What Makes Reading Comprehension Questions Easier?

Saku Sugawara♠, Kentaro Inui♣♢,
Satoshi Sekine♢, Akiko Aizawa♡♠

♠University of Tokyo, ♣Tohoku University,
♢RIKEN AIP, ♡National Institute of Informatics

November 4, EMNLP 2018

Machine Reading Comprehension & Related Datasets

Many many datasets!

2013 MCTest QA4MRE

Machine Reading Comprehension & Related Datasets

Many many datasets!

2013	MCTest	QA4MRE	
2015	bAbI	CNN/Daily Mail	CBT

Machine Reading Comprehension & Related Datasets

Many many datasets!

2013	MCTest	QA4MRE				
2015	bAbI	CNN/Daily Mail	CBT			
2016	SQuAD	WikiReading	LAMBADA	Who-did-What	NewsQA	MS MARCO

Machine Reading Comprehension & Related Datasets

Many many datasets!

2013	MCTest	QA4MRE					
2015	bAbI	CNN/Daily Mail	CBT				
2016	SQuAD	WikiReading	LAMBADA	Who-did-What	NewsQA	MS MARCO	
2017	TriviaQA	Quasar	AddSent	RACE	QAngaroo	NarrativeQA	MCScript

Machine Reading Comprehension & Related Datasets

Many many datasets!

2013	MCTest	QA4MRE					
2015	bAbI	CNN/Daily Mail	CBT				
2016	SQuAD	WikiReading	LAMBADA	Who-did-What	NewsQA	MS MARCO	
2017	TriviaQA	Quasar	AddSent	RACE	QAngaroo	NarrativeQA	MCScript
2018							

Machine Reading Comprehension & Related Datasets

Many many datasets!

2013	MCTest QA4MRE
2015	bAbI CNN/Daily Mail CBT
2016	SQuAD WikiReading LAMBADA Who-did-What NewsQA MS MARCO
2017	TriviaQA Quasar AddSent RACE QAngaroo NarrativeQA MCScript
2018	ARC CliCR MultiRC ProPara SQuAD2.0 DuoRC TextWorldsQA CLOTH emrQA RecipeQA OpenBookQA HotpotQA ShARC QuAC CoQA

Machine Reading Comprehension & Related Datasets

Many many datasets!

2013	MCTest QA4MRE
2015	bAbI CNN/Daily Mail CBT
2016	SQuAD WikiReading LAMBADA Who-did-What NewsQA MS MARCO
2017	TriviaQA Quasar AddSent RACE QAngaroo NarrativeQA MCScript
2018	ARC CliCR MultiRC ProPara SQuAD2.0 DuoRC TextWorldsQA CLOTH emrQA RecipeQA OpenBookQA HotpotQA ShARC QuAC CoQA

EMNLP2018

$>$ 30 datasets!

Reading Comprehension !!!

ID: MCTest, 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

Skills:

- Coreference resolution
- Commonsense reasoning
- Discourse understanding
- Spatiotemporal reasoning
- Logical reasoning
- Mathematical reasoning
- Causal reasoning
- Meta reasoning

...

Cf. Weston$^+$ (2015; bAbI)
Boratko$^+$ (2018; analysis on ARC)
Sugawara$^+$ (AAAI2017, ACL2017)

Reading Comprehension ???

Do MRC datasets really require *comprehension*?
$\simeq$ What kind of *comprehension* is required by MRC datasets?

Motivation: Understanding of Understanding

Annotation artifacts

Natural language understanding tasks contain *unintended patterns*

- **SNLI/MultiNLI**: Gururangan+ (2018), Poliak+ (2018), Glockner+ (2018)
  StoryClozeTest: Schwartz+ (2018)

→ Word patterns/features specific to certain answer classes

Motivation: Understanding of Understanding

Annotation artifacts

> Natural language understanding tasks contain *unintended patterns*

- **SNLI/MultiNLI**: Gururangan+ (2018), Poliak+ (2018), Glockner+ (2018)
  StoryClozeTest: Schwartz+ (2018)

→ Word patterns/features specific to certain answer classes

Adversarial examples

> MRC systems are fooled by manually injected *distracting sentences*

- SQuAD + AddSent: Jia and Liang (2017)

Motivation: Understanding of Understanding

Annotation artifacts

Natural language understanding tasks contain *unintended patterns*

- **SNLI/MultiNLI**: Gururangan+ (2018), Poliak+ (2018), Glockner+ (2018)
  StoryClozeTest: Schwartz+ (2018)

→ Word patterns/features specific to certain answer classes

Adversarial examples

MRC systems are fooled by manually injected *distracting sentences*

- SQuAD + AddSent: Jia and Liang (2017)

→ we don't know exactly what kind of understanding is required 😇

Example from SQuAD

What kind of understanding actually happens?

