Prerequisite skills
2. Readability metrics

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Prerequisite skills

Tracking or grasping of multiple objects

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

*Context: Tom ate apples.
Mary ate apples, too.*

Q: Who ate apples?

A: Tom and Mary

(Object: Tom, Mary)

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Statistical, mathematical and quantitative reasoning

*Context: Tom ate ten apples.
Mary ate eight apples.*

*Q: How many apples did
Tom and Mary eat?*

A: eighteen

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Detection and resolution of all possible demonstratives

*Context: Tom was hungry.
He ate ten apples.*

*Q: How many apples did
Tom eat?*

A: ten (Tom = He)

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Understanding of Predicate Logic

Context: All students have a pen. Tom is a student.

Q: Does Tom have a pen.

A: Yes (and object tracking)

Understanding metaphors

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. **Analogy**
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Context: The White House said Trump is open to ...

Q: Did the President of the United States and his staff say Trump is open to ...

A: Yes

(The White House said = POTUS and his staff said...)

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

“why,” “because,”

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Context: One day, Tom went to the park. After that, he went to the restaurant. Finally, he went to his grandma's house.

Q: Where did Tom go finally?

*A: his grandma's house
(Finally: temporal)*

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Recognizing implicit information

She is a smart student

=

She is a student

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Inference supported by
grammatical and lexical
knowledge

She loves sushi

=

She likes sushi

Inference using known facts, general knowledge

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

*The writer of Hamlet
was Shakespeare*



*Shakespeare wrote
Hamlet*

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Who are the principal characters of the story?

What is the main subject of this article?

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Understanding of complex sentences that have coordination or subordination

Context: Tom has a friend whose name is John.

Q: What is a name of Tom's friend?

A: John (whose = relative clause)

1. Object Tracking
2. Mathematical Reasoning
3. Coreference resolution
4. Logical Reasoning
5. Analogy
6. Causal relation
7. Spatiotemporal relation
8. Ellipsis
9. Bridging
10. Elaboration
11. Meta-Knowledge
12. Schematic clause relation
13. Punctuation

Understanding of punctuation marks

Context: The AFC champion (Denver Broncos) defeated the NFC champion (Carolina Panthers) in super bowl 50

Q: Which NFL team won Super Bowl 50?

A: Denver Broncos

Note: parentheses present the champion team's name

Readability
metrics

1. Lexical Features
2. Syntactic Features
3. Traditional Features

Annotation Procedure (100 Qu)

Step 1: annotators see simultaneously the context, question, and its answer

e.g. Q: Why Tom looked angry? A: His sister ate his cake.

Step 2: Select sentences (from the context)

e.g. Context:

(C1) Tom is a student.

(C2) Tom looks annoyed because his sister ate his cake.

(C3) His sister's name is Sylvia.

-> Select: C2

Step 3: Select skills required for answering the question

e.g.:

C2: Tom looks annoyed because his sister ate his cake.

→ Skill: causal relation ("because"), bridging (lexical knowledge of "annoyed = angry")

Results

1. Prerequisite skills required for each RC dataset
2. Prerequisite skills required per question
3. Readability of each RC dataset
4. Correlation between readability and prerequisite skills required.

Results

1- prerequisite skills required for each RC dataset

1. QA₄MRE (Highest score in all skills):

- Bridging
- Elaboration
- Clause Relation
- Punctuation

2. MCTest

- Casual Relation
- Meta Knowledge
- Coreference resolution
- Spatiotemoral Relation

