An Overview and Analysis of Reading Comprehension Tasks

Saku Sugawara @ Nagoya NLP seminar
2017-06-19

Introduction

Saku Sugawara / 菅原 朔

D1 (2017 Apr. -) @ Dept. of Computer Science, Univ. Tokyo
Aizawa lab. (National Institute of Informatics)

Interests:
natural language understanding, knowledge, inference, ...

saku (at) nii.ac.jp
http://penzant.net

Today

- Overview of reading comprehension (RC) tasks

- Evaluation methodology for reading comprehension
  - Prerequisite skills and MCTest analysis (AAAI 2017) [paper] [slide]
  - Prerequisite skills and readability metrics (ACL 2017) [paper] [slide]

- Observations
  - The more skills required to answer, the more difficult for systems
  - No correlation between “numbers of required skills” and “readability”

Task Example - MCTest (2013)

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

Context + Question + Answer

Task Example

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* → *wandered*)

Short history

- Deep Read (Hirschman 1999) [paper]: Textbooks and WH questions

↓

- 2000s: Question answering / Information retrieval / Textual entailment
  - QA tutorial @ NAACL2016 [slide] / TE @ ACLwiki [web] RTE [paper]

↓

- 2011-2013: Some reading comprehension tasks: QA4MRE, MCTest

↓

- 2015-2016: Large-scale & Automated
  - bAbI (not RC?), CNN/Daily Mail, Children's Book Test, Who-did-What

↓

- 2016-2017: Large-scale & Crowdsourced
  - SQuAD, LAMBADA, MS MARCO, NewsQA

Deep Read (1999)

Library of Congress Has Books for Everyone

(WASHINGTON, D.C., 1964) - It was 150 years ago this year that our nation's biggest library burned to the ground. Copies of all the written books of the time were kept in the Library of Congress. But they were destroyed by fire in 1814 during a war with the British.

That fire didn't stop book lovers. The next year, they began to rebuild the library. By giving it 6,457 of his books, Thomas Jefferson helped get it started.

The first libraries in the United States could be used by members only. But the Library of Congress was built for all the people. From the start, it was our national library.

Today, the Library of Congress is one of the largest libraries in the world. People can find a copy of just about every book and magazine printed.

Libraries have been with us since people first learned to write. One of the oldest to be found dates back to about 800 years B.C. The books were written on tablets made from clay. The people who took care of the books were called "men of the written tablets."

1. Who gave books to the new library?
2. What is the name of our national library?
3. When did this library burn down?
4. Where can this library be found?
5. Why were some early people called "men of the written tablets"?

Figure 1: Sample Remedia™ Reading Comprehension Story and Questions

- Elementary level textbook (60docs w/ 300Qs)
- Wh- questions / bag-of-words system

Recent Trends

QA4MRE (2011-2013) [paper1] [paper2] [web]
MCTest (2013) [paper] [web] (not found now)
bAbI (2015) [paper] [web] [slide]
CNN/Daily Mail (2015) [paper] [web]
Children's Book Test (2015) [paper] [web]
SQuAD (2016) [paper] [web]
LAMBADA (2016) [paper] [web]
Who-did-What (2016) [paper] [web]
MS MARCO (2016) [paper] [web]
NewsQA (2016) [paper] [web]

CLEF QA4MRE (2011-2013)

CLEF 2011-2013: Question Answering for Machine Reading Evaluation [paper1] [paper2] [web]

♠ Source: Technical documents
- Four topics: Alzheimer, AIDS, Climate Change, Music & Society (2013)

♣ Formulation: Multiple choice (5 options including “none of the above”)

♥ Pros
- Hard questions (several question types)
- Detailed analysis / Various languages
- Offering background knowledge
- Auxiliary (relaxed) questions

♦ Cons
- Small questions / Limited topics

CLEF QA4MRE (2011-2013)

```
<reading-test r_id="3">
▼<doc d_id="3">
    Alanna Shaikh: How I'm preparing to get Alzheimer's. I'd like to talk about my dad. My dad has Alzheimer's disease. He
    symptoms about 12 years ago, and he was officially diagnosed in 2005. Now he's really pretty sick. He needs help eat
    getting dressed, he doesn't really know where he is or when it is, and it's been really, really hard. My dad was my hero
    most of my life, and I've spent the last decade watching him disappear. My dad's not alone. There's about 35 million p
    with some kind of dementia, and by 2030 they're expecting that to double to 70 million. That's a lot of people. Demer
    confused faces and shaky hands of people who have dementia, the big numbers of people who get it, they frighten us.
    fear, we tend to do one of two things: We go into denial: "It's not me, it has nothing to do with me, it's never going to
    we decide that we're going to prevent dementia, and it will never happen to us because we're going to do everything
    come▼<q q_id="1">                                                                                    I'm doin
    can d  ▼<q_str>                                                                                      at the
    shoul      What would an Alzheimer relative envisage when practising an internal Chinese martial art?  vants y
    gonn   </q_str>                                                                                      chess,
    <answer a_id="1" correct="Yes">to improve their sense of balance</answer>
    <answer a_id="2">to learn origami</answer>
    <answer a_id="3">to get muscle tremors</answer>
    <answer a_id="4">to have a hobby</answer>
    <answer a_id="5">None of the above</answer>
  </q>
```

MCTest (2013)

By Microsoft Research / EMNLP2013 [paper] [web] (not found now)

♠ Source: Stories written by crowdworkers

♣ Formulation

- Multiple choice (4 options)
- Questions are also written by crowdworkers

♥ Pros

- Story-based RC
  - Characters' intentions, relations of events, commonsense...
- Limited vocabulary (written for children)

♦ Cons

- Not large: 660 stories with 4 questions each

MCTest (2013)

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He 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. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

bAbI tasks (2015)

By Facebook AI Research / ICLR2016 [paper] [web] [slide]

♠ Source: Automatically generated sentences

♣ Formulation
- Context sentences + questions
- 20 tasks

♥ Pros: Analysis is easy
- Many systems are not good at Path Finding

♦ Cons
- Very small vocabulary
- No scalability for further RC

Task 3: Three Supporting Facts

John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 19: Path Finding

The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.
How do you go from den to kitchen? A: west, north
How do you go from office to bathroom? A: north, west

CNN/Daily Mail (2015)

By DeepMind / NIPS2015 [paper] [web]

♠ Source: News articles (CNN/Daily Mail)

♣ Formulation

- Cloze in article titles
- Title is regarded as a summary of its content
- Entities are anonymized

♥ Pros: Large (1.4M) / Multiple topics

♦ Cons: Contain errors (cf. Chen+ (2016) [paper])

- Coreference errors and ambiguous/hard questions

CNN/Daily Mail - Example

Original Version	Anonymised Version
Context	
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query	
Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer	
Oisin Tymon	*ent193*

CNN/Daily Mail - Attention

by *ent423* , *ent261* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* , *ent85* , near *ent312* , a *ent119* official told *ent261* on wednesday . he was identified thursday as special warfare operator 3rd class *ent23* , 29 , of *ent187* , *ent265* . `` *ent23* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused

. . .

ent119 identifies deceased sailor as **X** , who leaves behind a wife

by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday , dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight . *ent164* and *ent21* , who are behind the *ent196* brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

X dedicated their fall fashion show to moms

Children's Book Test (2015)

By Facebook AI Research / ICLR2016 [paper] [web]

♠ Source: Children's books (from Project Gutenberg)

♣ Formulation: Cloze for named entities, nouns, verbs, prepositions

♥ Pros

- Large (688k) / Story-based reading

♦ Cons

- Language modeling task (rather than RC)

	Training	Validation	Test
Number of books	98	5	5
Number of questions	669,343	8,000	10,000
Average words in contexts	465	435	445
Average words in queries	31	27	29
Distinct candidates	37,242	5,485	7,108
Vocabulary size		53,628	

from wikipedia

Children's Book Test (2015)

