Evaluating Natural Language Understanding in Machine Reading Comprehension

Saku Sugawara

Department of Computer Science, University of Tokyo

PhD defense, January 16, 2020

Evaluation of Natural Language Understanding

Goal

- Developing a system that understand human languages

= Computationally modeling language understanding

→ Studying *human language understanding*

Evaluation of Natural Language Understanding

Goal

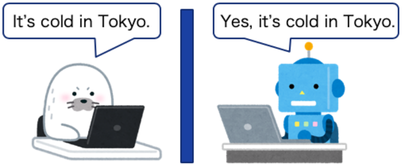
- Developing a system that understand human languages

= Computationally modeling language understanding

→ Studying *human language understanding*

Tasks

- Turing Test (1950)
- Question answering (1960s-)
- Recognizing textual entailment (2005-)
- Machine reading comprehension (2013-)



Turing Test (1950)

Premise:	A woman selling bamboo sticks talking to two men on a loading dock.
Hypothesis:	There are at least three people on the loading dock.
Entailment:	Yes

Recognizing Textual Entailment (2005)

Machine Reading Comprehension (MRC) 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

Machine Reading Comprehension (MRC) 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*)

MRC Datasets and Systems

	Datasets	Systems
2013	MCTest (2K) QA4MRE (240)	Feature-based models
2015	bAbI (10K) CNN/Daily Mail (1.4M) CBT (700K)	LSTM-based models (BiDAF: 2.5M)
2016	SQuAD (100K) WikiReading (18M) LAMBADA (10K) Who-did-What (200K) NewsQA (120K) MS MARCO (100K)	
2017	TriviaQA (650K) RACE (100K) QAngaroo (50K) NarrativeQA (50K) MCScript (30K) ...	
2018	ARC (8K) CliCR (100K) MultiRC (6K) SQuAD2.0 (100K) DuoRC (200K) HotpotQA (113K) QuAC (100K) CoQA (127K) ...	Transformer-based models (GPT-2/BERT/XLNet: 300-400M?)
2019	DROP (100K) ReCoRD (120K) MCScript2.0 (20K) ...	

MRC Datasets and Systems

	Datasets	Systems
2013	MCTest (2K) QA4MRE (240)	Feature-based models
2015	bAbI (10K) CNN/Daily Mail (1.4M) CBT (700K)	LSTM-based models (BiDAF: 2.5M)
2016	SQuAD (100K) WikiReading (18M) LAMBADA (10K) Who-did-What (200K) NewsQA (120K) MS MARCO (100K)	
2017	TriviaQA (650K) RACE (100K) QAngaroo (50K) NarrativeQA (50K) MCScript (30K) ...	
2018	ARC (8K) CliCR (100K) MultiRC (6K) SQuAD2.0 (100K) DuoRC (200K) HotpotQA (113K) QuAC (100K) CoQA (127K) ...	Transformer-based models (GPT-2/BERT/XLNet: 300-400M?)
2019	DROP (100K) ReCoRD (120K) MCScript2.0 (20K) ...	

More than 50 datasets

Huge models

Major Datasets

Dataset	Year	Domain	Sourcing	Ans Style	Focus
MCTest	2013	children stories	crowdsourced	multiple choice	first dataset
CNN/Daily Mail	2015	news articles	automated	entity cloze	first large-scale dataset
SQuAD	2016	Wikipedia articles	crowdsourced	extraction	large scale, written by humans
RACE	2017	English exams	experts	multiple choice	written by expert, various domains
HotpotQA	2018	open domain in Wikipedia	crowdsourced	extraction	multihop reasoning
DROP	2019	Wikipedia articles	crowdsourced	generation	discrete reasoning

Systems achieved human-level performance...

SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 May 21, 2019	**XLNet (single model)** *Google Brain & CMU*	**89.898**	**95.080**
2 Oct 05, 2018	**BERT (ensemble)** *Google AI Language* https://arxiv.org/abs/1810.04805	87.433	93.160
3	ATB (single model)	86.940	92.641

Issue 1: Evaluation Metrics

Systems achieved human-level performance. But the simple metric tells us...

Dataset A	System X
Q1	x
Q2	✓
Q3	x
⋮	⋮
Q10000	✓
Accuracy	75.0%

Issue 1: Evaluation Metrics

Systems achieved human-level performance. But the simple metric tells us...

Dataset A	System X
Q1	x
Q2	✓
Q3	x
⋮	⋮
Q10000	✓
Accuracy	75.0%

$\longrightarrow$

What skills the system has???

- Corefernece resolution?
- Commonsense reasoning?
- Logical reasoning?
- Understanding discourse relations?

Issue 1: Evaluation Metrics

Systems achieved human-level performance. But the simple metric tells us...

Dataset A	System X
Q1	x
Q2	✓
Q3	x
⋮	⋮
Q10000	✓
Accuracy	75.0%

$\rightarrow$

What skills the system has???

- Corefernece resolution?
- Commonsense reasoning?
- Logical reasoning?
- Understanding discourse relations?

No interpretability and explainability

Issue 2: Question Quality in Recent Datasets

Adversarial examples: [Jia and Liang, 2017]

MRC models are fooled by manually injected ***distracting sentences***

Context:	Peyton Manning is the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by ***John Elway***, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations.
Question:	What is the name of the quarterback who was 38 in Super Bowl XXXIII?
Predictions:	John Elway

Issue 2: Question Quality in Recent Datasets

Adversarial examples: [Jia and Liang, 2017]

MRC models are fooled by manually injected ***distracting sentences***

Context:	Peyton Manning is the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by ***John Elway***, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations. **Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.**
Question:	What is the name of the quarterback who was 38 in Super Bowl XXXIII?
Predictions:	John Elway $\rightarrow$ **Jeff Dean**

No validity and generalizability

Two Issues and Research Questions

1. Evaluation metrics
   - No explainability and interpretability

   → How to evaluate reading comprehension? (§3)

Two Issues and Research Questions

1. Evaluation metrics
   - No explainability and interpretability

   $\rightarrow$ How to evaluate reading comprehension? (§3)
2. Question quality
   - No validity and generalizability

   $\rightarrow$ How to ensure questions require precise NLU? (§4)

Two Issues and Research Questions

1. Evaluation metrics
   - No explainability and interpretability

   $\rightarrow$ How to evaluate reading comprehension? (§3)

2. Question quality
   - No validity and generalizability

   $\rightarrow$ How to ensure questions require precise NLU? (§4)

1 & 2 Benchmarking capability of MRC datasets

$\rightarrow$ How to specify high-quality questions with organized metrics? (§5)

Important for the explainability in NLU study and practical use

Issues and Motivation (in a Broad Sense)

Issues in Current NLP

- Reproducibility of findings [Bouthillier et al., 2019]

$=$ Findings in a dataset/task are transferable to other datasets/tasks?