Results

2- Number of prerequisite skills required per question

	QA ₄ MRE	MCTests	SQuAD	WDW	MS MARCO	News QA
Avg	3.25	1.56	1.28	2.43	1.19	1.99

Highest – technical documents – Qu handcrafted by experts

Results

Nonsense/Difficult Questions

	QA ₄ MRE	MCTest	SQuAD	WDW	MARCO	News QA
Non sense	10	1	3	27	14	1

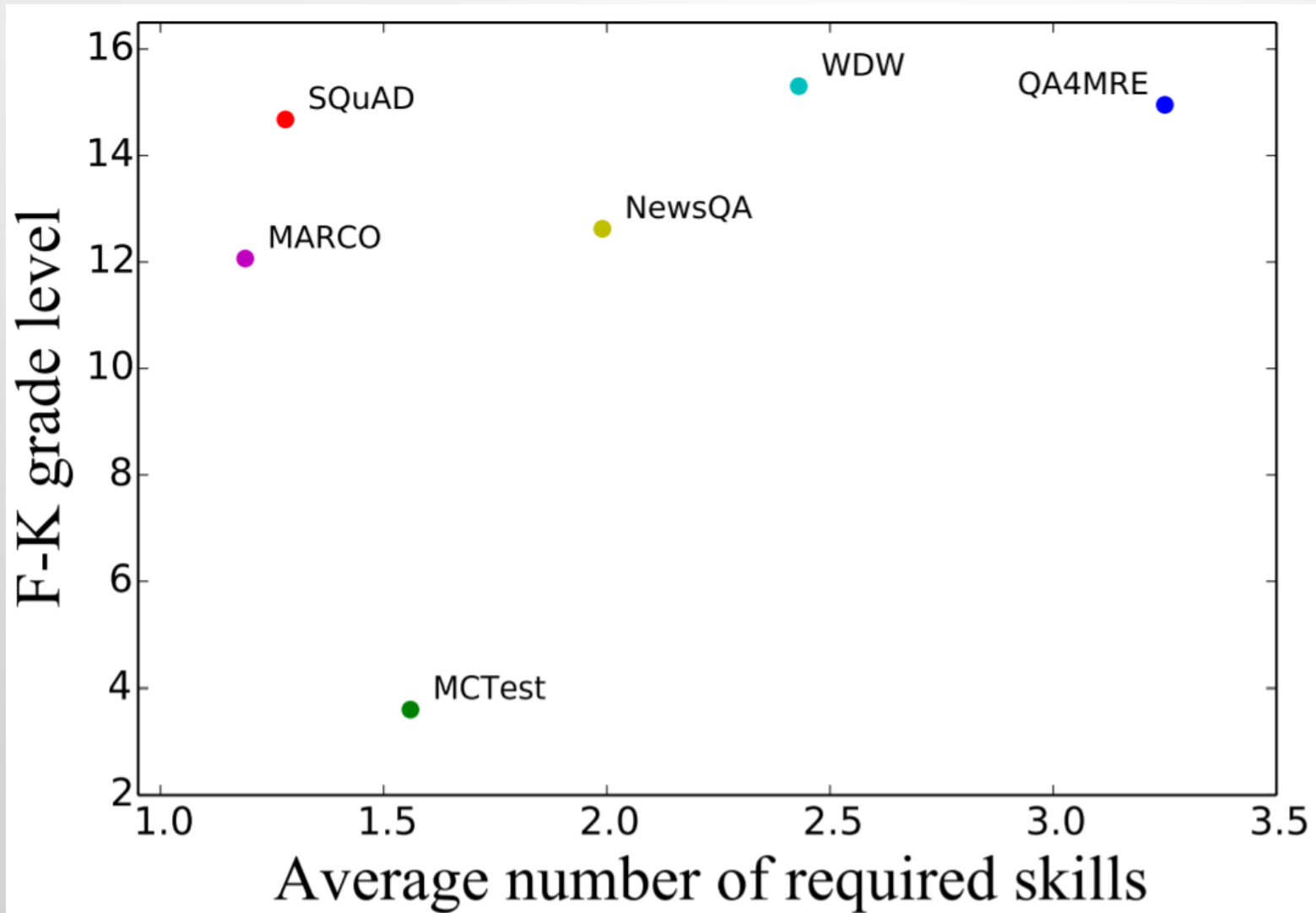
Results

3- Readability metrics for each RC dataset

	QA ₄ MRE	MCTests	SQuAD	WDW	MARCO	News QA
F-K	14.9	3.6	14.6	15.3	12.1	12.6

Results

4- Correlation between readability metrics and the number of required prerequisite skills



Results

4- Correlation between readability metrics and the number of required prerequisite skills

Metrics	<i>r</i>	<i>p</i>	Metrics	<i>r</i>	<i>p</i>
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

Summary

QA₄MRE

- Hard to read
- Hard to answer

MCTest

- Easy to read
- Hard to answer

SQuAD

- Hard to read
- Easy to answer

How to utilize this study

1. Preparing appropriate datasets for each step of RC dev:
 - I. easy-to-read and easy-to-answer
 - II. easy-to-read but difficult-to-answer dataset
 - III. difficult-to-read and difficult-to-answer datasets
2. Apply metrics to evaluate other datasets

Questions?