Evaluation Metrics for the Machine Reading Comprehension Task

Saku Sugawara

University of Tokyo

October 30, 2017 @ Google Tokyo

http://penzant.net

Agenda

1. Overview of the machine reading comprehension (RC) task
   “What is the RC task?”
2. Evaluation methodology for RC datasets/systems
   “How can we evaluate our systems/datasets?”
   Sugawara^{+} (2017a, AAAI) & Sugawara^{+} (2017b, ACL)
3. Discussion for constructing RC datasets:
   “How can we create *difficult but not too difficult* questions?”

Agenda

1. Overview of the machine reading comprehension (RC) task
   "What is the RC task?"
2. Evaluation methodology for RC datasets/systems
   "How can we evaluate our systems/datasets?"
   Sugawara^{+} (2017a, AAAI) & Sugawara^{+} (2017b, ACL)
3. Discussion for constructing RC datasets:
   "How can we create *difficult but not too difficult* questions?"

Reading Comprehension (RC) Task

ID: MCTest MC160.dev.29 (1) multiple:
Context: The princess climbed out the window of the high tower and climbed down the south wall when her mother was sleeping. She wandered out a good ways. Finally she went into the forest where there are no electric poles but where there are some caves.
Question: Where did the princess wander to after escaping?
Answer: A) Mountain *B) **Forest** C) Cave D) Castle

Task: context, question, and answer

Reading Comprehension (RC) Task

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*)

Definitions

Towards the Machine Comprehension of Text: An Essay (Burges 2013)

> "a machine comprehends a passage of text if, for any question regarding that text that can be answered correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question"

Overview of QA4MRE Main Task at CLEF 2013 (Sutcliffe^{+} 2013)

> "RC tests do not require only semantic understanding but they assume a cognitive process which involves using implications and presuppositions, retrieving the stored information, performing inferences to make implicit information explicit."

$\rightarrow$ the RC task is an evaluation method for language understanding systems in terms of their behavior, and tests a cognitive process that involves several skills, such as performing inferences using background knowledge, by letting the system answer questions about a given text.

Representative Datasets

- MCTest (2013)
- CNN/Daily Mail (2015)
- SQuAD (2016)

MCTest (2013)

- Richardson et al.
- Children stories and questions written by crowdworkers
- Multiple choice
- Pros:
  - Story-based RC = characters' intentions, relations of events, commonsense...
  - Limited vocabulary
- Cons
  - Not large: 660 stories with 4 questions each

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

CNN/Daily Mail (2015)

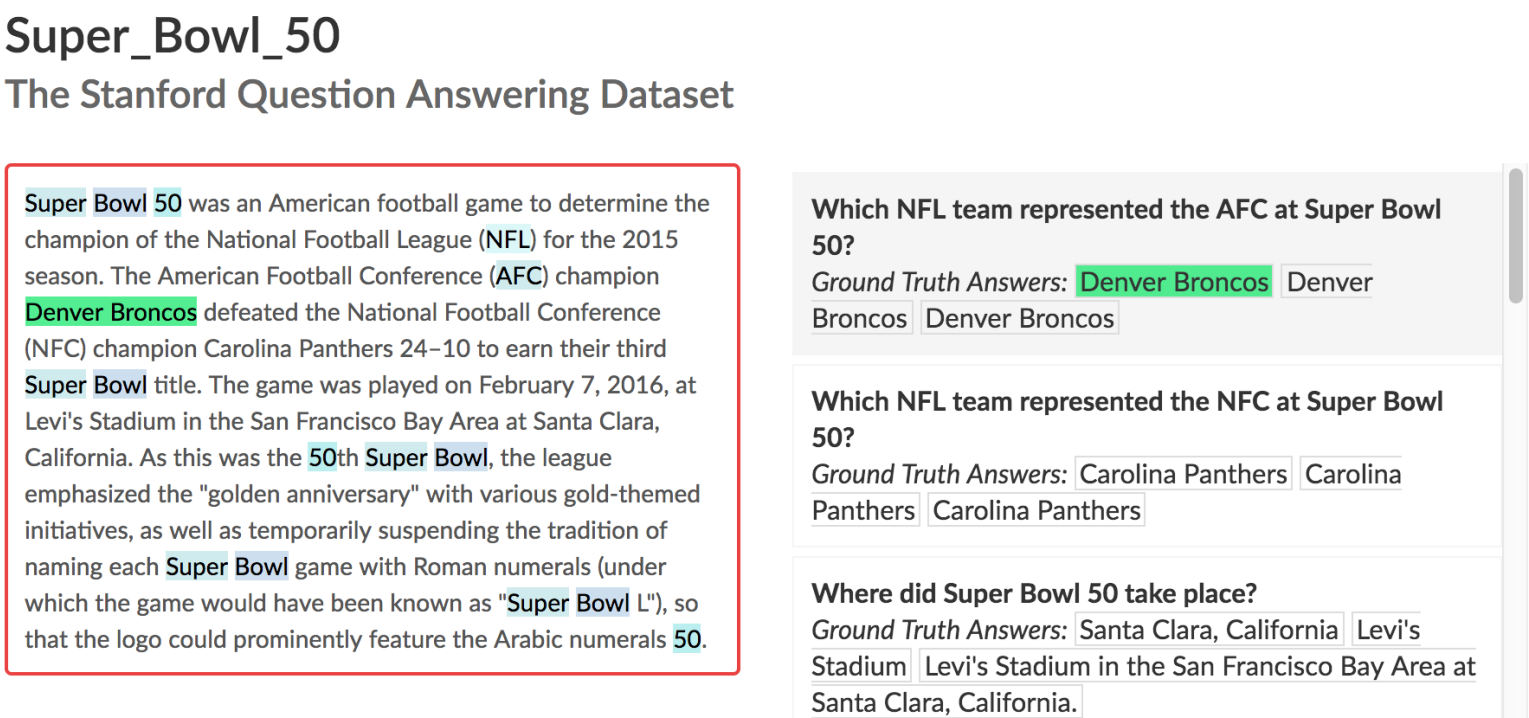
- Hermann et al.
- News articles
- Cloze (fiil-in-blank)
- Answer extraction