Question: When did hackers get into the Sony Pictures e-mail system?
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 [...].
Answer: *November 2014*

Example from SQuAD

What kind of understanding actually happens?

Question: When ~~did hackers get into the Sony Pictures e-mail system~~?
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 [...].
Answer: *November 2014*

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

Example from SQuAD

What kind of understanding actually happens?

Question: When ~~did hackers get into the Sony Pictures e-mail system~~?
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 [...].
Answer: *November 2014*

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

Example from SQuAD

What kind of understanding actually happens?

Question: Whem ~~did~~ **hackers** ~~get into the~~ **Sony Pictures e-mail** ~~system~~?
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 [...].
Answer: *November 2014*

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

What We Did in This Work

1. Propose two heuristics to identify Easy & Hard questions 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, mathematics, etc.

What We Did in This Work

1. Propose two heuristics to identify Easy & Hard questions 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, mathematics, etc.

What We Did in This Work

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

2. Analyze these subsets by annotating with varidity and requisite skills
   - Validity: solvability, multiple candidates answers, unambiguity
   - Skills: word matching, knowledge reasoning, mathematics, etc.

Two Heuristics

Two Heuristics

A. Entity type-based heuristic

to filter questions solved by recognizing the entity type of the answer or typical words around the answer

Q: How many questions are solved only with the first k tokens?
& Is there a moderate gap of performance between full and $k =$ small settings?

- e.g., *will I qualify for OSAP if I'm new in Canada?* (k = 3) (from MS MARCO)

Two Heuristics

A. Entity type-based heuristic

to filter questions solved by recognizing the entity type of the answer or typical words around the answer

Q: How many questions are solved only with the first k tokens?
& Is there a moderate gap of performance between full and $k =$ small settings?

- e.g., *will I qualify for OSAP if I'm new in Canada?* (k = 3) (from MS MARCO)

B. Attention-based heuristic

to filter questions solved by matching words between question and context

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

- We just compute unigram overlap to get intuitive results

Datasets

1. **SQuAD** (v1.1) [Rajpurkar et al., 2016]
2. **AddSent** [Jia and Liang, 2017]
3. **NewsQA** [Trischler et al., 2017]
4. **TriviaQA** (Wikipedia set) [Joshi et al., 2017]
5. **QAngaroo** (WikiHop) [Welbl et al., 2018]
6. **MS MARCO** (v2) [Nguyen et al., 2016]
7. **NarrativeQA** (summary) [Kočiský et al., 2018]
8. **MCTest** (160 + 500) [Richardson et al., 2013]
9. **RACE** (middle + high) [Lai et al., 2017]
10. **MCScript** [Ostermann et al., 2018]
11. **ARC Easy** [Clark et al., 2018]
12. **ARC Challenge** [Clark et al., 2018]

Datasets

1. **SQuAD** (v1.1) [Rajpurkar et al., 2016]
2. **AddSent** [Jia and Liang, 2017]
3. **NewsQA** [Trischler et al., 2017]
4. **TriviaQA** (Wikipedia set) [Joshi et al., 2017]
5. **QAngaroo** (WikiHop) [Welbl et al., 2018]
6. **MS MARCO** (v2) [Nguyen et al., 2016]
7. **NarrativeQA** (summary) [Kočiský et al., 2018]
8. **MCTest** (160 + 500) [Richardson et al., 2013]
9. **RACE** (middle + high) [Lai et al., 2017]
10. **MCScript** [Ostermann et al., 2018]
11. **ARC Easy** [Clark et al., 2018]
12. **ARC Challenge** [Clark et al., 2018]