$\rightarrow$ How can we accumulate findings in the study?

- Leaderboards tell us nothing about the task when both datasets and models are black-box.

Issues and Motivation (in a Broad Sense)

Issues in Current NLP

- Reproducibility of findings [Bouthillier et al., 2019]

$=$ Findings in a dataset/task are transferable to other datasets/tasks?

$\rightarrow$ How can we accumulate findings in the study?

- Leaderboards tell us nothing about the task when both datasets and models are black-box.

Underlying Motivation & Goal

- Create a theoretical foundation for ***the evaluation of NLU***.
- Contribute to make the NLU study (NLP in general?) more ***meaningful***.

Overview

§1, §2 Evaluating Natural Language Understanding in Machine Reading Comprehension

§3 Evaluation Metrics

- How to evaluate reading comprehension?

→ Requisite skills and text readability (AAAI 2017, ACL 2017)

§4 Question Quality

- How to ensure questions require precise NLU?

→ Heuristics for identifying easy & hard questions (EMNLP 2018)

§5 Benchmark Capacity

- How to specify high-quality questions with organized metrics?

→ Automated analysis methods for requisite skill (AAAI 2020)

Background: System Analysis by *Accuracy* (Issue 1)

Dataset A	System X
Q1	x
Q2	✓
Q3	x
⋮	⋮
Q10000	✓
Accuracy	75.0%

$\rightarrow$

What skills the system has???

- Corefernece resolution?
- Commonsense reasoning?
- Logical reasoning?
- Understanding discourse relations?

No interpretability and explainability

Motivation: System Analysis by *Skills* as Metrics

Dataset A	System X
Q1	x
Q2	✓
Q3	x
⋮	⋮
Q10000	✓
Accuracy	75.0%

Motivation: System Analysis by *Skills* as Metrics

Dataset A	System X
Q1	x
Q2	✓
Q3	x
⋮	⋮
Q10000	✓
Accuracy	75.0%

→

Dataset A					System X
	Prerequisite Skills				
Question	*Skill 1*	*Skill 2*	⋯	*Skill 13*	
Q1	x	-	⋯	x	x
Q2	-	✓	⋯	-	✓
Q3	x	x	⋯	-	x
⋮	⋮	⋮	⋮	⋮	⋮
Q10000	✓	✓	⋯	✓	✓
Accuracy	40.0%	90.0%	⋯	70.0%	75.0%

Motivation: System Analysis by *Skills* as Metrics

Dataset A	System X
Q1	x
Q2	✓
Q3	x
⋮	⋮
Q10000	✓
Accuracy	75.0%

→

Dataset A					System X
	Prerequisite Skills				
Question	*Skill 1*	*Skill 2*	⋯	*Skill 13*	
Q1	x	-	⋯	x	x
Q2	-	✓	⋯	-	✓
Q3	x	x	⋯	-	x
⋮	⋮	⋮	⋮	⋮	⋮
Q10000	✓	✓	⋯	✓	✓
Accuracy	40.0%	90.0%	⋯	70.0%	75.0%

Decompose the performance into *skills* → Detailed analysis

Chapter 3: Evaluation Metrics for MRC (ACL 2017)

SQuAD (2016)
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 professed their own faith*.
Question: What are members *who have been baptized as an infant or child but who have not professed their own faith?*
Answer: **Baptized Members**

MCTest (2013)
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

Chapter 3: Evaluation Metrics for MRC (ACL 2017)

Difficult-to-read & Easy-to-answer

SQuAD (2016)
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 professed their own faith*.
Question: What are members *who have been baptized as an infant or child but who have not professed their own faith?*
Answer: **Baptized Members**

Easy-to-read & Difficult-to-answer

MCTest (2013)
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

Chapter 3: Evaluation Metrics for MRC (ACL 2017)

Difficult-to-read & Easy-to-answer

SQuAD (2016)
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 professed their own faith*.
Question: What are members *who have been baptized as an infant or child but who have not professed their own faith?*
Answer: **Baptized Members**

Easy-to-read & Difficult-to-answer

MCTest (2013)
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

What we did in this chapter:

1. Defined two classes of metrics: requisite skills and readability
2. Annotated questions with the skills (multi labeling)
3. Analyzed & compared datasets

Chapter 3: Evaluation Metrics for MRC (ACL 2017)

Difficult-to-read & Easy-to-answer

SQuAD (2016)
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 professed their own faith*.
Question: What are members *who have been baptized as an infant or child but who have not professed their own faith?*
Answer: **Baptized Members**

Easy-to-read & Difficult-to-answer

MCTest (2013)
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

What we did in this chapter:

1. Defined two classes of metrics: requisite skills and readability
2. Annotated questions with the skills (multi labeling)
3. Analyzed & compared datasets

Provide fine-grained evaluation metrics

Requisite 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	

- Skills are newly defined for MRC, based on existing NLP tasks.
- Related works in NLU tasks don't cover discourse level skills.
  - Knowledge types in RTE [LoBue and Yates, 2011]
  - Reasoning types in science QA [Jansen et al., 2016]

Numbers of Required Skills (AAAI 2017)

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

Numbers of Required Skills (AAAI 2017)

	MCTest	Accuracy		
		Baseline	Smith^{+}	Yin^{+}
#Skills	Freq.	SW+D	LexMatch	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

- Previous study (Sugawara^{+} 2017a) observed that *"the more skills are required, the more difficult to answer (lower accuracy)."*

$\rightarrow$ # requisite skills in a question $=$ the difficulty of answering it

Result: Frequencies (%) of Requisite Skills

Skill \ Dataset	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

- 100 Qs * 6 datasets across several answering styles
- Asked to annotate with skills needed for answering
- Agreement > 90%

Result: Frequencies (%) of Requisite Skills

Skill \ Dataset	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

- 100 Qs * 6 datasets across several answering styles
- Asked to annotate with skills needed for answering
- Agreement > 90%
- MCTest

= narrative
$\rightarrow$ coreference?

Result: Frequencies (%) of Requisite Skills

Skill \ Dataset	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

- 100 Qs * 6 datasets across several answering styles
- Asked to annotate with skills needed for answering
- Agreement > 90%
- MCTest
  = narrative
  → coreference?
- QA4MRE
  = written by experts
  → reasoning?

Calculation of Readability

- Avg. Num. of characters per word (*NumChar*)
- Avg. Num. of syllables per word (*NumSyll*)
- Avg. sentence length in words (*MLS*)
- Proportion of words in Academic Word List (*AWL*)
- Modifier variation (*ModVar*)
- Num. of coordinate phrases per sentence (*CoOrd*)
- Coleman-Liau index (computed by #letters and #sentences) (*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

Measures	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.