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. _____ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

SQuAD (2016)

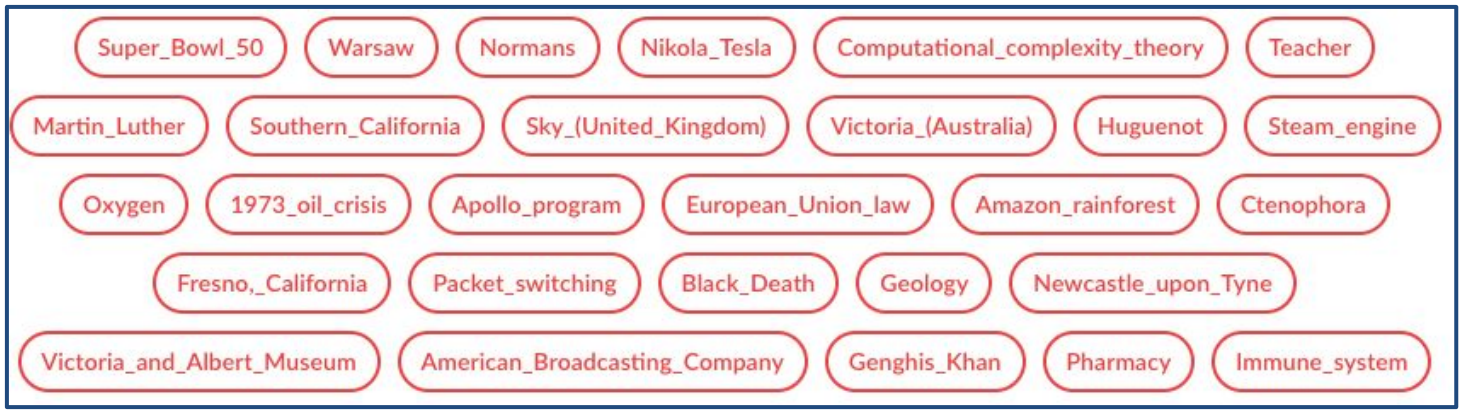
By Stanford / EMNLP2016 [paper] [web]

♠ Source: Wikipedia articles

♣ Formulation: Text span selection

♥ Pros: Large (100k) / Various topics

♦ Cons: Not many questions that require multiple sentence reasoning



SQuAD (2016) - Example

Super_Bowl_50

The Stanford Question Answering Dataset

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Which NFL team represented the AFC at Super Bowl 50?

Ground Truth Answers: Denver Broncos | Denver Broncos | Denver Broncos

Prediction: National Football League

Which NFL team represented the NFC at Super Bowl 50?

Ground Truth Answers: Carolina Panthers | Carolina Panthers | Carolina Panthers

Prediction: National Football League

Where did Super Bowl 50 take place?

LAMBADA (2016)

By Univ. Trento and Univ. Amsterdam / ACL2016 [paper] [web]

♠ Source: Narratives (from Book Corpus by Zhu+ (2015) [paper])

♣ Formulation: Cloze in the last sentence of a context

- More than 80% of passages include target words in the context

♥ Pros: Low baseline (SOTA as of 2017S1: 0.49 by Chu+ (2017) [paper])

♦ Cons

- Too difficult…? / Contextual word prediction task

Context: He shook his head, took a step back and held his hands up as he tried to smile without losing a cigarette. "Yes you can," Julia said in a reassuring voice. "I 've already focused on my friend. You just have to click the shutter, on top, here."
Target sentence: He nodded sheepishly, through his cigarette away and took the _____.
Target word: camera

Who-did-What (2016)

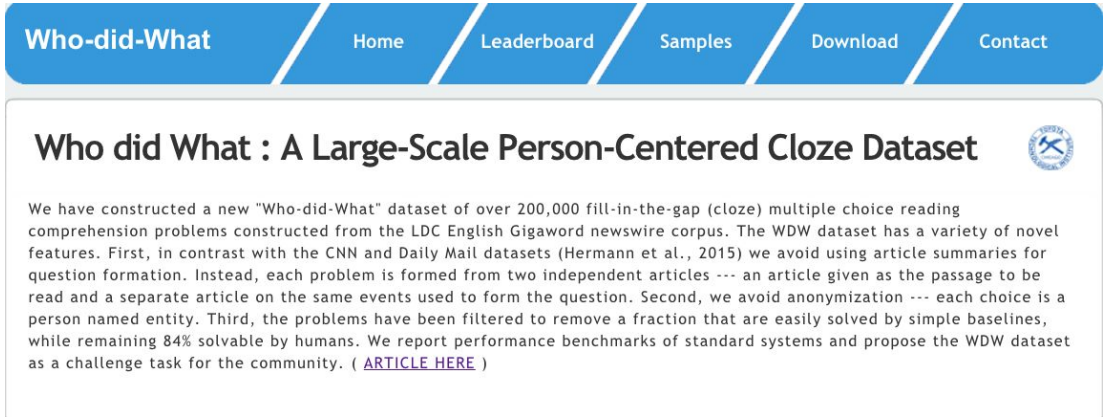
By Toyota Technological Institute at Chicago / EMNLP2016 [paper] [web]

♠ Source: News articles (Gigaword corpus v5)

♣ Formulation: Cloze using two articles regarding the same topic and persons

♥ Pros: Large (200K) / Automatically generated queries

♦ Cons: Including nonsense questions



Who-did-What - Example

Q1

Passage : Britain's decision on Thursday to drop extradition proceedings against Gen. Augusto Pinochet and allow him to return to Chile is understandably frustrating ... Jack Straw, the home secretary, said the 84-year-old former dictator's ability to understand the charges against him and to direct his defense had been seriously impaired by a series of strokes. ... Chile's president-elect, Ricardo Lagos, has wisely pledged to let justice run its course. But the outgoing government of President Eduardo Frei is pushing a constitutional reform that would allow Pinochet to step down from the Senate and retain parliamentary immunity from prosecution. ...

Question : Sources close to the presidential palace said that Fujimori declined at the last moment to leave the country and instead he will send a high level delegation to the ceremony , at which Chilean President Eduardo Frei will pass the mandate to XXX.

Choices : (1) Augusto Pinochet (2) Jack Straw (3) Ricardo Lagos

MS MARCO (2016)

By Microsoft Research [paper] [web]