- Pros
  - 140M Qs, various topics
- Cons
  - Contains errors in coreference (Chen + 2016)

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*

SQuAD (2016)

- Rajpurkar et al.
- Wikipedia articles
- Answer extraction
- Pros
  - 100K Qs, various topics
- Cons
  - Goldberg (2017) says “pattern matching”
  - Easily fooled by Jia+ (2017)’s adversarial examples



Datasets

Dataset	Year	Question/ Answer	Given text	Genre or Source	Question sourcing	Question size	Reader
QA4MRE	2013	Complete/ Multichoice	Document	Technical documents	Expert	284	Student
MCTest	2013	Complete/ Multichoice	Paragraphs	Crafted story	Crowd-worker	2640	Child
CNN/ Daily Mail	2015	Cloze/ Extraction	Document	News article	Auto-mated	1.4M	Adult
SQuAD	2016	Complete/ Extraction	Paragraph	Wikipedia	Crowd-worker	100K	Adult
Who-did-What	2016	Cloze/ Extraction	Document	News article	Auto-mated	200K	Adult
MS MARCO	2016	Complete/ Extraction	Paragraphs	Web page	Search engine	100K	Adult
NewsQA	2016	Complete/ Extraction	Document	News article	Crowd-worker	120K	Adult
TriviaQA	2017	Complete/ Extraction	Paragraph	Wikipedia/ Web page	Quiz website	650K	Adult
RACE	2017	Complete/ Multichoice	Document	Exam	Expert	100K	Student

Datasets

Dataset	Year	Question/ Answer	Given text	Genre or Source	Question sourcing	Question size	Reader
QA4MRE	2013	Complete/ Multichoice	Document	Technical documents	Expert	284	Student
MCTest	2013	Complete/ Multichoice	Paragraphs	Crafted story	Crowd-worker	2640	Child
CNN/ Daily Mail	2015	Cloze/ Extraction	Document	News article	Auto-mated	1.4M	Adult
SQuAD	2016	Complete/ Extraction	Paragraph	Wikipedia	Crowd-worker	100K	Adult
Who-did-What	2016	Cloze/ Extraction	Document	News article	Auto-mated	200K	Adult
MS MARCO	2016	Complete/ Extraction	Paragraphs	Web page	Search engine	100K	Adult
NewsQA	2016	Complete/ Extraction	Document	News article	Crowd-worker	120K	Adult
TriviaQA	2017	Complete/ Extraction	Paragraph	Wikipedia/ Web page	Quiz website	650K	Adult
RACE	2017	Complete/ Multichoice	Document	Exam	Expert	100K	Student

Datasets

Dataset	Year	Question/ Answer	Given text	Genre or Source	Question sourcing	Question size	Reader
QA4MRE	2013	Complete/ Multichoice	Document	Technical documents	Expert	284	Student
MCTest	2013	Complete/ Multichoice	Paragraphs	Crafted story	Crowd-worker	2640	Child
CNN/ Daily Mail	2015	Cloze/ Extraction	Document	News article	Auto-mated	1.4M	Adult
SQuAD	2016	Complete/ Extraction	Paragraph	Wikipedia	Crowd-worker	100K	Adult
Who-did-What	2016	Cloze/ Extraction	Document	News article	Auto-mated	200K	Adult
MS MARCO	2016	Complete/ Extraction	Paragraphs	Web page	Search engine	100K	Adult
NewsQA	2016	Complete/ Extraction	Document	News article	Crowd-worker	120K	Adult
TriviaQA	2017	Complete/ Extraction	Paragraph	Wikipedia/ Web page	Quiz website	650K	Adult
RACE	2017	Complete/ Multichoice	Document	Exam	Expert	100K	Student

Datasets

Dataset	Year	Question/ Answer	Given text	Genre or Source	Question sourcing	Question size	Reader
QA4MRE	2013	Complete/ Multichoice	Document	Technical documents	Expert	284	Student
MCTest	2013	Complete/ Multichoice	Paragraphs	Crafted story	Crowd-worker	2640	Child
CNN/ Daily Mail	2015	Cloze/ Extraction	Document	News article	Auto-mated	1.4M	Adult
SQuAD	2016	Complete/ Extraction	Paragraph	Wikipedia	Crowd-worker	100K	Adult
Who-did-What	2016	Cloze/ Extraction	Document	News article	Auto-mated	200K	Adult
MS MARCO	2016	Complete/ Extraction	Paragraphs	Web page	Search engine	100K	Adult
NewsQA	2016	Complete/ Extraction	Document	News article	Crowd-worker	120K	Adult
TriviaQA	2017	Complete/ Extraction	Paragraph	Wikipedia/ Web page	Quiz website	650K	Adult
RACE	2017	Complete/ Multichoice	Document	Exam	Expert	100K	Student

Datasets

Dataset	Year	Question/ Answer	Given text	Genre or Source	Question sourcing	Question size	Reader
QA4MRE	2013	Complete/ Multichoice	Document	Technical documents	Expert	284	Student
MCTest	2013	Complete/ Multichoice	Paragraphs	Crafted story	Crowd-worker	2640	Child
CNN/ Daily Mail	2015	Cloze/ Extraction	Document	News article	Auto-mated	1.4M	Adult
SQuAD	2016	Complete/ Extraction	Paragraph	Wikipedia	Crowd-worker	100K	Adult
Who-did-What	2016	Cloze/ Extraction	Document	News article	Auto-mated	200K	Adult
MS MARCO	2016	Complete/ Extraction	Paragraphs	Web page	Search engine	100K	Adult
NewsQA	2016	Complete/ Extraction	Document	News article	Crowd-worker	120K	Adult
TriviaQA	2017	Complete/ Extraction	Paragraph	Wikipedia/ Web page	Quiz website	650K	Adult
RACE	2017	Complete/ Multichoice	Document	Exam	Expert	100K	Student