Result: Readability Metrics

Measures	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

- QA4MRE, SQuAD, WDW
- = e.g., news & Wikipedia articles

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

Result: Readability Metrics

Measures	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

- QA4MRE, SQuAD, WDW
  = e.g., news & Wikipedia articles
- MCTest
  = stories for children

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

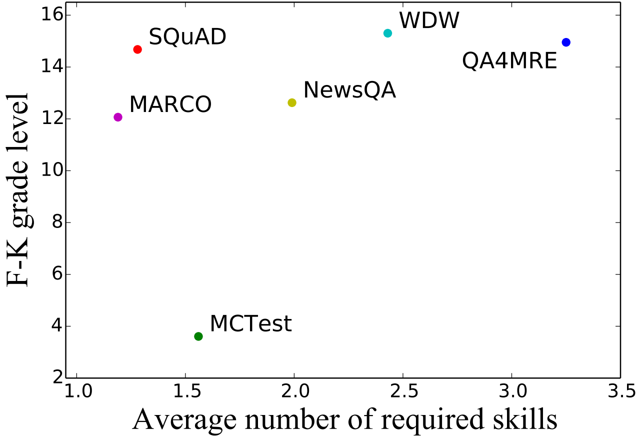
Result: Readability Metrics

Measures	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

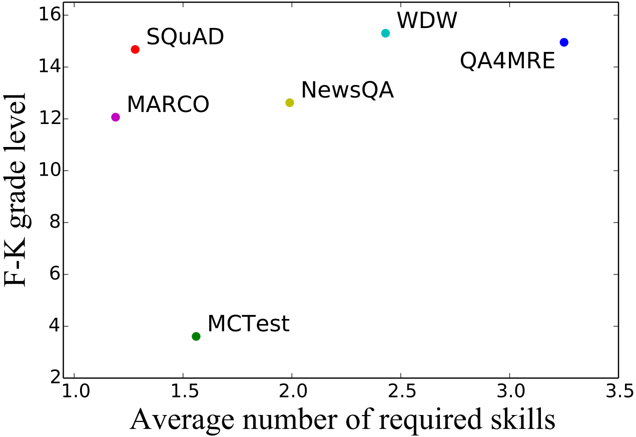
- QA4MRE, SQuAD, WDW
  = e.g., news & Wikipedia articles
- MCTest
  = stories for children

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

Relation between Skills and Readability



Relation between Skills and Readability



There is only a weak correlation $\rightarrow$ readability $\neq$ difficulty

Summary of Chapter 3

Observations

- *Difficult texts* do not necessarily make *difficult questions*

$\rightarrow$ Text readability $\neq$ question difficulty

- When controlling the question difficulty, we can focus on *easy* texts (e.g., story for children) rather than *difficult* texts (e.g., news articles)

Summary of Chapter 3

Observations

- *Difficult texts* do not necessarily make *difficult questions*

$\rightarrow$ Text readability $\neq$ question difficulty

- When controlling the question difficulty, we can focus on *easy* texts (e.g., story for children) rather than *difficult* texts (e.g., news articles)

Research Question and Contribution

Q: How to evaluate reading comprehension beyond simple accuracy?

A: Define a comprehensive set of requisite skills and readability measures

$\rightarrow$ Provide fine-grained and human-based evaluation metrics for MRC

Overview

§1, §2 Evaluating Natural Language Understanding in Machine Reading Comprehension

§3 Evaluation Metrics

- How to evaluate reading comprehension?

→ Requisite skills and text readability (AAAI 2017, ACL 2017)

§4 Question Quality

- How to ensure questions require precise NLU?

→ Heuristics for identifying easy & hard questions (EMNLP 2018)

§5 Benchmark Capacity

- How to specify high-quality questions with organized metrics?

→ Automated analysis methods for requisite skill (AAAI 2020)

Background: Question Quality in NLU Tasks (Issue 2)

Annotation artifacts

NLU tasks contain ***unintended patterns*** specific to certain answer classes

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are ***at least*** three ***people*** on a loading dock.
Neutral	A woman is selling bamboo sticks ***to help provide for her family***.
Contradiction	A woman is ***not*** taking money for any of her sticks.

- **SNLI/MultiNLI** [Gururangan et al., 2018], **StoryClozeTest** [Schwartz et al., 2017]

Background: Question Quality in NLU Tasks (Issue 2)

Annotation artifacts

NLU tasks contain ***unintended patterns*** specific to certain answer classes

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are ***at least*** three ***people*** on a loading dock.
Neutral	A woman is selling bamboo sticks ***to help provide for her family***.
Contradiction	A woman is ***not*** taking money for any of her sticks.

- **SNLI/MultiNLI** [Gururangan et al., 2018], **StoryClozeTest** [Schwartz et al., 2017]

Adversarial examples

- **SQuAD** [Jia and Liang, 2017], **SST/SNLI/SQuAD** [Wallace et al., 2019]

Background: Question Quality in NLU Tasks (Issue 2)

Annotation artifacts

NLU tasks contain ***unintended patterns*** specific to certain answer classes

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are ***at least*** three ***people*** on a loading dock.
Neutral	A woman is selling bamboo sticks ***to help provide for her family***.
Contradiction	A woman is ***not*** taking money for any of her sticks.

- **SNLI/MultiNLI** [Gururangan et al., 2018], **StoryClozeTest** [Schwartz et al., 2017]

Adversarial examples

- **SQuAD** [Jia and Liang, 2017], **SST/SNLI/SQuAD** [Wallace et al., 2019]

Questions may fail to require precise understanding?

Motivation: What Makes Questions Easier?

What kind of understanding actually happens?

Context:	In *November 2014*, Sony Pictures Entertainment was targeted by hackers who released details of confidential e-mails between Sony executives regarding several high-profile film projects. Included within these were several memos relating to the production of Spectre , claiming that [...]. Eon Productions issued a statement [...].
Question:	When did hackers get into the Sony Pictures e-mail system?
Answer:	*November 2014*

Motivation: What Makes Questions Easier?

What kind of understanding actually happens?

Context:	In *November 2014*, Sony Pictures Entertainment was targeted by hackers who released details of confidential e-mails between Sony executives regarding several high-profile film projects. Included within these were several memos relating to the production of Spectre , claiming that [...]. Eon Productions issued a statement [...].
Question:	When ~~did hackers get into the Sony Pictures e-mail system~~?
Answer:	*November 2014*

1. Recognizing entity type
   - Single candidate answer November 2014 for when

Motivation: What Makes Questions Easier?

What kind of understanding actually happens?