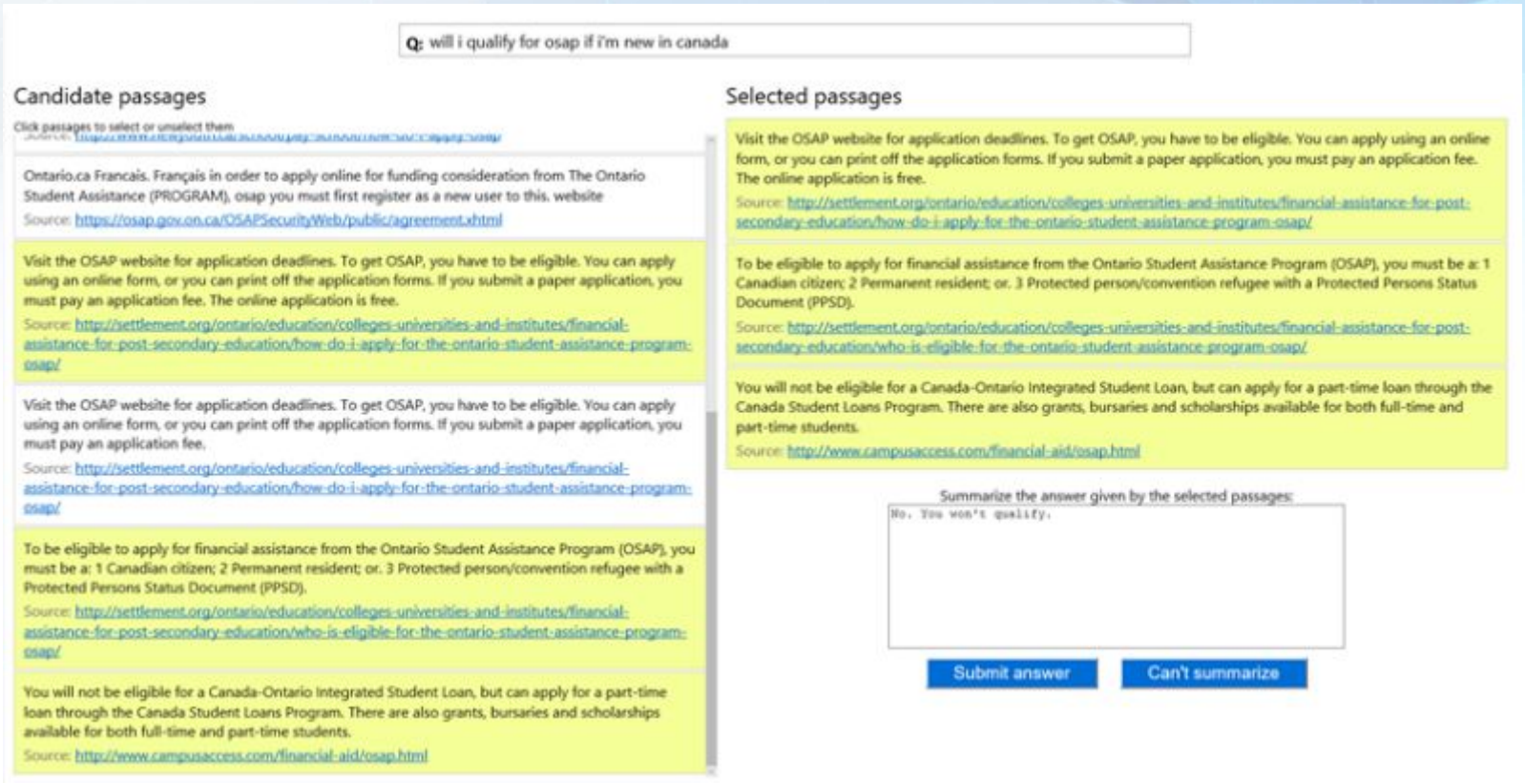
♠ Source
- Search query and web pages
- Human generated answers

♣ Formulation
- Question + some passages
- Five question types: *description*, *numeric*, *entity*, *person*, and *location*

♥ Pros
- Large (100K) / Including candidate (not directly related) passages

♦ Cons
- Difficult to evaluate (not exact match but ROUGE/BLEU)

MS MARCO - Answer generation



NewsQA (2016)

By Maluuba (MS) [paper] [web]

- ♠ Source
  - CNN articles
- ♣ Formulation
  - Crowdsourced QA
  - Text span selection
- ♥ Pros
  - Large (120K) / “More challenging than existing datasets, e.g., SQuAD”
- ♦ Cons
  - Issue remains in human performance (See openreview of the paper)
  - (However, what is “human performance”?)

NewsQA - Example

Immigration Groups to Challenge New Law

(CNN) -- Four groups that advocate for immigrant rights said Thursday they will challenge Arizona's new immigration law, which allows police to ask anyone for proof of legal U.S. residency.

The Mexican American Legal Defense and Educational Fund, the American Civil Liberties Union, the ACLU of Arizona and the National Immigration Law Center held a news conference Thursday in Phoenix to announce the legal challenge.

"The Arizona community can be assured that a vigorous and sophisticated legal challenge will be mounted, in advance of SB1070's implementation, seeking to prevent this unconstitutional and discriminatory law from ever taking

Are there any groups that support the law?
Answer 1: Brewer and others
Answer 2: Four

What do supporters say?
Answer 1: it does not involve racial profiling or any other illegal acts.

Dataset at a Glance

Dataset	Year	Source	Formulation	Query	Size (Q)
QA4MRE	2013	Technical doc	Multiple choice (5 ops)	Experts	240
MCTest	2013	Narrative	Multiple choice (4 ops)	Crowdworkers	2640
bAbI	2015	Simple text	Word selection (context)	Automated	10K (*20)
CNN/Daily Mail	2015	News article	Cloze (word from context)	Automated	1.4M
Children's Book Test	2015	Narrative	Cloze (word from candidates)	Automated	688K
SQuAD	2016	Wikipedia	Text span selection (context)	Crowdworkers	100K
Who-did-What	2016	News article	Cloze (word from context)	Automated	200K
LAMBADA	2016	Narrative	Cloze (word from context*0.8)	Crowdworkers	10K
MS MARCO	2016	Web page	Text span selection* (context)	Search engine	100K
NewsQA	2016	News article	Text span selection (context)	Crowdworkers	120K

Dataset at a Glance

Dataset	Year	Source	Formulation	Query	Size (Q)
QA4MRE	2013	Technical doc	Multiple choice (5 ops)	Experts	240
MCTest	2013	Narrative	Multiple choice (4 ops)	Crowdworkers	2640
bAbI	2015	Simple text	Word selection (context)	Automated	10K (*20)
CNN/Daily Mail	2015	News article	Cloze (word from context)	Automated	1.4M
Children's Book Test	2015	Narrative	Cloze (word from candidates)	Automated	688K
SQuAD	2016	Wikipedia	Text span selection (context)	Crowdworkers	100K
Who-did-What	2016	News article	Cloze (word from context)	Automated	200K
LAMBADA	2016	Narrative	Cloze (word from context*0.8)	Crowdworkers	10K
MS MARCO	2016	Web page	Text span selection* (context)	Search engine	100K
NewsQA	2016	News article	Text span selection (context)	Crowdworkers	120K

Dataset at a Glance

Dataset	Year	Source	Formulation	Query	Size (Q)
QA4MRE	2013	Technical doc	Multiple choice (5 ops)	Experts	240
MCTest	2013	Narrative	Multiple choice (4 ops)	Crowdworkers	2640
bAbI	2015	Simple text	Word selection (context)	Automated	10K (*20)
CNN/Daily Mail	2015	News article	Cloze (word from context)	Automated	1.4M
Children's Book Test	2015	Narrative	Cloze (word from candidates)	Automated	688K
SQuAD	2016	Wikipedia	Text span selection (context)	Crowdworkers	100K
Who-did-What	2016	News article	Cloze (word from context)	Automated	200K
LAMBADA	2016	Narrative	Cloze (word from context*0.8)	Crowdworkers	10K
MS MARCO	2016	Web page	Text span selection* (context)	Search engine	100K
NewsQA	2016	News article	Text span selection (context)	Crowdworkers	120K

Dataset at a Glance

Dataset	Year	Source	Formulation	Query	Size (Q)
QA4MRE	2013	Technical doc	Multiple choice (5 ops)	Experts	240
MCTest	2013	Narrative	Multiple choice (4 ops)	Crowdworkers	2640
bAbI	2015	Simple text	Word selection (context)	Automated	10K (*20)
CNN/Daily Mail	2015	News article	Cloze (word from context)	Automated	1.4M
Children's Book Test	2015	Narrative	Cloze (word from candidates)	Automated	688K
SQuAD	2016	Wikipedia	Text span selection (context)	Crowdworkers	100K
Who-did-What	2016	News article	Cloze (word from context)	Automated	200K
LAMBADA	2016	Narrative	Cloze (word from context*0.8)	Crowdworkers	10K
MS MARCO	2016	Web page	Text span selection* (context)	Search engine	100K
NewsQA	2016	News article	Text span selection (context)	Crowdworkers	120K