Datasets

Dataset	Year	Question/ Answer	Given text	Genre or Source	Question sourcing	Question size	Reader
QA4MRE	2013	Complete/ Multichoice	Document	Technical documents	Expert	284	Student
MCTest	2013	Complete/ Multichoice	Paragraphs	Crafted story	Crowd-worker	2640	Child
CNN/ Daily Mail	2015	Cloze/ Extraction	Document	News article	Auto-mated	1.4M	Adult
SQuAD	2016	Complete/ Extraction	Paragraph	Wikipedia	Crowd-worker	100K	Adult
Who-did-What	2016	Cloze/ Extraction	Document	News article	Auto-mated	200K	Adult
MS MARCO	2016	Complete/ Extraction	Paragraphs	Web page	Search engine	100K	Adult
NewsQA	2016	Complete/ Extraction	Document	News article	Crowd-worker	120K	Adult
TriviaQA	2017	Complete/ Extraction	Paragraph	Wikipedia/ Web page	Quiz website	650K	Adult
RACE	2017	Complete/ Multichoice	Document	Exam	Expert	100K	Student

Datasets

Dataset	Year	Question/ Answer	Given text	Genre or Source	Question sourcing	Question size	Reader
QA4MRE	2013	Complete/ Multichoice	Document	Technical documents	Expert	284	Student
MCTest	2013	Complete/ Multichoice	Paragraphs	Crafted story	Crowd-worker	2640	Child
CNN/ Daily Mail	2015	Cloze/ Extraction	Document	News article	Auto-mated	1.4M	Adult
SQuAD	2016	Complete/ Extraction	Paragraph	Wikipedia	Crowd-worker	100K	Adult
Who-did-What	2016	Cloze/ Extraction	Document	News article	Auto-mated	200K	Adult
MS MARCO	2016	Complete/ Extraction	Paragraphs	Web page	Search engine	100K	Adult
NewsQA	2016	Complete/ Extraction	Document	News article	Crowd-worker	120K	Adult
TriviaQA	2017	Complete/ Extraction	Paragraph	Wikipedia/ Web page	Quiz website	650K	Adult
RACE	2017	Complete/ Multichoice	Document	Exam	Expert	100K	Student

Datasets

Dataset	Year	Question/ Answer	Given text	Genre or Source	Question sourcing	Question size	Reader
QA4MRE	2013	Complete/ Multichoice	Document	Technical documents	Expert	284	Student
MCTest	2013	Complete/ Multichoice	Paragraphs	Crafted story	Crowd-worker	2640	Child
CNN/ Daily Mail	2015	Cloze/ Extraction	Document	News article	Auto-mated	1.4M	Adult
SQuAD	2016	Complete/ Extraction	Paragraph	Wikipedia	Crowd-worker	100K	Adult
Who-did-What	2016	Cloze/ Extraction	Document	News article	Auto-mated	200K	Adult
MS MARCO	2016	Complete/ Extraction	Paragraphs	Web page	Search engine	100K	Adult
NewsQA	2016	Complete/ Extraction	Document	News article	Crowd-worker	120K	Adult
TriviaQA	2017	Complete/ Extraction	Paragraph	Wikipedia/ Web page	Quiz website	650K	Adult
RACE	2017	Complete/ Multichoice	Document	Exam	Expert	100K	Student

RC and Question Answering

Question answering

- Q + A without the explicit context

  e.g., Q: *'What is solid CO2 commonly called?'* A: *dry ice*
- Include "searching" from the resources

RC and Question Answering

Question answering

- Q + A without the explicit context

  e.g., Q: *'What is solid CO2 commonly called?'* A: *dry ice*
- Include "searching" from the resources

Reading comprehension as question answering

- Q + A with the explicit context
- Inquire not only general knowledge but the context-dependent information about temporal situations/stories

RC and Textual Entailment

Textual entailment

- Recognizing and testing:

  premise $\rightarrow$ hypothesis

  e.g., Premise: *'alcohol reduces blood pressure'*

  $\rightarrow$ Hypothesis: *'alcohol affects blood pressure'*

RC and Textual Entailment

Textual entailment

- Recognizing and testing:
  premise $\rightarrow$ hypothesis
  e.g., Premise: *'alcohol reduces blood pressure'*
  $\rightarrow$ Hypothesis: *'alcohol affects blood pressure'*

Reading comprehension as textual entailment

- Recognizing and testing:
  multiple premises (from context) $\rightarrow$ hyposhesis (from Q+A)
- Cf. FraCaS (Cooper, 1996) (issues: simple & small)
- Etzioni^{+} (2006): "Machine reading [...] combines multiple textual entailment steps to form a coherent set of beliefs based on the text." (see also Manning (2006))

Agenda

1. Overview of the machine reading comprehension (RC) task
   “What is the RC task?”
2. Evaluation methodology for RC datasets/systems
   “How can we evaluate our systems/datasets?”
3. Discussion for constructing RC datasets:
   “How can we create *difficult but not too difficult* questions?”