Context:	In *November 2014*, Sony Pictures Entertainment was targeted by hackers who released details of confidential e-mails between Sony executives regarding several high-profile film projects. Included within these were several memos relating to the production of Spectre (2015), claiming that [...]. In February 2015, Eon Productions issued a statement [...].
Question:	When ~~did hackers get into the Sony Pictures e-mail system~~?
Answer:	*November 2014*

1. Recognizing entity type
   - Single candidate answer November 2014 for when

Motivation: What Makes Questions Easier?

What kind of understanding actually happens?

Context:	In *November 2014*, **Sony Pictures** Entertainment was targeted by **hackers** who released details of confidential **e-mails** between **Sony** executives regarding several high-profile film projects. Included within these were several memos relating to the production of Spectre (2015), claiming that [...]. In February 2015, Eon Productions issued a statement [...].
Question:	When ~~did~~ **hackers** ~~get into the~~ **Sony Pictures e-mail** ~~system~~?
Answer:	*November 2014*

1. Recognizing entity type
   - Single candidate answer November 2014 for when
2. Attending words between **Context** and **Question**
   - **Sony**, **Pictures**, **hackers**, **emails**, **Sony**...

Motivation: What Makes Questions Easier?

What kind of understanding actually happens?

Context:	In *November 2014*, Sony Pictures Entertainment was targeted by hackers who released details of confidential e-mails between Sony executives regarding several high-profile film projects. Included within these were several memos relating to the production of Spectre (2015), claiming that [...]. In February 2015, Eon Productions issued a statement [...].
Question:	When ~~did~~ hackers ~~get into the~~ Sony Pictures e-mail ~~system~~?
Answer:	*November 2014*

1. Recognizing entity type
   - Single candidate answer November 2014 for when
2. Attending words between **Context** and **Question**
   - **Sony**, **Pictures**, **hackers**, **emails**, **Sony**...

→ Use these information to classify Easy & Hard questions

Chapter 4: What Makes Questions Easier? (EMNLP 2018)

What we did in this chapter:

Chapter 4: What Makes Questions Easier? (EMNLP 2018)

What we did in this chapter:

1. Propose two heuristics to identify Easy & Hard questions of 12 datasets with regard to the baseline performance
   - Entity type-based heuristic
   - Attention-based heuristic

Chapter 4: What Makes Questions Easier? (EMNLP 2018)

What we did in this chapter:

1. Propose two heuristics to identify Easy & Hard questions of 12 datasets with regard to the baseline performance
   - Entity type-based heuristic
   - Attention-based heuristic
2. Analyze these subsets by annotating with validity and requisite skills
   - Validity: solvability, multiple candidates answers, unambiguity
   - Skills: word matching, knowledge reasoning, etc. (single labeling)

Chapter 4: What Makes Questions Easier? (EMNLP 2018)

What we did in this chapter:

1. Propose two heuristics to identify Easy & Hard questions of 12 datasets with regard to the baseline performance
   - Entity type-based heuristic
   - Attention-based heuristic
2. Analyze these subsets by annotating with validity and requisite skills
   - Validity: solvability, multiple candidates answers, unambiguity
   - Skills: word matching, knowledge reasoning, etc. (single labeling)

Enable to collect questions that require a deeper understanding of texts

Two Heuristics $\rightarrow$ *Easy* and *Hard* Subsets

A. Entity type-based heuristic

Q: How many questions are solved only with the first k tokens? (for simplicity)

Two Heuristics $\rightarrow$ *Easy* and *Hard* Subsets

A. Entity type-based heuristic

Q: How many questions are solved only with the first k tokens? (for simplicity)

B. Attention-based heuristic

Q: How many questions have their answers in the most similar sentence?

- We compute unigram overlap to get intuitive results

Two Heuristics → *Easy* and *Hard* Subsets

A. Entity type-based heuristic

Q: How many questions are solved only with the first k tokens? (for simplicity)

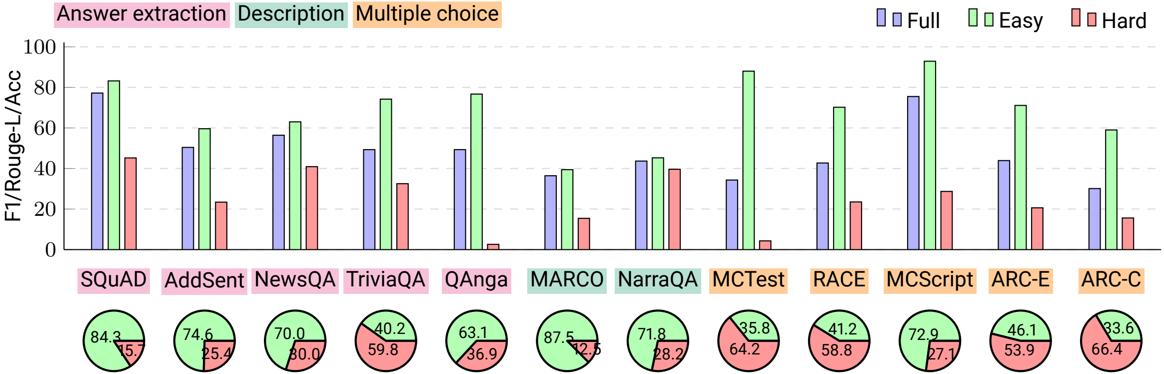
B. Attention-based heuristic

Q: How many questions have their answers in the most similar sentence?

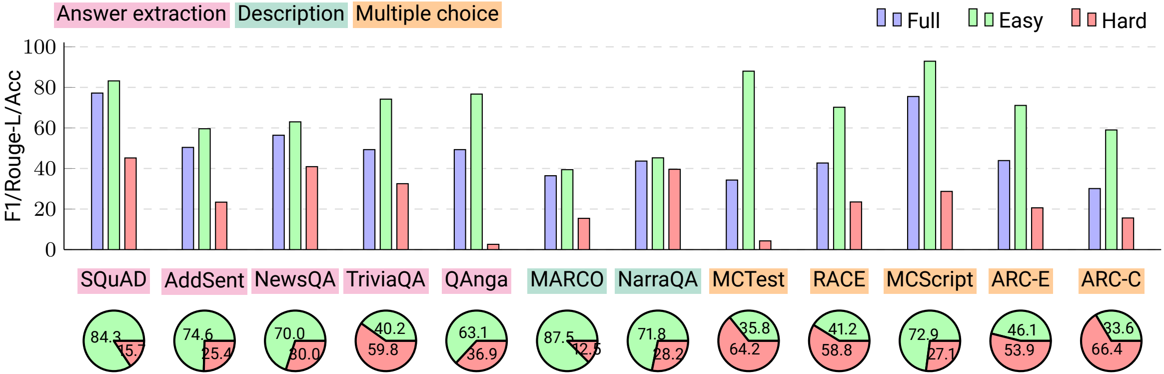
- We compute unigram overlap to get intuitive results