Dataset at a Glance

Dataset	Year	Source	Formulation	Query	Size (Q)
QA4MRE	2013	Technical doc	Multiple choice (5 ops)	Experts	240
MCTest	2013	Narrative	Multiple choice (4 ops)	Crowdworkers	2640
bAbI	2015	Simple text	Word selection (context)	Automated	10K (*20)
CNN/Daily Mail	2015	News article	Cloze (word from context)	Automated	1.4M
Children's Book Test	2015	Narrative	Cloze (word from candidates)	Automated	688K
SQuAD	2016	Wikipedia	Text span selection (context)	Crowdworkers	100K
Who-did-What	2016	News article	Cloze (word from context)	Automated	200K
LAMBADA	2016	Narrative	Cloze (word from context*0.8)	Crowdworkers	10K
MS MARCO	2016	Web page	Text span selection* (context)	Search engine	100K
NewsQA	2016	News article	Text span selection (context)	Crowdworkers	120K

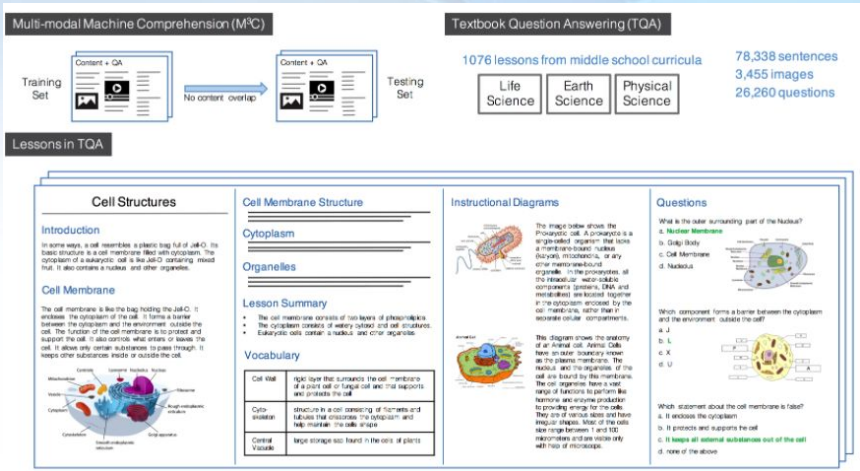
Others...

By Allen Institute for AI (AI2) [web]

- Elementary level science exam questions (Aristo) [web] [paper]
- Textbook Question Answering [web] [paper]
  - Diagrams and images

By Stanford

- Stanford Natural Language Inference Corpus [web] [paper]
  - Automatically generated textual entailment corpus
  - 570k human-written pairs



Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

Others...

- FraCaS (1996) [web]

fracas-005 answer: yes

P1	The really ambitious tenors are Italian.
Q	Are there really ambitious tenors who are Italian?
H	There are really ambitious tenors who are Italian.

fracas-006 answer: no

P1	No really great tenors are modest.
Q	Are there really great tenors who are modest?
H	There are really great tenors who are modest.

fracas-007 answer: yes

P1	Some great tenors are Swedish.
Q	Are there great tenors who are Swedish?
H	There are great tenors who are Swedish.

- Winograd Schema Challenge (2011) [web] [paper]

I. The trophy would not fit in the brown suitcase because it was too **big** (*small*). What was too **big** (*small*)?
Answer 0: the trophy
Answer 1: the suitcase

II. The town councilors refused to give the demonstrators a permit because they **feared** (*advocated*) violence. Who **feared** (*advocated*) violence?
Answer 0: the town councilors
Answer 1: the demonstrators

Issue

- Large dataset
  - How can we analyze our systems?
- Only with accuracy, we cannot tell what the systems understand and what they don’t.

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

Research Question

Q: How can we evaluate and analyze our RC systems?
→ We propose an evaluation methodology for RC

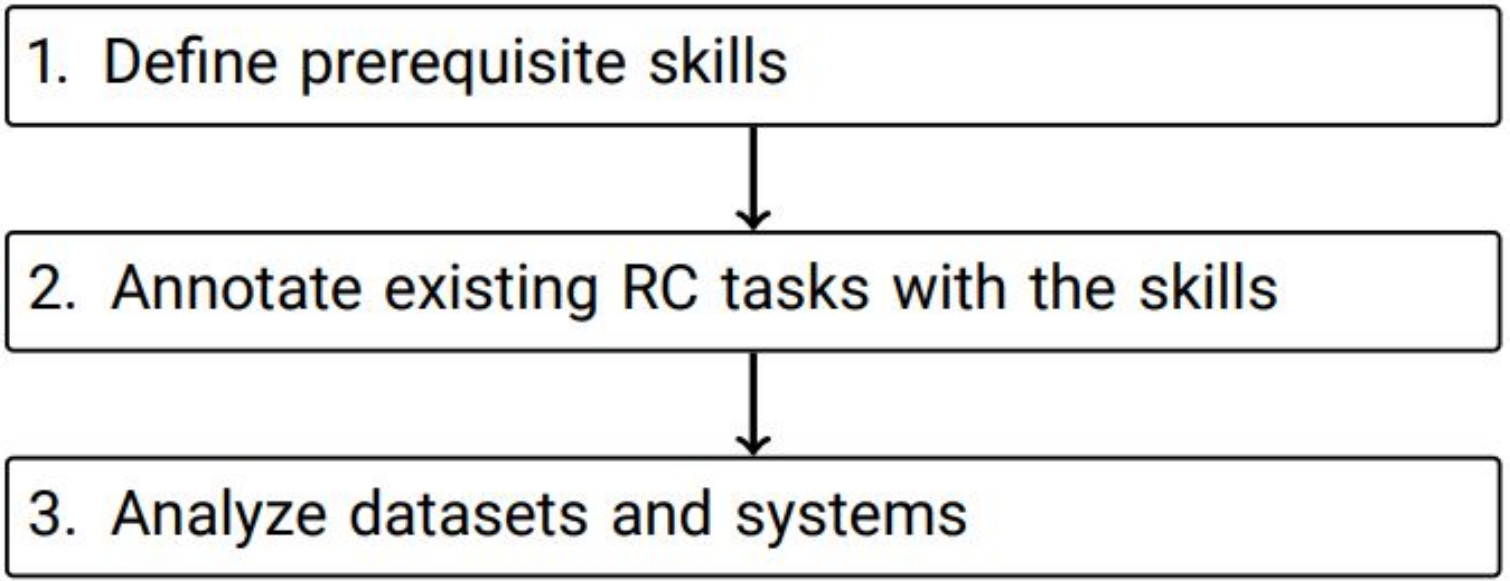
Research Question

Sub-questions:

Q1. How can we measure the *difficulty of RC questions*?
→ Propose "*prerequisite skills*"
→ Analyze one RC dataset and systems

Q2. What is the *relation between the readability of texts and the difficulty of questions*?
→ Use "*readability metrics*" for dataset analysis
→ Refine prereq. skills and analyze multiple RC datasets