Answer extraction

Description

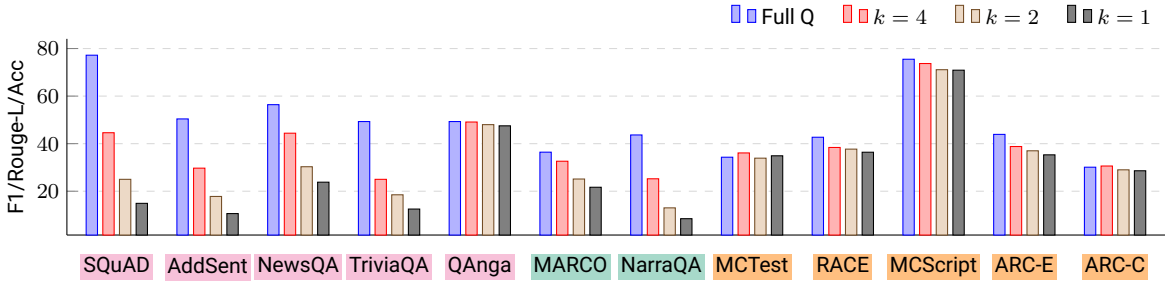
Multiple choice

Baseline Models

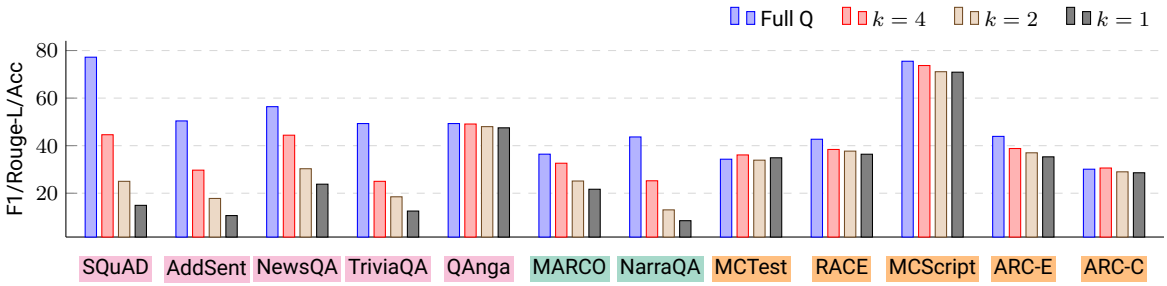
- **Bi-directional Attention Flow** (BiDAF) [Seo et al., 2017]
  - for the answer extraction and description styles
  - previous SOTA on SQuAD
- **Gated Attention Reader** (GAR) [Dhingra et al., 2017]
  - for the multiple choice style
  - previous SOTA on CNN/Daily Mail [Hermann et al., 2015] and Who-did-What [Onishi et al., 2016]

Hyperparameters are tuned on each dataset

Analysis: Questions with the first k tokens

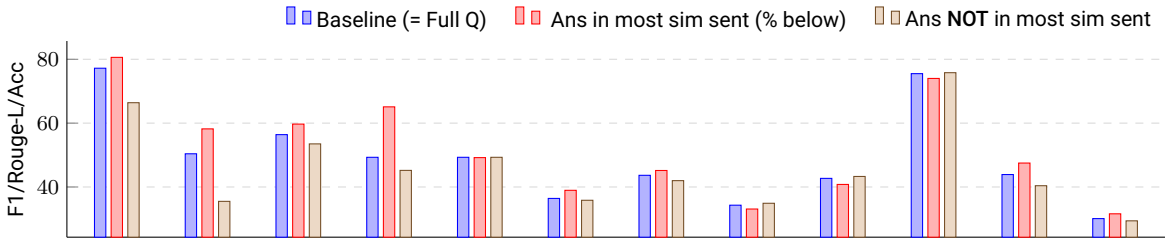


Analysis: Questions with the first k tokens

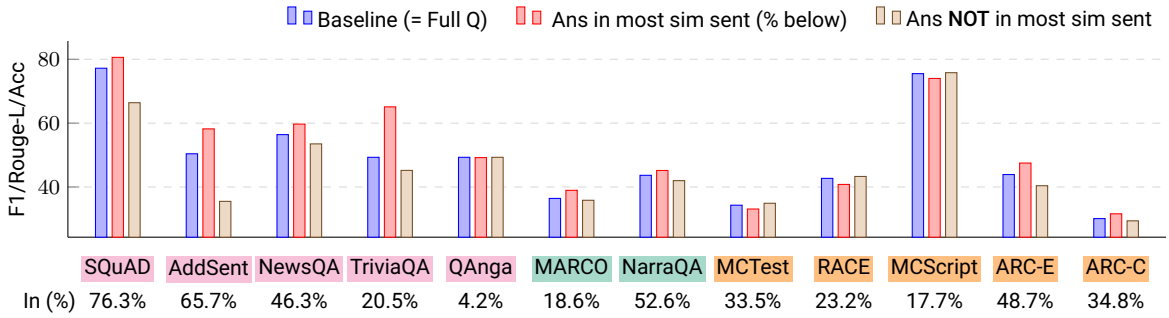


- QAngaroo : a small gap between Full and $k = 1$
- The multiple choice datasets: not enough high performance to see its difference?