Heuristics		Score on the first two question tokens ($k = 2$) > 0	Score on the first two question tokens ($k = 2$) 0
Answer in most sim sentence	*Yes*	Easy	Easy
	No	Easy	Hard

Easy and *Hard* Subsets

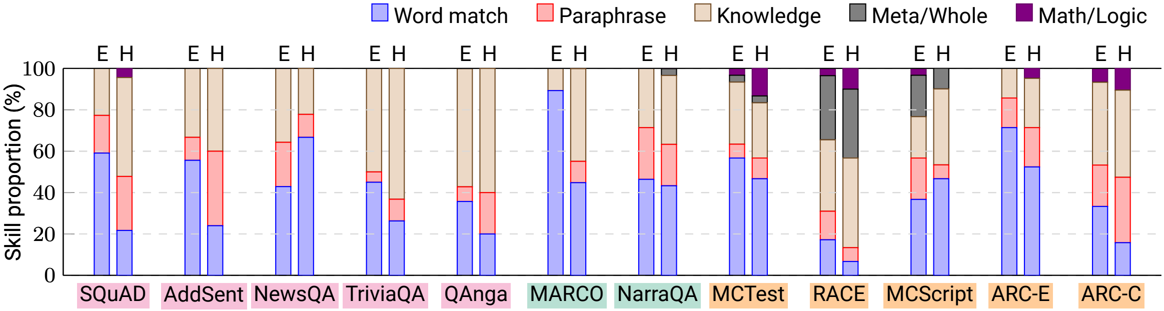


Easy and *Hard* Subsets

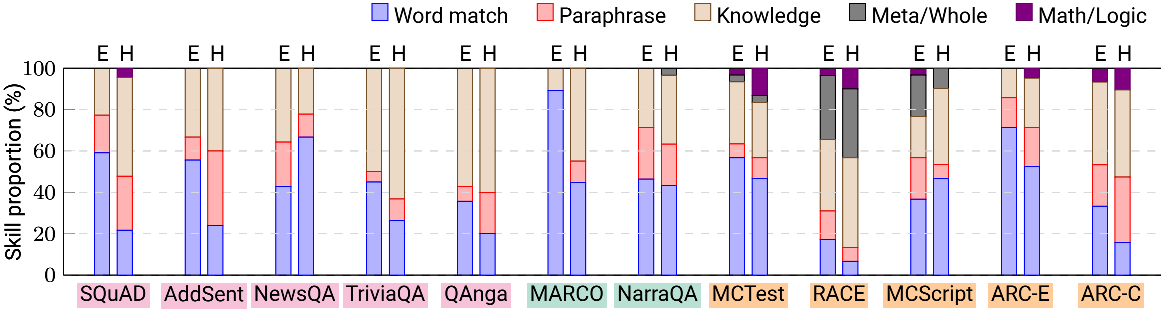


- The baseline performances: Easy >>> Hard
  - → *We overestimate the performance?*

Annotation Results: Requisite Skills

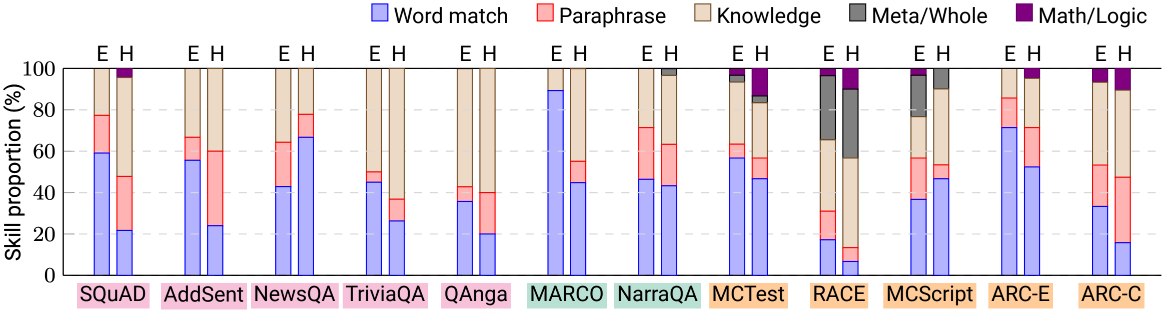


Annotation Results: Requisite Skills



- Word matching: Easy > Hard

Annotation Results: Requisite Skills



- Word matching: Easy > Hard
- Knowledge reasoning: Hard > Easy

Summary of Chapter 4

Observations

- The baseline performances: Easy $>>>$ Hard

$\rightarrow$ Overestimate the current performance?

- Knowledge reasoning & multi sentence reasoning: Hard $>$ Easy
- Multiple choice datasets are better in validity and reasoning types

Summary of Chapter 4

Observations

- The baseline performances: Easy $>>>$ Hard

$\rightarrow$ Overestimate the current performance?

- Knowledge reasoning & multi sentence reasoning: Hard $>$ Easy
- Multiple choice datasets are better in validity and reasoning types

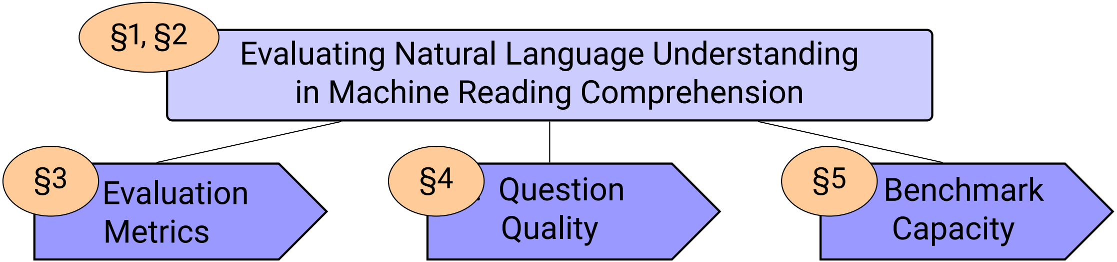
Research Question and Contribution

Q: How to ensure questions require precise NLU?

A: Develop a filtering method using two simple heuristics

$\rightarrow$ Enable to collect questions that require a deeper understanding of texts

Overview



- How to evaluate reading comprehension?

→ Requisite skills and text readability (AAAI 2017, ACL 2017)

- How to ensure questions require precise NLU?

→ Heuristics for identifying easy & hard questions (EMNLP 2018)

- How to specify high-quality questions with organized metrics?