Our Methodology



Our Methodology

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

Our Methodology

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%

Our Methodology

2. Annotation RC questions with defined skills

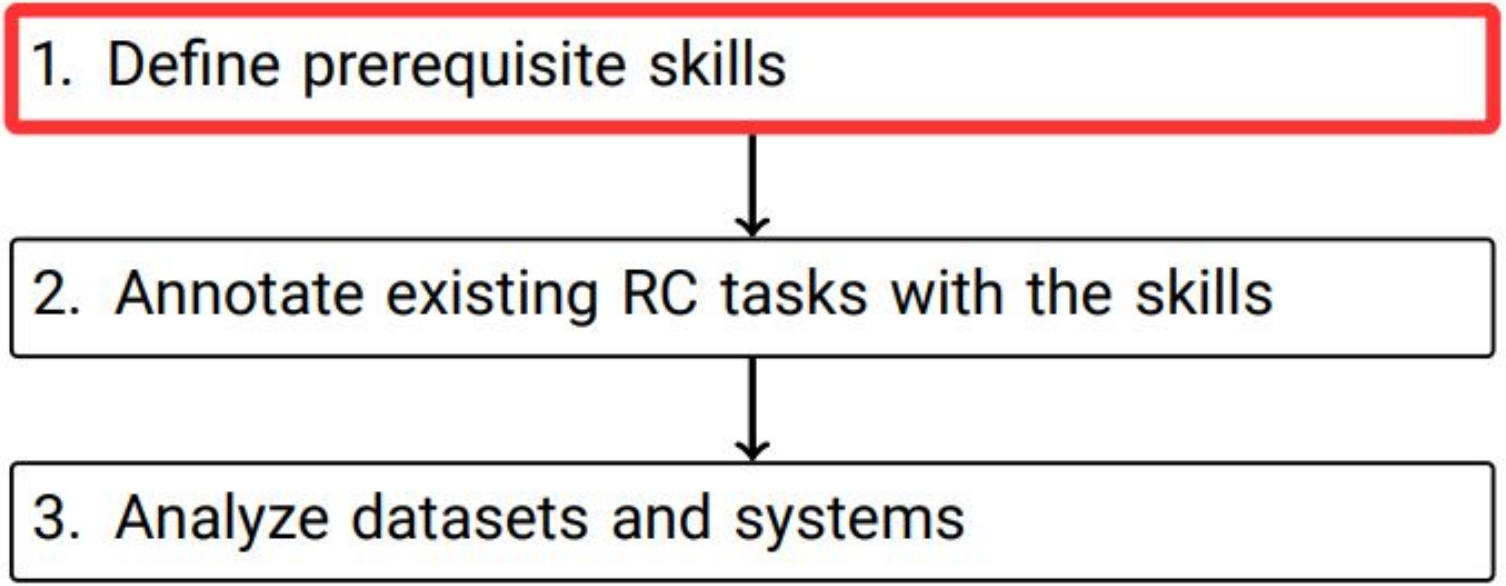
Question	Dataset A Skill 1	Skill 2	...	Skill 10	System X
Q1	1	0	...	1	x
Q2	0	1	...	0	o
Q3	1	1	...	0	x
⋮	⋮	⋮	⋮	⋮	⋮
Q100	1	1	...	1	o
Accuracy	-	-	...	-	75.0%

Our Methodology

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%

Our Methodology



1. Prerequisite Skills

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

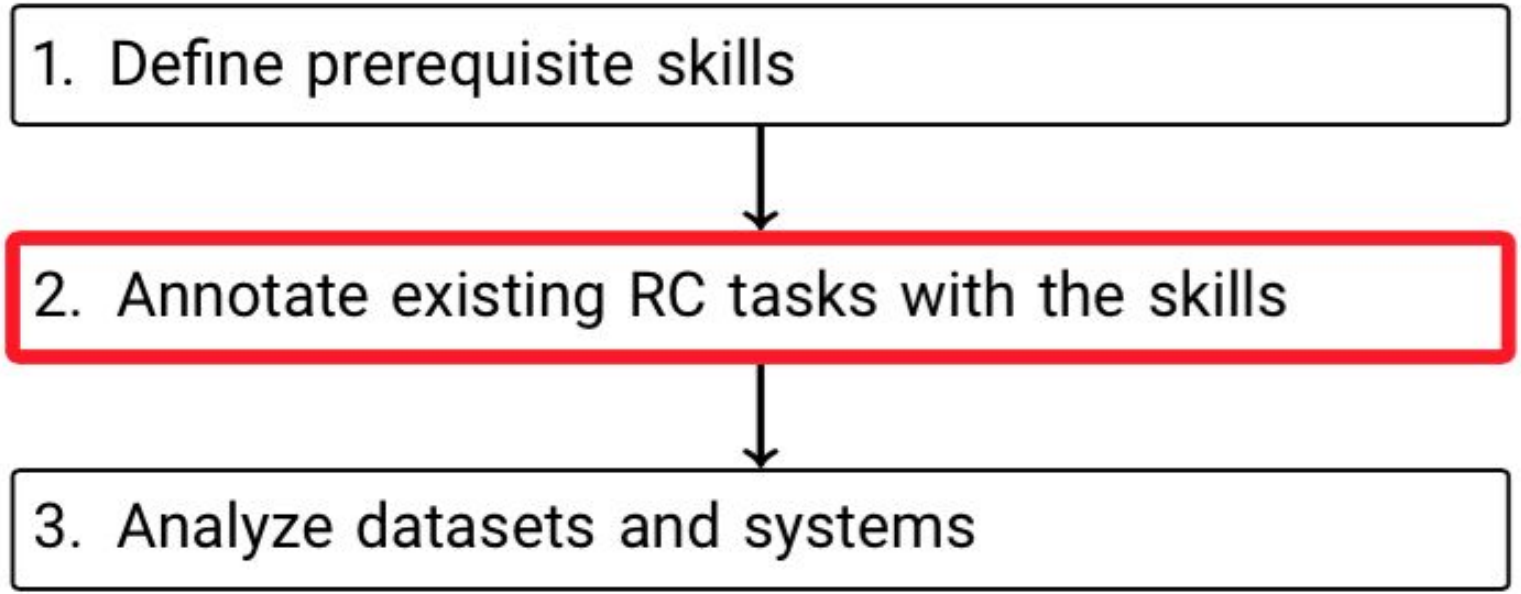
The asterisks (*) with items represent "understanding of."

1. Prerequisite Skills

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

The asterisks (*) with items represent "understanding of."

Our Methodology



2. Annotation: MCTest (2013)

- By Microsoft Research / EMNLP2013
- Source: Stories written by crowdworkers
- Formulation
  - Multiple choice (4 options)
  - Questions are also written by crowdworkers
- Annotation: 320 questions (MC160+MC500)
- Agreement: 85% (two annotators; for sampled questions)

2. Annotation: MCTest (2013)

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

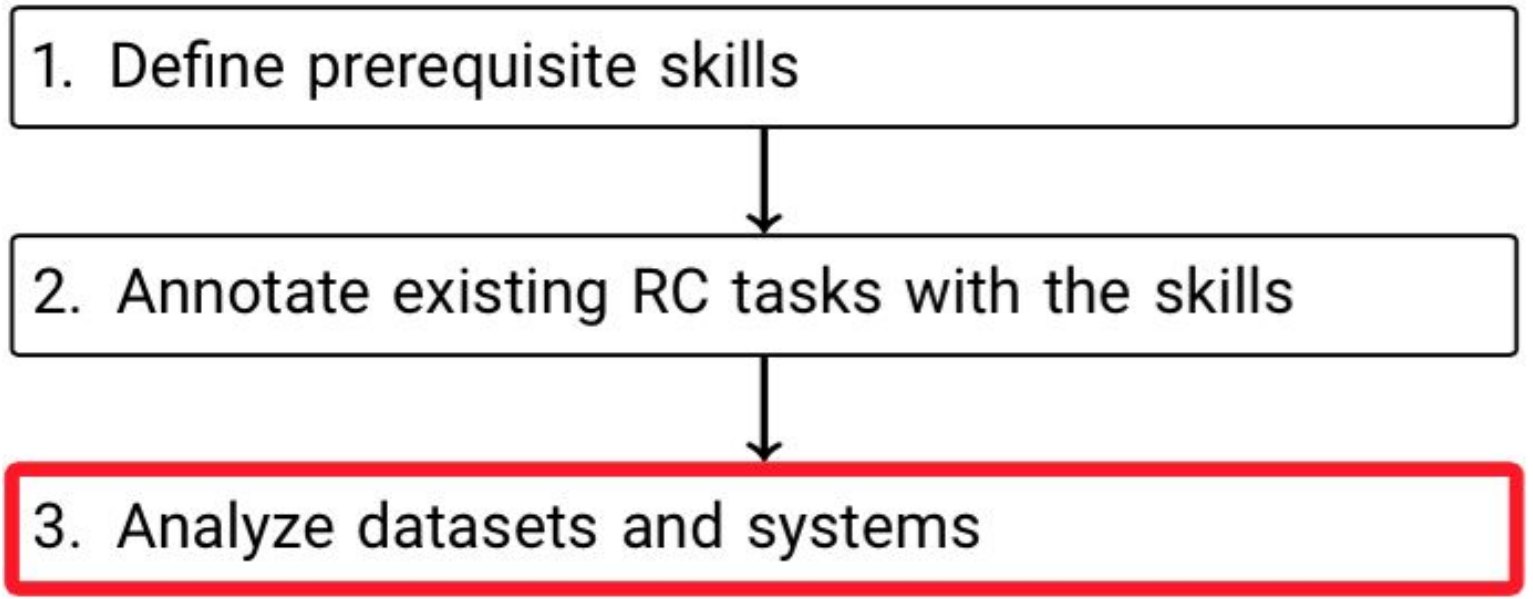
- **Coreference resolution**:
  - *She* in C2 = *the princess* in C1
  - *She* in C3 = *the princess* in C1
- **Temporal relation**:
  - the actions in C1 → *wandered out* ... in C2 → *went into* ... in C3

