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
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
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

<table>
<thead>
<tr>
<th>RC dataset</th>
<th>Genre</th>
<th>Query sourcing</th>
<th>Task formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA4MRE (2013)</td>
<td>Technical documents</td>
<td>Handcrafted by experts</td>
<td>Multiple choice</td>
</tr>
<tr>
<td>MCTest (2013)</td>
<td>Narratives by crowd workers</td>
<td>Crowdsourced</td>
<td>Multiple choice</td>
</tr>
<tr>
<td>SQuAD (2016)</td>
<td>Wikipedia articles</td>
<td>Crowdsourced</td>
<td>Text span selection</td>
</tr>
<tr>
<td>MS MARCO (2016)</td>
<td>Segmented web pages</td>
<td>Search engine queries</td>
<td>Description</td>
</tr>
</tbody>
</table>
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
Prerequisite skills

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
Tracking or grasping of multiple objects

Context: Tom ate apples. Mary ate apples, too.
Q: Who ate apples?
A: Tom and Mary
(Object: Tom, Mary)
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
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)
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

**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...)
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Object Tracking</td>
</tr>
<tr>
<td>2.</td>
<td>Mathematical Reasoning</td>
</tr>
<tr>
<td>3.</td>
<td>Coreference resolution</td>
</tr>
<tr>
<td>4.</td>
<td>Logical Reasoning</td>
</tr>
<tr>
<td>5.</td>
<td>Analogy</td>
</tr>
<tr>
<td>6.</td>
<td>Causal relation</td>
</tr>
<tr>
<td>7.</td>
<td>Spatiotemporal relation</td>
</tr>
<tr>
<td>8.</td>
<td>Ellipsis</td>
</tr>
<tr>
<td>9.</td>
<td>Bridging</td>
</tr>
<tr>
<td>10.</td>
<td>Elaboration</td>
</tr>
<tr>
<td>11.</td>
<td>Meta-Knowledge</td>
</tr>
<tr>
<td>12.</td>
<td>Schematic clause relation</td>
</tr>
<tr>
<td>13.</td>
<td>Punctuation</td>
</tr>
</tbody>
</table>

“why,” “because,”
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)
Recognizing implicit information

She is a smart student

= 

She is a student
Inference supported by grammatical and lexical knowledge

She loves sushi

= 

She likes sushi
Inference using known facts, general knowledge

The writer of Hamlet was Shakespeare

Shakespeare wrote Hamlet
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)
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

<table>
<thead>
<tr>
<th></th>
<th>QA4MRE</th>
<th>MCTests</th>
<th>SQuAD</th>
<th>WDW</th>
<th>MS MARCO</th>
<th>News QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg</td>
<td>3.25</td>
<td>1.56</td>
<td>1.28</td>
<td>2.43</td>
<td>1.19</td>
<td>1.99</td>
</tr>
</tbody>
</table>

Highest – technical documents – Qu handcrafted by experts
Results
Nonsense/Difficult Questions

<table>
<thead>
<tr>
<th></th>
<th>QA4MRE</th>
<th>MCTest</th>
<th>SQuAD</th>
<th>WDW</th>
<th>MARCO</th>
<th>News QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsense</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td>27</td>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>
## Results

3- Readability metrics for each RC dataset

<table>
<thead>
<tr>
<th></th>
<th>QA4MRE</th>
<th>MCTests</th>
<th>SQuAD</th>
<th>WDW</th>
<th>MARCO</th>
<th>News QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-K</td>
<td>14.9</td>
<td>3.6</td>
<td>14.6</td>
<td>15.3</td>
<td>12.1</td>
<td>12.6</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>$r$</th>
<th>$p$</th>
<th>Metrics</th>
<th>$r$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NumChar</td>
<td>0.068</td>
<td>0.095</td>
<td>CoOrd</td>
<td>0.166</td>
<td>0.000</td>
</tr>
<tr>
<td>NumSyll</td>
<td>0.057</td>
<td>0.161</td>
<td>Coleman</td>
<td>0.140</td>
<td>0.001</td>
</tr>
<tr>
<td>MLS</td>
<td>0.416</td>
<td>0.000</td>
<td>DC/C</td>
<td>0.188</td>
<td>0.000</td>
</tr>
<tr>
<td>AWL</td>
<td>0.114</td>
<td>0.005</td>
<td>CN/C</td>
<td>0.131</td>
<td>0.001</td>
</tr>
<tr>
<td>ModVar</td>
<td>0.025</td>
<td>0.545</td>
<td>AdvVar</td>
<td>0.026</td>
<td>0.515</td>
</tr>
</tbody>
</table>
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?