Motivation: Accuracy is not Enough

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

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

⊗ Chen^{+} (2016) shows: CNN/Daily Mail datasets contain unanswerable or ambiguous questions

Our Research Question

- How can we evaluate and analyze our RC systems?
  - → Propose evaluation metrics for RC
  - → Focus on prerequisite skills and readability

Motivation: Two Types of Difficulties

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.29
Context: The princess climbed out the window of the high tower and climbed down the south wall when her mother was sleeping. She wandered out a good ways. Finally she went into the forest where there are no electric poles but where there are some caves.
Question: Where did the princess wander to after escaping?
Answer: A) Mountain *B) Forest C) Cave D) Castle

Motivation: Two Types of Difficulties

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.29
Context: The princess climbed out the window of the high tower and climbed down the south wall when her mother was sleeping. She wandered out a good ways. Finally she went into the forest where there are no electric poles but where there are some caves.
Question: Where did the princess wander to after escaping?
Answer: A) Mountain *B) Forest C) Cave D) Castle

Motivation: Two Types of Difficulties

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 r*
Answer: **Baptized M**

→ **Answerble simply by noticing one sentence**

ID: MCTest (2013), mc160.dev.29
Context: The princess climbed out the window of the high tower and climbed down the south wall when her mother was sleeping. She wandered out a good ways. Finally she went into the forest where there are no electric poles but where there are some caves.
Question: Where did the princess wander to after escaping?
Answer: A) Mountain *B) Forest C) Cave D) Castle

Motivation: Two Types of Difficulties

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 r*
Answer: **Baptized M**

→ **Answerble simply by noticing one sentence**

ID: MCTest (2013), mc160.dev.29
Context: The *princess climbed out* the window of the high tower and *climbed down* the south wall when her mother was sleeping. *She wandered out* a good ways. *Finally she went into the forest* where there are no electric poles but where there are some caves.
Question: Where did the princess wander to *after escaping*?
Answer: A) Mountain *B) **Forest** C) Cave D) Castle

Motivation: Two Types of Difficulties

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 r*
Answer: **Baptized M**

→ **Answerble simply by noticing one sentence**

ID: MCTest (2013), mc160.dev.29
Context: The *princess climbed out* the window of the high tower and *climbed down* the south wall when her mother was sleeping. *She wandered out* a good ways. *Finally she went into the forest* where there are no electric poles but where there are some caves.
Question: Where d
Answer: A) Mount

→ **Requre reading multile sentence with skills**

Motivation: Two Types of Difficulties

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 subseque*
Answer: **Baptized Members**

Difficult-to-read & Easy-to-answer

ID: MCTest (2013), mc160.dev.29
Context: The *princess climbed out* the window of the high tower and *climbed down* the south wall when her mother was sleeping. *She wandered out* a good ways. *Finally she went into the forest* where there are no electric poles but where there are some caves.
Question: Where did the princes
Answer: A) Mountain *B) **Fores**

Easy-to-read & Difficult-to-answer

Our study: Evaluation Metrics for RC

1. Define prerequisite skills and readability metrics
   - 13 prerequisite skills and 10 readability measures

Our study: Evaluation Metrics for RC

1. Define prerequisite skills and readability metrics
   - 13 prerequisite skills and 10 readability measures
2. Annotate RC datasets with skills
   - 6 datasets: QA4MRE, MCTest, SQuAD, Who-did-What, MS MARCO, NewsQA

Our study: Evaluation Metrics for RC

1. Define prerequisite skills and readability metrics
   - 13 prerequisite skills and 10 readability measures
2. Annotate RC datasets with skills
   - 6 datasets: QA4MRE, MCTest, SQuAD, Who-did-What, MS MARCO, NewsQA
3. Calculate readability of the datasets
   - Readability of “context sentences necessary for answering” (selected in the annotation) ($\neq$ whole context)

Our study: Evaluation Metrics for RC

1. Define prerequisite skills and readability metrics
   - 13 prerequisite skills and 10 readability measures
2. Annotate RC datasets with skills
   - 6 datasets: QA4MRE, MCTest, SQuAD, Who-did-What, MS MARCO, NewsQA
3. Calculate readability of the datasets
   - Readability of "context sentences necessary for answering" (selected in the annotation) ($\neq$ whole context)
4. Analyze the datasets on two types of difficulties
   - See the relation between skills and readability

System Analysis by Accuracy

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

System Analysis by the Skills and Readability

	Dataset A								System X
	Prerequisite Skills				*Readability Metrics*				
Question	*Skill 1*	*Skill 2*	···	*Skill 13*	*RM 1*	*RM 2*	···	*RM 10*	
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%

This study shows statistics of datasets & an observation on the relation between skills (difficulty in answering) and readability (difficulty in reading)

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

- New knowledge reasoning skills in this study

  “Commonsense reasoning” is updated to new 4 skills for more detailed analysis

Prerequisite Skills

1. Object tracking	8. Ellipsis
2. Mathematical reasoning	9. Bridging
3. Coreference resolution	10. Elaboration
4. Logical reasoning	11. Meta-knowledge
5. Analogy	12. Schematics clause relation
6. Causal relation	13. Punctuation
7. Spatiotemporal relation	

- Previous study (Sugawara^{+}, AAAI 2017): skills are based on existing NLU tasks in NLP

→ Analyzed MCTest dataset and three systems, and showed that *"the more skills are required, the more difficult to answer (lower accuracy)."*

- We regard this as *the difficulty of answering*

Annotated RC Datasets: 100 Qs for each

RC dataset	Genre	Query sourcing	Task formulation
QA4MRE (2013)	Technical documents	Handcrafted by experts	Multiple choice
MCTest (2013)	Narratives by crowd workers	Crowd sourced	Multiple choice
SQuAD (2016)	Wikipedia articles	Crowd sourced	Text span selection
Who-did-What (2016)	News articles (Gigaward v5)	Automated from other articles	Cloze
MS MARCO (2016)	Segmented web pages	Search engine queries	Description
NewsQA (2016)	News articles	Crowd sourced	Text span selection

Annotation with the Prerequisite Skills

- 13 skillls: mult-label annotation
- 6 datasets: QA4MRE, MCTest, SQuAD, Who-did-What, MS MARCO, NewsQA
- 100 questions for each dataset
- 4 annotators: graduate NLP students
- For 62 randomly sampled questions, 90.1% agreement

- Annotation: choose skill labels and “necessary context sentences” for answering
- Sentences are used to calculate readability measures