2. Annotation: MCTest (2013)

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 Methodology



3. Three Analyzed Systems

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

3. Annotation Result / 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
Ave. accuracy	-	50.9	66.2	48.1

3. Annotation Result / 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

3. Annotation Result / 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	-			

MCTest requires:

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

3. Annotation Result / 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**

3. Annotation Result / 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
A				3.1

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

3. Number of Required Skills

#Skills	MCTest Freq.	Accuracy Baseline SW+D	Accuracy Smith LexMatch	Accuracy 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

3. Number of Required Skills

#Skills	MCTest Freq.	Accuracy Baseline SW+D	Accuracy Smith LexMatch	Accuracy 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

→ The more skills, the more difficult?

3. Annotation Result / Skills

#Skills	MCTest Freq.	Accuracy Baseline SW+D	Accuracy Smith LexMatch	Accuracy 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

→ The more skills, the more difficult?

Research Question - Next

Sub-questions:

Q1. How can we measure the *difficulty of RC questions*?
A1. Required numbers of prerequisite skills

Next:

Q2. What is the *relation between the readability of texts and the difficulty of questions*?
→ Use "*readability metrics*" for dataset analysis
→ Refine prereq. skills and analyze multiple RC datasets

Why Readability?

ID: SQuAD (2016), United_Methodist_Church
Context: The United Methodist Church (UMC) practices infant and adult baptism. Baptized Members are those who have been baptized as an infant or child, but who have not subsequently professed their own faith.
Question: What are members who have been baptized as an infant or child but who have not subsequently professed their own faith?
Answer: Baptized Members

ID: MCTest (2013), mc160.dev.8
Context: Sara wanted to play on a baseball team. She had never tried to swing a bat and hit a baseball before. Her Dad gave her a bat and together they went to the park to practice.
Question: Why was Sara practicing?
Answer: She wanted to play on a team

Why Readability?

ID: SQuAD (2016), United_Methodist_Church
Context: The United Methodist Church (UMC) practices infant and adult baptism. **Baptized Members** are those *who have been baptized as an infant or child, but who have not subsequently professed their own faith.*
Question: What are members *who have been baptized as an infant or child but who have not subsequently professed their own f*
Answer: **Baptized Members**

→ Answerble simply by noticing one sentence

ID: MCTest (2013), mc160.dev.8
Context: **Sara wanted to play on a *baseball* team**. She had never tried to swing a bat and hit a baseball before. Her Dad gave her a bat and together they *(= Sara and her Dad)* went to the park to *practice*.
Question: Why was Sara *practicing*?
Answer: **She wanted to play on a team**

→ causal relation, coreference, ellipsis, etc.

Why Readability?

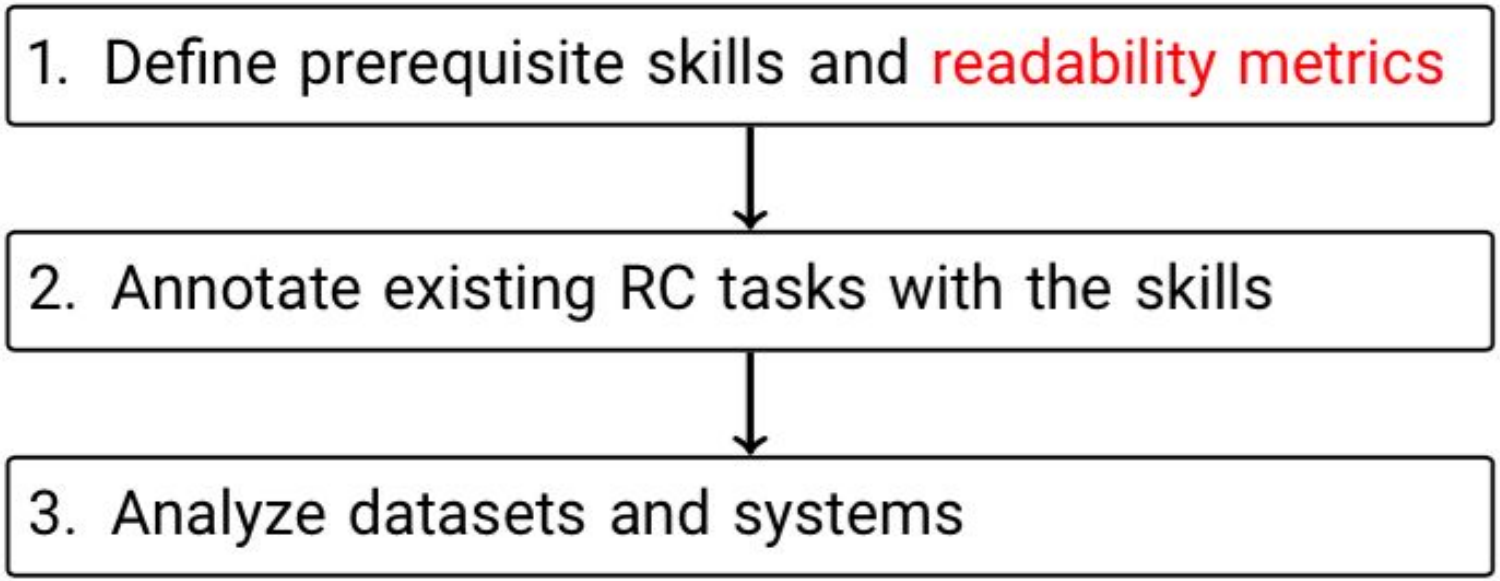
ID: SQuAD (2016), United_Methodist_Church
Context: The United Methodist Church (UMC) practices infant and adult baptism. Baptized Members are those who have been baptized as an infant or child, but who have not subsequently professed their own faith.
Question: What are members who have been baptized as an infant or child but who have not subsequently professed their own faith?
Answer: Baptized Members

Hard-to-read & Easy-to-answer

ID: MCTest (2013), mc160.dev.8
Context: Sara wanted to play on a baseball team. She had never tried to swing a bat and hit a baseball before. Her Dad gave her a bat and together they went to the park to practice.
Question: Why was Sara practicing?
Answer: She wanted to play on a team

Easy-to-read & Hard-to-answer

Prereq. Skills and Readability



Readability Metrics

10 readability metrics from Vajjala and Meurers (2012) [paper]