→ Automated analysis methods for requisite skill (AAAI 2020)

Motivation: Skill-based & Automated Analysis

§3 Requisite skills for MRC

- Manual annotation of *requisite skills* in the MRC task
- Enable to detailed evaluation but necessitate much annotation cost

Motivation: Skill-based & Automated Analysis

§3 Requisite skills for MRC

- Manual annotation of *requisite skills* in the MRC task
- Enable to detailed evaluation but necessitate much annotation cost

§4 Heuristics for MRC

- *Automated analysis* of question difficulty in the MRC task
- Easy to use, but still don't provide detailed information

Motivation: Skill-based & Automated Analysis

§3 Requisite skills for MRC

- Manual annotation of *requisite skills* in the MRC task
- Enable to detailed evaluation but necessitate much annotation cost

§4 Heuristics for MRC

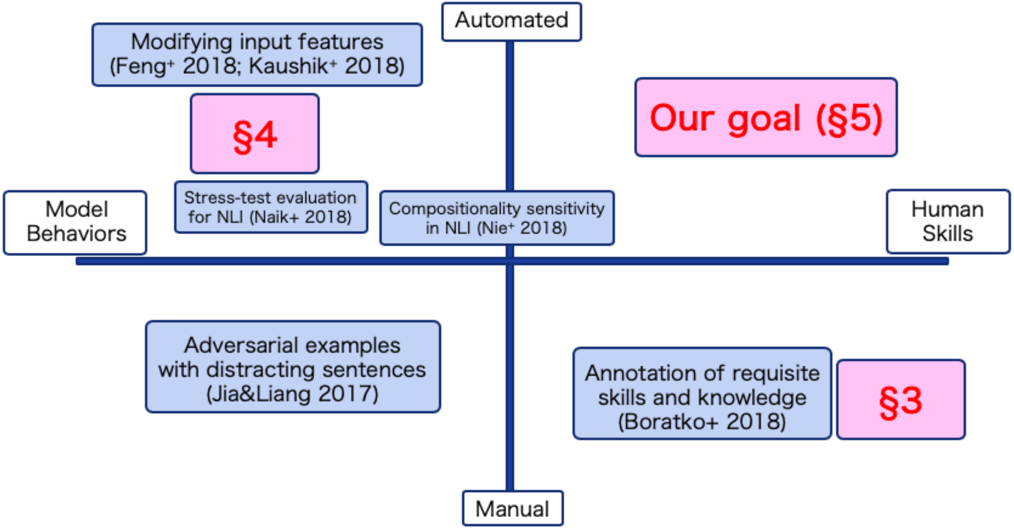
- *Automated analysis* of question difficulty in the MRC task
- Easy to use, but still don't provide detailed information

$\rightarrow$ §5 Skill-based & automated analysis for MRC

- Machines may find *bypass* solutions.

$\rightarrow$ Simple human annotation $\neq$ *true* requisite skills? $\Rightarrow$ low explainability

Analysis Methods in MRC



Intuition: Ablation of Features as Dataset Analysis

Previous Work: analyzing model behavior by input modification

- Drop tokens [Kaushik and Lipton, 2018]
- Replace tokens [Cirik et al., 2018]
- Shuffle tokens [Nie et al., 2019]

Intuition: Ablation of Features as Dataset Analysis

Previous Work: analyzing model behavior by input modification

- Drop tokens [Kaushik and Lipton, 2018]
- Replace tokens [Cirik et al., 2018]
- Shuffle tokens [Nie et al., 2019]

↓

Intuition: ablation of features

If a solved question can be still solved ***even after removing features associated with a skill***, the question do not require that skill.

Ablation of Features: Shuffling Sentence Words

Context

What colour is your name ? One person says her name is the colour red . ***Synesthesia is not a common condition*** . For these people , the everyday world can be a colourful and interesting place .

Context with *shuffled sentence words*

is colour your What name ? One is person colour her says red the name . ***Synesthesia a common is not condition*** . world the colourful , can For be people place everyday and a interesting these .

Question

What is this passage mainly about?

Options

(A) **An unusual condition.** (B) People who like colour. (C) The colour of pain. (D) Music and art.

Prediction before and after shuffling

(A) An unusual condition. $\rightarrow$ (A) An unusual condition.

Ablation of Features: Shuffling Sentence Words

Context

What colour is your name ? One person says her name is the colour red . ***Synesthesia is not a common condition*** . For these people , the everyday world can be a colourful and interesting place .

Context with *shuffled sentence words*

is colour your What name ? One is person colour her says red the name . ***Synesthesia a common is not condition*** . world the colourful , can For be people place everyday and a interesting these .

Question

What is this passage mainly about?

Options

(A) **An unusual condition.** (B) People who like colour. (C) The colour of pain. (D) Music and art.

Prediction before and after shuffling

(A) An unusual condition. $\rightarrow$ (A) An unusual condition.

Does this question require the syntax-level information?

Chapter 5: Assessing the Benchmarking Capacity of Datasets (AAAI2020)

What we did in this chapter:

Chapter 5: Assessing the Benchmarking Capacity of Datasets (AAAI2020)

What we did in this chapter:

1. Propose a semi-automated, ablation-based methodology for analyzing the benchmarking capacity of MRC datasets.
   - 12 skills and corresponding ablation methods

Chapter 5: Assessing the Benchmarking Capacity of Datasets
(AAAI2020)

What we did in this chapter:

1. Propose a semi-automated, ablation-based methodology for analyzing the benchmarking capacity of MRC datasets.
   - 12 skills and corresponding ablation methods
2. Evaluate to what degree the questions do not require the skills
   - 10 existing MRC datasets from the answer extraction and multiple choice

Chapter 5: Assessing the Benchmarking Capacity of Datasets (AAAI2020)

What we did in this chapter:

1. Propose a semi-automated, ablation-based methodology for analyzing the benchmarking capacity of MRC datasets.
   - 12 skills and corresponding ablation methods
2. Evaluate to what degree the questions do not require the skills
   - 10 existing MRC datasets from the answer extraction and multiple choice

Enable to precisely evaluate the benchmarking capacity of datasets.

12 Skills and Ablation Methods: Reading Level (1–6)

1. Recognizing the whole question except for ***interrogatives***
   - Drop all words except interrogatives (wh- words and how) in a question.
2. Recognizing ***content words***
   - Drop content words in the context.
3. Recognizing ***function words***
   - Drop function words in the context.
4. Recognizing ***vocabulary***
   - Anonymize context and questions words with their part-of-speech tag.
5. Attending the whole context other than ***similar sentences***
   - Keep the sentences that are the most similar to the question.
6. Recognizing the ***word order***
   - Randomly *shuffle* all words in the context.

12 Skills and Methods: Reasoning Level (7–12)

7. Grasping ***sentence-level compositionality***
   - Randomly shuffle the words in all the sentences except the last token.
8. ***Bridging reasoning***
   - Randomly shuffle the order of the sentences in the context.
9. Performing basic ***arithmetic operations***
   - Replace numerical expressions with random numbers.
10. ***Explicit logical reasoning***
    - Drop logical terms such as not, every, and if.
11. Resolving ***pronoun coreferences***
    - Drop personal and possessive pronouns.
12. Reasoning about ***explicit causality***
    - Drop causal terms/clauses such as because and therefore.