Result: Frequencies (%) of Prerequisite Skills

Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1. Tracking	11.0	6.0	3.0	8.0	6.0	2.0
2. Math.	4.0	4.0	0.0	3.0	0.0	1.0
3. Coref. resol.	32.0	49.0	13.0	19.0	15.0	24.0
4. Logical rsng.	15.0	2.0	0.0	8.0	1.0	2.0
5. Analogy	7.0	0.0	0.0	7.0	0.0	3.0
6. Causal rel.	1.0	6.0	0.0	2.0	0.0	4.0
7. Sptemp rel.	26.0	9.0	2.0	2.0	0.0	3.0
8. Ellipsis	13.0	4.0	3.0	16.0	2.0	15.0
9. Bridging	69.0	26.0	42.0	59.0	36.0	50.0
10. Elaboration	60.0	8.0	13.0	57.0	18.0	36.0
11. Meta	1.0	1.0	0.0	0.0	0.0	0.0
12. Clause rel.	52.0	40.0	28.0	42.0	27.0	34.0
13. Punctuation	34.0	1.0	24.0	20.0	14.0	25.0
Nonsense	10.0	1.0	3.0	27.0	14.0	1.0

Result: Frequencies (%) of Prerequisite Skills

Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1. Tracking	11.0	6.0	3.0	8.0	6.0	2.0
2. Math.	4.0	4.0	0.0	3.0	0.0	1.0
3. Coref. resol.	32.0	49.0	13.0	19.0	15.0	24.0
4. Logical rsng.	15.0	2.0	0.0	8.0	1.0	2.0
5. Analogy	7.0	0.0	0.0	7.0	0.0	3.0
6. Causal rel.	1.0	6.0	0.0	2.0	0.0	4.0
7. Sptemp rel.	26.0	9.0	2.0	2.0	0.0	3.0
8. Ellipsis	13.0	4.0	3.0	16.0	2.0	15.0
9. Bridging	69.0	26.0	42.0	59.0	36.0	50.0
10. Elaboration	60.0	8.0	13.0	57.0	18.0	36.0
11. Meta	1.0	1.0	0.0	0.0	0.0	0.0
12. Clause rel.	52.0	40.0	28.0	42.0	27.0	34.0
13. Punctuation	34.0	1.0	24.0	20.0	14.0	25.0
Nonsense	10.0	1.0	3.0	27.0	14.0	1.0

Result: Frequencies (%) of Prerequisite Skills

Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1. Tracking	11.0	6.0	3.0	8.0	6.0	2.0
2. Math.	4.0	4.0	0.0	3.0	0.0	1.0
3. Coref. resol.	32.0	49.0	13.0	19.0	15.0	24.0
4. Logical rsng.	15.0	2.0	0.0	8.0	1.0	2.0
5. Analogy	7.0	0.0	0.0	7.0	0.0	3.0
6. Causal rel.	1.0	6.0	0.0	2.0	0.0	4.0
7. Sptemp rel.	26.0	9.0	2.0	2.0	0.0	3.0
8. Ellipsis	13.0	4.0	3.0	16.0	2.0	15.0
9. Bridging	69.0	26.0	42.0	59.0	36.0	50.0
10. Elaboration	60.0	8.0	13.0	57.0	18.0	36.0
11. Meta	1.0	1.0	0.0	0.0	0.0	0.0
12. Clause rel.	52.0	40.0	28.0	42.0	27.0	34.0
13. Punctuation	34.0	1.0	24.0	20.0	14.0	25.0
Nonsense	10.0	1.0	3.0	27.0	14.0	1.0

Result: Frequencies (%) of Prerequisite Skills

Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
1. Tracking	11.0	6.0	3.0	8.0	6.0	2.0
2. Math.	4.0	4.0	0.0	3.0	0.0	1.0
3. Coref. resol.	32.0	49.0	13.0	19.0	15.0	24.0
4. Logical rsng.	15.0	2.0	0.0	8.0	1.0	2.0
5. Analogy	7.0	0.0	0.0	7.0	0.0	3.0
6. Causal rel.	1.0	6.0	0.0	2.0	0.0	4.0
7. Sptemp rel.	26.0	9.0	2.0	2.0	0.0	3.0
8. Ellipsis	13.0	4.0	3.0	16.0	2.0	15.0
9. Bridging	69.0	26.0	42.0	59.0	36.0	50.0
10. Elaboration	60.0	8.0	13.0	57.0	18.0	36.0
11. Meta	1.0	1.0	0.0	0.0	0.0	0.0
12. Clause rel.	52.0	40.0	28.0	42.0	27.0	34.0
13. Punctuation	34.0	1.0	24.0	20.0	14.0	25.0
Nonsense	10.0	1.0	3.0	27.0	14.0	1.0

Result: Numbers of Required Skills

#Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
0	2.0	18.0	27.0	2.0	15.0	13.0
1	13.0	36.0	33.0	5.0	35.0	26.0
2	13.0	24.0	24.0	14.0	29.0	23.0
3	20.0	15.0	6.0	22.0	6.0	25.0
4	14.0	4.0	6.0	16.0	2.0	9.0
5	13.0	1.0	1.0	6.0	0.0	2.0
6	10.0	1.0	0.0	6.0	0.0	1.0
7	1.0	0.0	0.0	2.0	0.0	0.0
8	1.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0
10	3.0	0.0	0.0	0.0	0.0	0.0
Ave.	3.25	1.56	1.28	2.43	1.19	1.99

Result: Numbers of Required Skills

#Skills	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
0	2.0	18.0	27.0	2.0	15.0	13.0
1	13.0	36.0	33.0	5.0	35.0	26.0
2	13.0	24.0	24.0	14.0	29.0	23.0
3	20.0	15.0	6.0	22.0	6.0	25.0
4	14.0	4.0	6.0	16.0	2.0	9.0
5	13.0	1.0	1.0	6.0	0.0	2.0
6	10.0	1.0	0.0	6.0	0.0	1.0
7	1.0	0.0	0.0	2.0	0.0	0.0
8	1.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0
10	3.0	0.0	0.0	0.0	0.0	0.0
Ave.	3.25	1.56	1.28	2.43	1.19	1.99