- Ave. Num. of characters per word (*NumChar*)
- Ave. Num. of syllables per word (*NumSyll*)
- Ave. sentence length in words (*MLS*)
- Proportion of words in AWL (*AWL*)
- Modifier variation (*ModVar*)
- Num. of coordinate phrases per sentence (*CoOrd*)
- Coleman-Liau index (*Coleman*)
- Dependent clause to clause ratio (*DC/C*)
- Complex nominals per clause (*CN/C*)
- Adverb variation (*AdvVar*)

Readability Metrics

	Dataset A								System X
Question	Skill 1	Skill 2	⋯	Skill 10	*NumChar*	*MLS*	⋯	*ModVar*	
Q1	x	-	⋯	x	5.1	27.1	⋯	0.17	x
Q2	-	o	⋯	-	3.9	13.5	⋯	0.11	o
Q3	x	x	⋯	-	4.6	26.9	⋯	0.08	x
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Q100	o	o	⋯	o	4.3	16.9	⋯	0.12	o
Accuracy	40.0%	90.0%	⋯	70.0%	-	-	⋯	-	75.0%

Refined 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. Schematics clause relation
13. Punctuation

Four Reasoning Skills

- Ellipsis
  - Recognizing implicit/omitted information (argument, predicate, quantifier, time, or place)
  - e.g. *She is a smart student → She is a student*

- Bridging
  - Inferences between two facts supported by grammatical and conceptual knowledge (synonymy, hypernymy, thematic role, part of events, idioms, and apposition)
  - e.g. *She loves sushi. → She likes sushi.*

Four Reasoning Skills

- Elaboration
  - Inference using known facts, general knowledge (e.g., kinship, exchange, typical event sequence, and naming), and implicit relations (e.g., noun compounds and possessives)
  - e.g. *The writer of Hamlet was Shakespeare* → *Shakespeare wrote Hamlet*

- Meta-knowledge
  - Inference using external knowledge including a reader, writer, character, and genre
  - e.g. *Who is the main character in this story?*

Analyzed Six Datasets

RC dataset	Genre	Query sourcing	Task formulation
QA4MRE	Technical documents	Handcrafted by experts	Multiple choice
MCTest	Narratives by crowd workers	Crowd sourced	Multiple choice
SQuAD	Wikipedia articles	Crowd sourced	Text span selection
Who-did-What	News articles (Gigaward v5)	Automated from other articles	Cloze
MS MARCO	Segmented web pages	Search engine queries	Description
NewsQA	News articles	Crowd sourced	Text span selection

Annotation:
6*100 questions
13 skills
4 annotators
90% agreement

Skill Frequencies (%)

Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1. Tracking	14.3	7.1	7.1	10.0	7.1	4.3
2. Math.	5.7	0.0	0.0	5.7	0.0	1.4
3. Coref. resol.	30.0	51.4	14.3	25.7	12.9	24.3
4. Logical rsng.	20.0	2.9	0.0	12.9	1.4	2.9
5. Analogy	10.0	0.0	0.0	11.4	0.0	4.3
6. Causal rel.	1.4	7.1	0.0	2.9	0.0	2.9
7. Sptemp rel.	37.1	8.6	5.7	7.1	0.0	4.3
8. Ellipsis	15.7	4.3	4.3	17.1	2.9	4.3
9. Bridging	81.4	24.3	50.0	72.9	21.4	47.1
10. Elaboration	77.1	10.0	14.3	67.1	18.6	35.7
11. Meta	0.0	1.4	0.0	0.0	0.0	0.0
12. Clause rel.	55.7	44.3	28.6	37.1	31.4	31.4
13. Punctuation	45.7	1.4	28.6	24.3	15.7	25.7
Nonsense	5.7	1.4	2.9	15.7	14.3	0.0

Skill Frequencies (%)

Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1. Tracking	14.3	7.1	7.1	10.0	7.1	4.3
2. Math.	5.7	0.0	0.0	5.7	0.0	1.4
3. Coref. resol.	30.0	**51.4**	14.3	25.7	12.9	24.3
4. Logical rsng.	20.0	2.9				
5. Analogy	10.0	0.0				
6. Causal rel.	1.4	7.1				
7. Sptemp rel.	37.1	8.6				
8. Ellipsis	15.7	4.3	4.3	17.1	2.9	4.3
9. Bridging	81.4	24.3	50.0	72.9	21.4	47.1
10. Elaboration	77.1	10.0	14.3	67.1	18.6	35.7
11. Meta	0.0	1.4	0.0	0.0	0.0	0.0
12. Clause rel.	55.7	44.3	28.6	37.1	31.4	31.4
13. Punctuation	45.7	1.4	28.6	24.3	15.7	25.7
Nonsense	5.7	1.4	2.9	15.7	14.3	0.0

MCTest:
Narratives with characters

Skill Frequencies (%)

Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1. Tracking	14.3	7.1	7.1	10.0	7.1	4.3
2. Math.	5.7	0.0	0.0	5.7	0.0	1.4
3. Coref. resol.	30.0	51.4	14.3	25.7	12.9	24.3
4. Logical rsng.	20.0	2.9	0.0	12.9	1.4	2.9
5. Analogy	10.0	0.0	0.0	11.4	0.0	4.3
6. Causal rel.	1.4	7.1	0.0	2.9	0.0	2.9
7. Sptemp rel.	37.1	8.6	5.7	7.1	0.0	4.3
8. Ellipsis	15.7	4.3	4.3	17.1	2.9	4.3
9. Bridging	**81.4**	24.3	50.0	72.9	21.4	47.1
10. Elaboration	**77.1**	10.0	14.3	67.1	18.6	35.7
11. Meta	0.0					
12. Clause rel.	55.7					
13. Punctuation	45.7					
Nonsense	5.7	1.4	2.9	15.7	14.3	0.0

QA4MRE:
Questions are written by experts

Skill Frequencies (%)

Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1. Tracking	14.3	7.1	7.1	10.0	7.1	4.3
2. Math.	5.7	0.0	0.0	5.7	0.0	1.4
3. Coref. resol.	30.0	51.4	14.3	25.7	12.9	24.3
4. Logical rsng.	20.0	2.9	0.0	12.9	1.4	2.9
5. Analogy	10.0	0.0	0.0	11.4	0.0	4.3
6. Causal rel.	1.4					
7. Sptemp rel.	37.1					
8. Ellipsis	15.7					
9. Bridging	81.4					
10. Elaboration	77.1					
11. Meta	0.0					
12. Clause rel.	55.7					
13. Punctuation	45.7	1.4	28.6	24.3	15.7	25.7
Nonsense	5.7	1.4	2.9	**15.7**	**14.3**	0.0

Who did What (WDW):
Automatically generated queries

MS MARCO:
Search queries and web pages

Number of Required Skills

#Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
0	1.4	15.7	20.0	1.4	20.0	15.7
1	5.7	37.1	34.3	7.1	32.9	27.1
2	11.4	21.4	27.1	11.4	27.1	27.1
3	20.0	18.6	8.6	27.1	4.3	18.6
4	15.7	2.9	2.9	17.1	2.9	7.1
5	18.6	1.4	1.4	8.6	0.0	2.9
6	14.3	1.4	0.0	8.6	0.0	1.4
7	1.4	0.0	2.9	2.9	0.0	0.0
8	1.4	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0
10	4.3	0.0	0.0	0.0	0.0	0.0
Ave.	3.94	1.63	1.53	2.94	1.11	1.89

Number of Required Skills

#Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
0	1.4	15.7	20.0	1.4	20.0	15.7
1	5.7	37.1	34.3	7.1	32.9	27.1
2	11.4	21.4	27.1	11.4	27.1	27.1
3	20.0	18.6	8.6	27.1	4.3	18.6
4	15.7	2.9	2.9	17.1	2.9	7.1
5	18.6	1.4	1.4	8.6	0.0	2.9
6	14.3	1.4	0.0	8.6	0.0	1.4
7	1.4	0.0	2.9	2.9	0.0	0.0
8	1.4	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0
10	4.3	0.0	0.0	0.0	0.0	0.0
Ave.	3.94	1.63	1.53	2.94	1.11	1.89