Other Examples

Context word shuffle (solved!)

C: Chris Ulmer, the 26-year-old teacher in Jacksonville starts his class by calling up each student individually to give them much admiration and a high-five. I couldn't help but be reminded of Syona's teacher and how she supports each kid in a very similar way.
Q: What can we learn about Chris Ulmer?
A: He praises his students one by one (multiple choice)

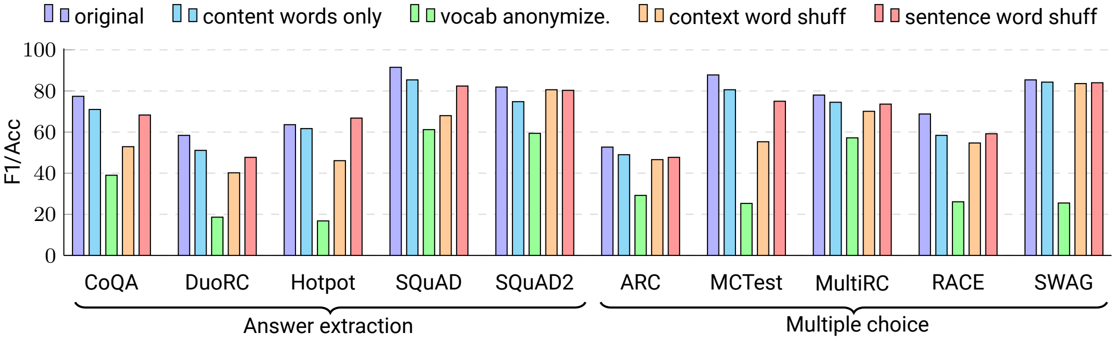
C: his help a in calling class but Syona's starts each 26-year-old similar individually Ulmer, and Chris admiration way. Jacksonville kid much I by couldn't them the a to supports of in student and teacher each be teacher reminded give how she high-five. up very
Q: What can we learn about Chris Ulmer?
A: He praises his students one by one (multiple choice)

Vocabulary anonymization (solved!)

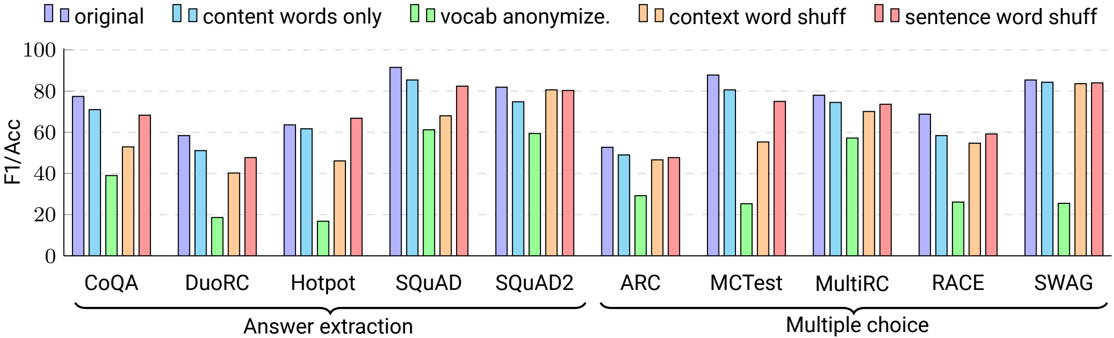
C: Immediately behind the basilica is the Grotto, a Marian place of prayer. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858.
Q: To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?
A: Saint Bernadette Soubirous

C: @adverb1 @prep5 @other0 @noun17 @verb2 @other0 @noun20 […] @other0 @noun20 @prep6 @noun25 @punct0 @noun26 @wh0 @other0 @noun7 @noun8 @adverb3 @verb4 **@prep4** **@noun27** **@noun28** **@noun29** **@prep2** **@number0** @period0
Q: **@prep4** @wh2 @verb6 @other0 @noun7 @noun8 @adverb4 @verb4 **@prep2** **@number0** @prep2 @noun25 @noun26
A: @noun27 @noun28 @noun29

Results: Trained & Evaluated on the Ablated Data

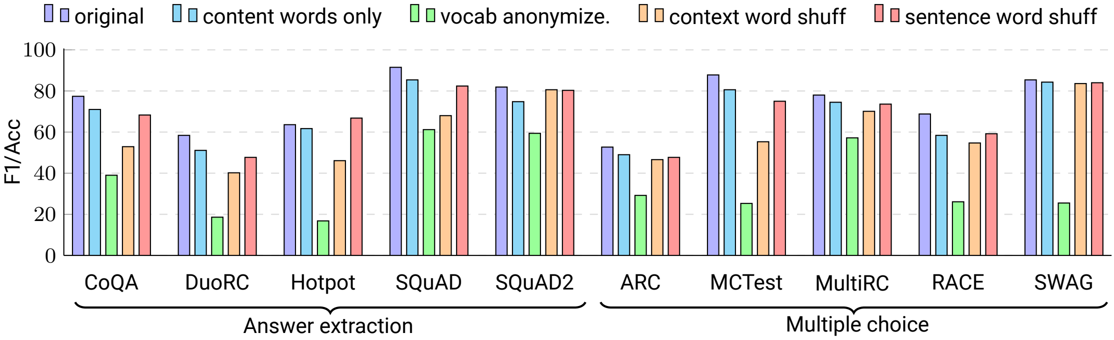


Results: Trained & Evaluated on the Ablated Data



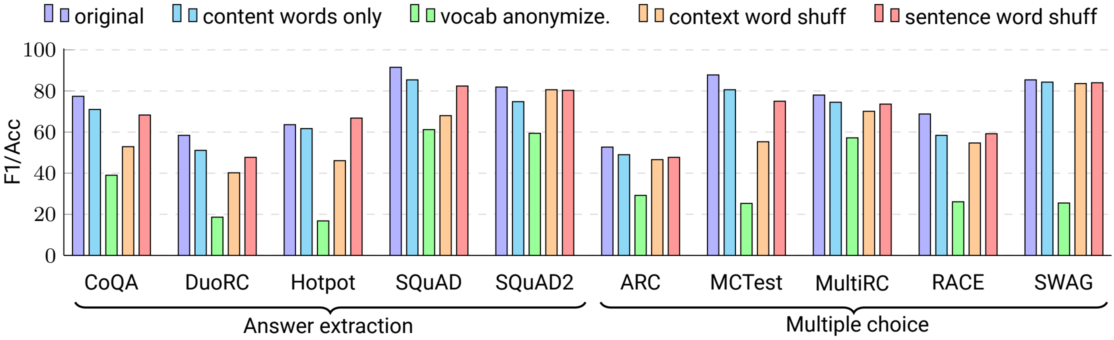
- Content words only & Sentence word shuffle : small drop from the original

Results: Trained & Evaluated on the Ablated Data



- Content words only & Sentence word shuffle: small drop from the original
- CoQA, SQuAD, SQuAD2: relatively high performance on Vocabulary anonymization

Results: Trained & Evaluated on the Ablated Data



- Content words only & Sentence word shuffle: small drop from the original
- CoQA, SQuAD, SQuAD2: relatively high performance on Vocabulary anonymization
- Multiple choice datasets: high performance on Context word shuffle

Summary of Chapter 5

Observations

- Most of the questions already answered correctly by the baseline model do not necessarily require ***lexical***, ***grammatical*** and ***complex reasoning***.