Calculation of Readability

- 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*)

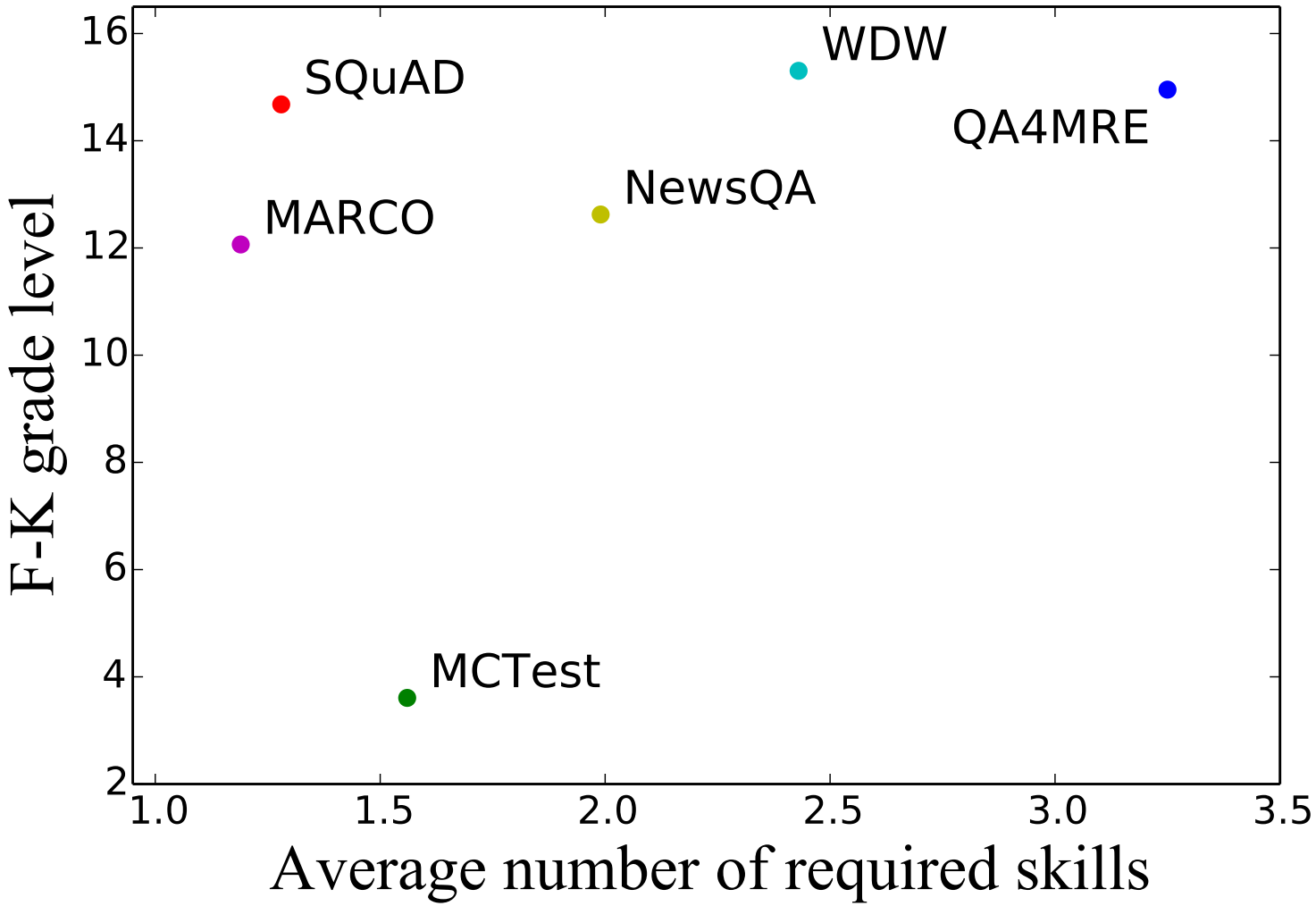
Figure: 10 readability measure from Vajjala and Meurers (2012).

Result: Readability Metrics

Mesures	QA4MRE	MCTest	SQuAD	WDW	MARCO	NewsQA
NumChar	5.026	*3.892*	***5.378***	4.988	5.016	5.017
NumSyll	1.663	*1.250*	***1.791***	1.657	1.698	1.635
MLS	28.488	*11.858*	23.479	***29.146***	19.634	22.933
AWL	0.067	*0.003*	***0.071***	0.033	0.047	0.038
ModVar	0.174	*0.114*	***0.188***	0.150	0.186	0.138
CoOrd	***0.922***	*0.309*	0.722	0.467	0.651	0.507
Coleman	12.553	*4.333*	***14.095***	12.398	11.836	12.138
DC/C	***0.343***	0.223	0.243	0.254	*0.220*	0.264
CN/C	1.948	*0.614*	1.887	***2.310***	1.935	1.702
AdvVar	***0.038***	0.035	0.032	*0.019*	0.022	*0.019*
F–K	14.953	*3.607*	14.678	***15.304***	12.065	12.624
Words	***1545.7***	174.1	130.4	253.7	*70.7*	638.4