Readability Metrics

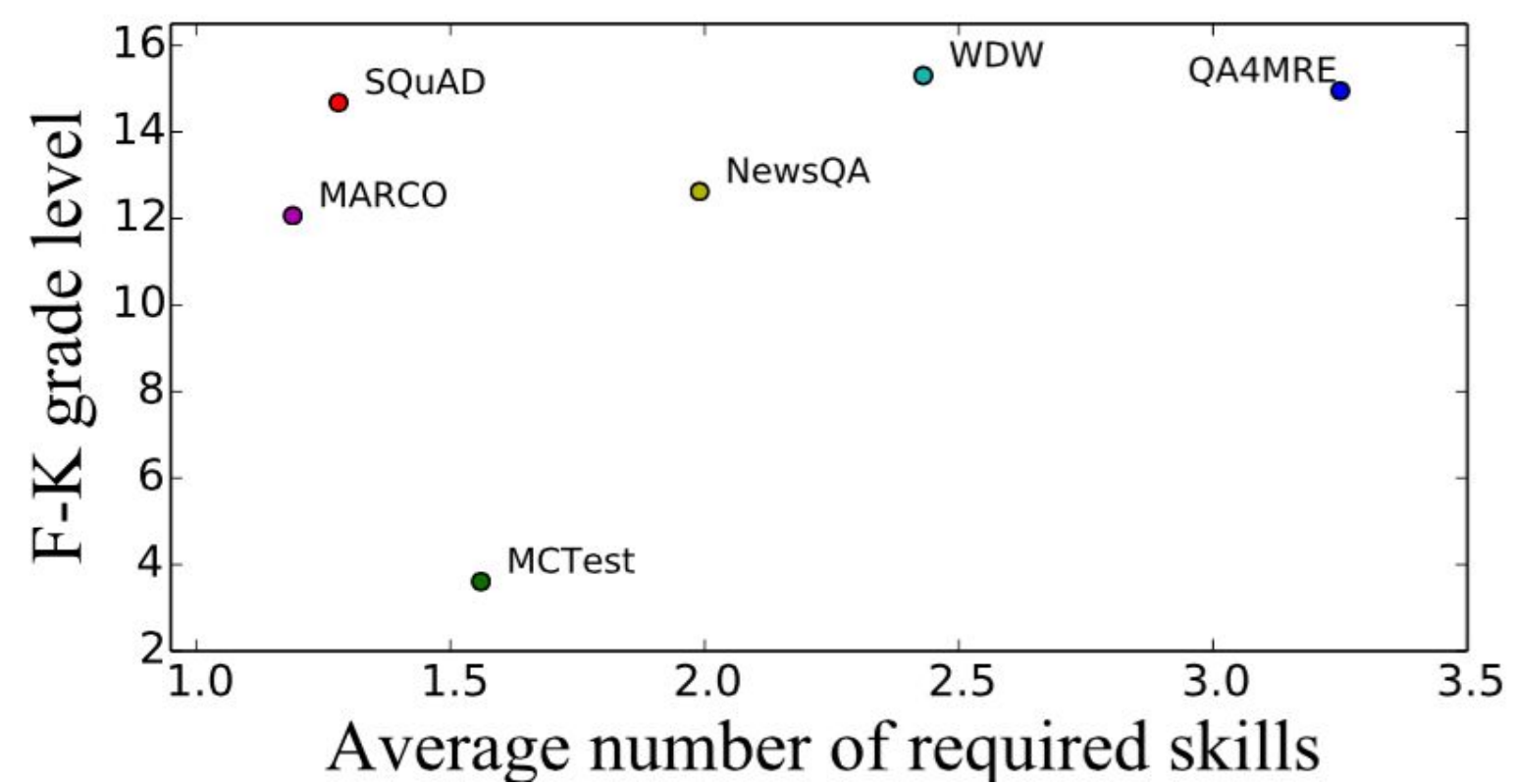
Metrics	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
NumChar	5.199	3.899	5.339	4.991	5.019	4.984
NumSyll	1.721	1.249	1.787	1.660	1.698	1.623
MLS	31.743	11.887	26.064	25.598	19.898	22.833
AWL	0.077	0.004	0.069	0.033	0.046	0.038
ModVar	0.167	0.117	0.175	0.153	0.158	0.139
CoOrd	1.074	0.299	0.857	0.412	0.636	0.514
Coleman	13.755	4.413	13.927	12.075	11.792	11.922
DC/C	0.368	0.228	0.266	0.234	0.234	0.280
CN/C	2.306	0.578	1.973	2.112	1.908	1.647
AdvVar	0.032	0.037	0.030	0.019	0.017	0.020
Flesch*	16.943	3.637	15.634	13.935	12.167	12.449
Words	1557.2	176.7	137.8	268.6	73.6	619.7

Readability Metrics

Metrics	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
NumChar	5.199	3.899	5.339	4.991	5.019	4.984
NumSyll	1.721	1.249	1.787	1.660	1.698	1.623
MLS	31.743	11.887	26.064	25.598	19.898	22.833
AWL	0.077	0.004	0.069	0.033	0.046	0.038
ModVar	0.167	0.117	0.175	0.153	0.158	0.139
CoOrd						
Coleman						
DC/C						
CN/C						
AdvVar	0.032	0.037	0.030	0.019	0.017	0.020
Flesch*	16.943	3.637	15.634	13.935	12.167	12.449
Words	1557.2	176.7	137.8	268.6	73.6	619.7

Flesch
= Flesch-Kincaid grade level
= the number of years required to understand the text

Skills and Readability



Skills and Readability

Metrics	*r*	*p*
NumChar	0.067	0.161
NumSyll	0.057	0.235
MLS	0.411	0.000
AWL	0.160	0.001
ModVar	0.063	0.189
CoOrd	0.194	0.000
Coleman	0.147	0.002
DC/C	0.174	0.000
CN/C	0.167	0.000
AdvVar	0.007	0.882
Flesch	0.348	0.000

Table: Pearson's correlation coefficients (r) with the p-values (p) in all RC datasets

Obsearvation

There is only a weak correlation between readability metrics and numbers of required skills

- "*Difficult to read*" □does not mean "*difficult to answer*"
- It is possible to create a dataset that consists of an easy-to-read context and difficult-to-answer questions

Q2. What is the *relation between the readability of texts and the difficulty of questions*?
A2. There is only a weak correlation between them

Textual Entailment and RC

Textual entailment (cf. RTE, FraCaS, SNLI)
Premise(s) → hypothesis

Reading comprehension
Multiple premises in context → hypothesis as Q+A

Our methodology cannot evaluate the following processes in RC:

- Multiple premises
  - Skill for gathering premises from context sentences
- Hypothesis
  - Skill for choosing and generating hypothesis from answer candidates

Research Question Summary

Sub-questions:

Q1. How can we measure the *difficulty of RC questions*?
A1. Required numbers of prerequisite skills

Q2. What is the *relation between the readability of texts and the difficulty of questions*?
A2. There is only a weak correlation between them

Summary

- Overview of reading comprehension (RC) tasks
  - (Deep Read,) QA4MRE, MCTest, bAbI, CNN/Daily Mail, CBT, SQuAD, LAMBADA, Who-did-What, MS MARCO, NewsQA ----- 10 datasets!

- Evaluation methodology for reading comprehension
  - Prerequisite skills and system/dataset analysis (AAAI 2017)
  - Prerequisite skills and readability metrics (ACL 2017)

- Observations
  - The more skills required to answer, the more difficult for systems
  - No correlation between “numbers of required skills” and “readability”