$\rightarrow$ Existing questions may fail to require complex understanding of texts

- For precise benchmarking, MRC datasets will need to take extra care in their design to ensure that questions require the intended skills.

Summary of Chapter 5

Observations

- Most of the questions already answered correctly by the baseline model do not necessarily require ***lexical***, ***grammatical*** and ***complex reasoning***.

→ Existing questions may fail to require complex understanding of texts

- For precise benchmarking, MRC datasets will need to take extra care in their design to ensure that questions require the intended skills.

Research Question and Contribution

Q: How to specify high-quality questions with organized metrics?

A: Proposed analysis methods for datasets using ***feature ablation***

→ Enable to precisely evaluate the benchmarking capacity of datasets.

Overview

§1, §2 Evaluating Natural Language Understanding in Machine Reading Comprehension

§3 Evaluation Metrics

- How to evaluate reading comprehension?

→ Requisite skills and text readability (AAAI 2017, ACL 2017)

§4 Question Quality

- How to ensure questions require precise NLU?

→ Heuristics for identifying easy & hard questions (EMNLP 2018)

§5 Benchmark Capacity

- How to specify high-quality questions with organized metrics?

→ Automated analysis methods for requisite skill (AAAI 2020)

Contributions

§3. How to evaluate reading comprehension?

- Defined a comprehensive set of requisite skills and readability measures

$\rightarrow$ Provide fine-grained evaluation metrics for NLU beyond simple accuracy.

Contributions

§3. How to evaluate reading comprehension?

- Defined a comprehensive set of requisite skills and readability measures

→ Provide fine-grained evaluation metrics for NLU beyond simple accuracy.

§4. How to ensure questions require precise NLU?

- Developed a question-filtering method using two simple heuristics

→ Enable to automatically collect hard questions that require a deeper understanding of texts beyond using superficial cues.

Contributions

§3. How to evaluate reading comprehension?

- Defined a comprehensive set of requisite skills and readability measures

$\rightarrow$ Provide fine-grained evaluation metrics for NLU beyond simple accuracy.

§4. How to ensure questions require precise NLU?

- Developed a question-filtering method using two simple heuristics

$\rightarrow$ Enable to automatically collect hard questions that require a deeper understanding of texts beyond using superficial cues.

§5. How to specify high-quality questions with organized metrics?

- Proposed analysis methods of datasets using feature ablation

$\rightarrow$ Enable to assess the capabilities of datasets for benchmarking NLU.

Summary of the Thesis

Requirements of MRC Datasets

- ***Explainability*** (cf. psychological study of human text understanding)
  - Evaluation metrics reflect the question intention in human terms.

  $\rightarrow$ Explain *what is evaluated* and *what successful models can do*.
- ***Validiation*** (cf. validity in psychometrics)
  - Questions reliably evaluate the intended skills without bypass solutions.

  $\rightarrow$ Ensure that the intended skills are evaluated.

Summary of the Thesis

Requirements of MRC Datasets

- ***Explainability*** (cf. psychological study of human text understanding)
  - Evaluation metrics reflect the question intention in human terms.
  - → Explain *what is evaluated* and *what successful models can do*.
- ***Validiation*** (cf. validity in psychometrics)
  - Questions reliably evaluate the intended skills without bypass solutions.
  - → Ensure that the intended skills are evaluated.

Why Important?

- Hypothesis verification for the scientific study of NLU
- Accountability in practical applications such as assisting human intelligent activities

Publications

1. **Saku Sugawara**, Stenetorp Pontus, Kentaro Inui, and Akiko Aizawa. 2020. *Assessing the benchmarking capacity of machine reading comprehension datasets*. In Proceedings of AAAI Conference on Artificial Intelligence (**AAAI 2020**), to appear.
2. **Saku Sugawara**, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. *What makes reading comprehension questions easier?*. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2018**), pages4028-4219.
3. **Saku Sugawara**, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017a. *Evaluation metrics for machine reading comprehension: Prerequisite skills and readability*. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (**ACL 2017**), pages 816-817.
4. **Saku Sugawara**, Hikaru Yokono, and Akiko Aizawa. 2017b. *Prerequisite skills for reading comprehension: Multi-perspective analysis of mctest datasets and systems*. In AAAI Conference on Artificial Intelligence (**AAAI 2017**), pages 3089-3096.

References I

Xavier Bouthillier, César Laurent, and Pascal Vincent. 2019. Unreproducible research is reproducible. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 725–734, Long Beach, California, USA. PMLR.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1870–1879. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 781–787, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

References II

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5927–5934, Hong Kong, China. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems (NIPS), pages 1693–1701.

References III

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Peter Jansen, Niranjan Balasubramanian, Mihai Surdeanu, and Peter Clark. 2016. What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2956–2965, Osaka, Japan. The COLING 2016 Organizing Committee.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2011–2021. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5010–5015. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 252–262. Association for Computational Linguistics.

References IV

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. Transactions of the Association for Computational Linguistics, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:453–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 796–805. Association for Computational Linguistics.

Peter LoBue and Alexander Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 329–334, Portland, Oregon, USA. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. CoRR, abs/1611.09268.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of NLI models. In Proceedings of AAAI Conference on Artificial Intelligence.

References V

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did What: A large-scale person-centered cloze dataset. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2230–2235. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7:249–266.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1683–1693. Association for Computational Linguistics.

References VI

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. Story cloze task: UW NLP system. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 52–55, Valencia, Spain. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. Transactions of the Association for Computational Linguistics, 7:217–231.

Simon Suster and Walter Daelemans. 2018. CliCR: a dataset of clinical case reports for machine reading comprehension. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1551–1563. Association for Computational Linguistics.

Richard Sutcliffe, Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2013. Overview of QA4MRE main task at CLEF 2013. Working Notes, CLEF.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 163–173. Association for Computational Linguistics.

References VII

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. Transactions of the Association for Computational Linguistics, 6:287–302.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In International Conference on Learning Representations.