**F–K* = Flesch-Kincaid grade level
= education level required to understand the text.

Relation btwn Skills and Readability



Observation

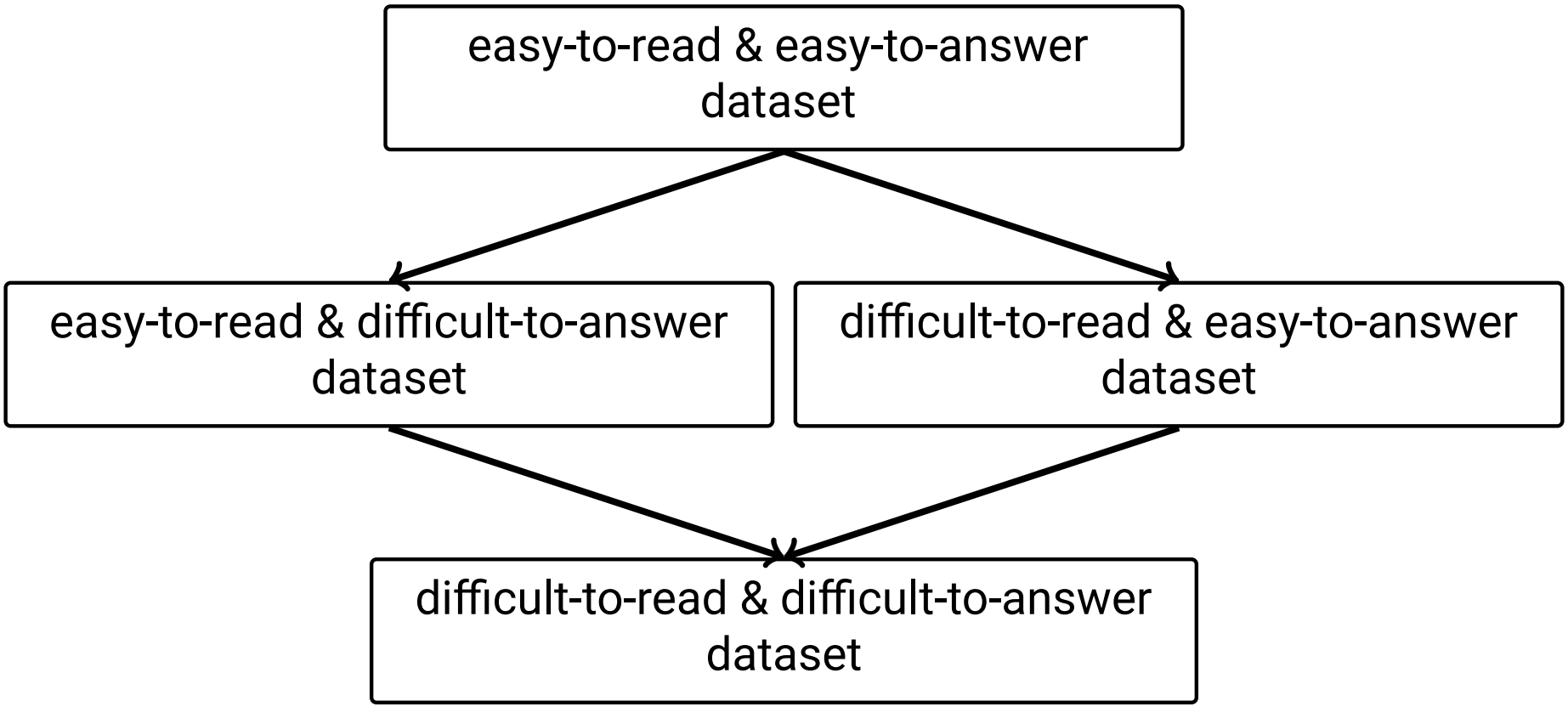
- ⊗ There is only a weak correlation between readability metrics and numbers of required skills

→ Difficult to read $\not\Rightarrow$ difficult to answer (and vice versa)

→ It is possible to create a dataset that consists of an easy-to-read context and difficult-to-answer questions.

How to Use This Observation?

For development of RC systems, select datasets in the following steps:



Short Discussion

Answerability of questions

- We cannot evaluate the difference among “truely difficult,” “non sense,” and “no answer” when we encounter systems’ incorrect answers (and even in human performance!)
- It is not easy to maintain the quality of questions especially in crowd-based sourcing.

Short Discussion

Answerability of questions

- We cannot evaluate the difference among “truely difficult,” “non sense,” and “no answer” when we encounter systems’ incorrect answers (and even in human performance!)
- It is not easy to maintain the quality of questions especially in crowd-based sourcing.

Corpus genre

- Are Wikipedia or news articles enough?

  Example: narratives are close to our everyday experience (characters’ emotions, intentions, and actions)

Summary of the Second Part

Proposed evaluation methodology for RC

1. Defined two classes of metrics:
   prerequisite skills and readability
2. Annotated RC datasets with the skills
3. Calculated readability of datasets
4. Analyzed datasets

$\rightarrow$ Results can be used for evaluation of systems

Observation

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

Agenda

1. Overview of the machine reading comprehension (RC) task
   “What is the RC task?”
2. Evaluation methodology for RC datasets/systems
   “How can we evaluate our systems/datasets?”
3. Discussion for constructing RC datasets:
   “How can we create *difficult but not too difficult* questions?”

Agenda

1. Overview of the machine reading comprehension (RC) task
   "What is the RC task?"
2. Evaluation methodology for RC datasets/systems
   "How can we evaluate our systems/datasets?"
3. Discussion for constructing RC datasets:
   "How can we create *difficult but not too difficult* questions?"

$\rightarrow$ What factors affect the difficulty of RC questions?

How do we solve RC questions?

1. Read a given text and understand the meaning of a question about it
   - e.g. which type of entitiy is required.

How do we solve RC questions?

1. Read a given text and understand the meaning of a question about it
   - e.g. which type of entitiy is required.
2. If needed (e.g., in answer extraction questions), find answer candidates from the context

How do we solve RC questions?

1. Read a given text and understand the meaning of a question about it
   - e.g. which type of entitiy is required.
2. If needed (e.g., in answer extraction questions), find answer candidates from the context
3. Make hypotheses using the question and each answer candidate

How do we solve RC questions?

1. Read a given text and understand the meaning of a question about it
   - e.g. which type of entitiy is required.
2. If needed (e.g., in answer extraction questions), find answer candidates from the context
3. Make hypotheses using the question and each answer candidate
4. Test each hypothesis whether it can be entailed from a given text or its sub-sentences

How do we solve RC questions?

1. Read a given text and understand the meaning of a question about it
   - e.g. which type of entitiy is required.
2. If needed (e.g., in answer extraction questions), find answer candidates from the context
3. Make hypotheses using the question and each answer candidate
4. Test each hypothesis whether it can be entailed from a given text or its sub-sentences
5. Choose a hypothesis that is most likely to be entailed, and return its corresponding candidate

Parameters of the Difficulty

A. Understanding a question itself
   - ⌖ Involving some skills such as coreference resolution
   - → Easier if a question does not require background knowledge (difficult example: “what is the main theme of this passage?”)

Parameters of the Difficulty

A. Understanding a question itself
   - ⌖ Involving some skills such as coreference resolution
   - → Easier if a question does not require background knowledge (difficult example: “what is the main theme of this passage?”)

B. Identifying answer candidates
   - ⌖ Recognizing entity types (wh-) and clause (why/how)
   - → Easier if there are few answer candidates (e.g., there is only one temporal expression in the context for a *when* question) Multiple choice seems better (cf. Levesque (2013))

Parameters of the Difficulty

C. Choosing evidential sentence(s) from a given text
- ⊕ Which sentence is informative for examining a hypothesis

→ Easier if the number of sentences is small (like TE)

Parameters of the Difficulty

C. Choosing evidential sentence(s) from a given text
   - ⊕ Which sentence is informative for examining a hypothesis
   - → Easier if the number of sentences is small (like TE)

D. Testing entailment between evidnece and hypothesis
   - ⊕ How many prerequisite skills are required
     (Partly revealed by our evaluation methodology)
   - → Easier if no skill is required (e.g., the hypothesis is almost similar to the evidence)

Summary

1. Overview of the machine reading comprehension task
   “What is the reading comprehension (RC) task?”
2. Evaluation methodology for RC datasets/systems
   “How can we evaluate our systems/datasets?”
3. Discussion for constructing RC datasets:
   “How can we create *difficult but not too difficult* questions?”

Reference I

- Burges (2013) Towards the Machine Comprehension of Text: An Essay, technical report.
- Chen et al. (2016) A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task, in ACL.
- Cooper et al. (1996) Using the Framework, by the FraCaS Consortium.
- Etzioni et al. (2006) Machine Reading, in AAAI.
- Goldberg (2017) Near human performance in question answering? (`http://u.cs.biu.ac.il/~yogo/on-squad.pdf`).
- Hermann et al. (2015) Teaching Machines to Read and Comprehend, in NIPS.
- Jia et al. (2017) Adversarial Examples for Evaluating Reading Comprehension Systems, in EMNLP.
- Lai et al. (2017) RACE: Large-scale ReAding Comprehension Dataset From Examinations, in EMNLP.
- Levesque (2013) On our best behaviour, in IJCAI.
- Joshi et al. (2017) TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, in ACL.
- Manning (2006) Local textual inference: Its hard to circumscribe, but you know it when you see it–and NLP needs it, unpublished manuscript.

Reference II

- Nguyen et al. (2016) MS MARCO: A Human Generated MAchine Reading COmprehension Dataset.
- Onishi et al. (2016) Who did What: A Large-Scale Person-Centered Cloze Dataset, in EMNLP.
- Rajpurkar et al. (2016) SQuAD: 100,000+ Questions for Machine Comprehension of Text, in EMNLP.
- Richardson et al. (2013) MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text, in EMNLP.
- Sugawara et al. (2017a) Prerequisite Skills for Reading Comprehension: Multi-perspective Analysis of MCTest Datasets and Systems, in AAAI.
- Sugawara et al. (2017b) Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability, in ACL.
- Sutcliffe et al. (2013) Overview of QA4MRE Main Task at CLEF 2013, in CLEF.
- Trischler et al. (2016) NewsQA: A Machine Comprehension Dataset, in REP4NLP.

Appendix: Correlation btwn Readability and the Skills

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
F–K	0.348	0.000

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

Appendix: Knowledge reasoning

- Ellipsis
  - Recognizing implicit/omitted information
  - e.g. *She is a smart student* $\rightarrow$ *She is a student*

Appendix: Knowledge reasoning

- Ellipsis
  - Recognizing implicit/omitted information
  - e.g. *She is a smart student* $\rightarrow$ *She is a student*
- Bridging
  - Inferences supported by grammatical and lexical information
  - e.g. *She loves sushi.* $\rightarrow$ *She likes sushi.*

Appendix: Knowledge reasoning

- Ellipsis
  - Recognizing implicit/omitted information
  - e.g. *She is a smart student* $\rightarrow$ *She is a student*
- Bridging
  - Inferences supported by grammatical and lexical information
  - e.g. *She loves sushi.* $\rightarrow$ *She likes sushi.*
- Elaboration
  - Inference using known facts and general knowledge
  - e.g. *The writer of Hamlet was Shakespeare* $\rightarrow$ *Shakespeare wrote Hamlet*

Appendix: Knowledge reasoning

- Ellipsis
  - Recognizing implicit/omitted information
  - e.g. *She is a smart student* $\rightarrow$ *She is a student*
- Bridging
  - Inferences supported by grammatical and lexical information
  - e.g. *She loves sushi.* $\rightarrow$ *She likes sushi.*
- Elaboration
  - Inference using known facts and general knowledge
  - e.g. *The writer of Hamlet was Shakespeare* $\rightarrow$ *Shakespeare wrote Hamlet*
- Meta-knowledge
  - Inference using external knowledge including a reader, writer, and text genre
  - e.g. *Who is the main character in this story?*