Analysis: Answer in the most similar sentence

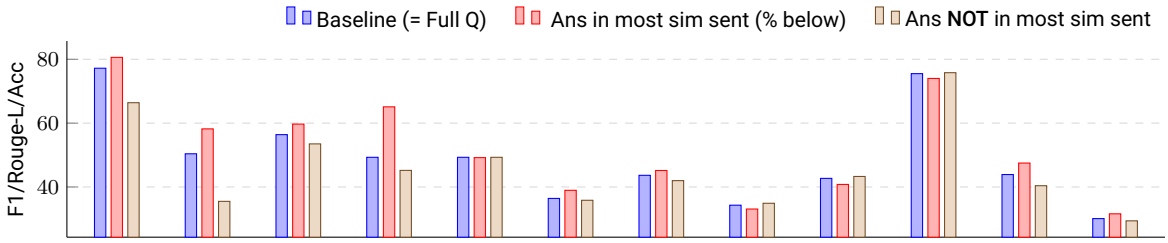


Analysis: Answer in the most similar sentence



- Answer in the most similar sentence: improve the performance in some datasets

Analysis: Answer in the most similar sentence

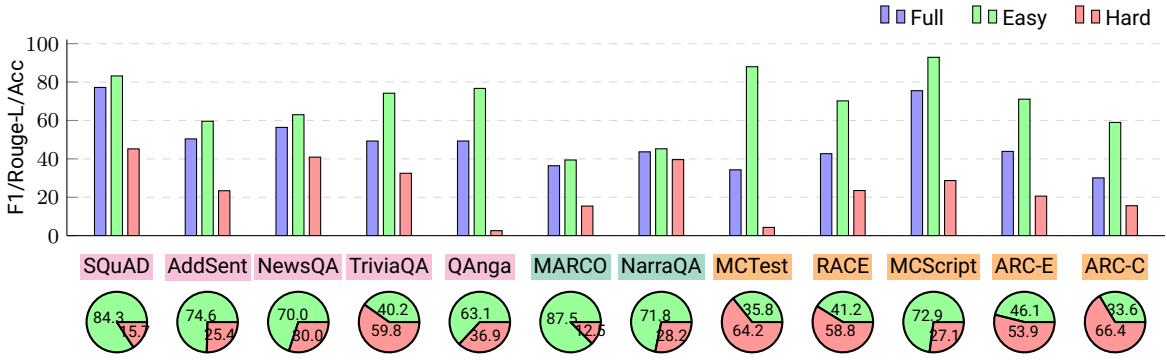


- Answer in the most similar sentence: improve the performance in some datasets
- SQuAD AddSent NewsQA NarraQA and ARC-E: >45%

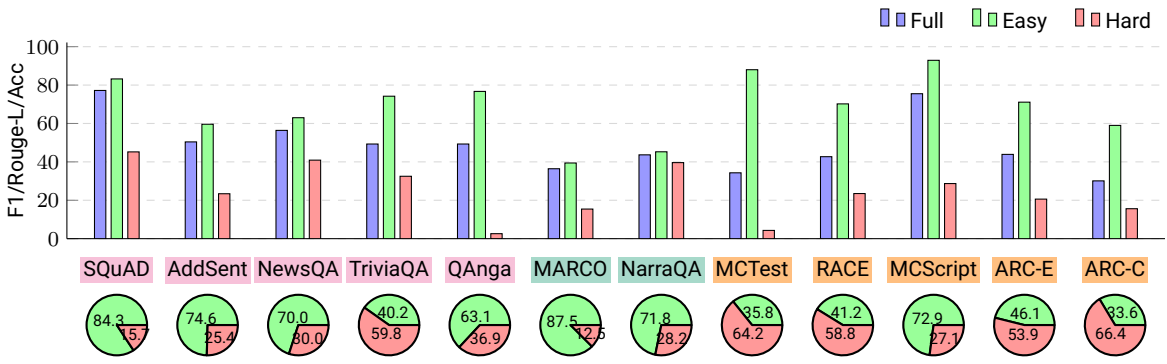
Easy and *Hard* Subsets

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

Easy and *Hard* Subsets

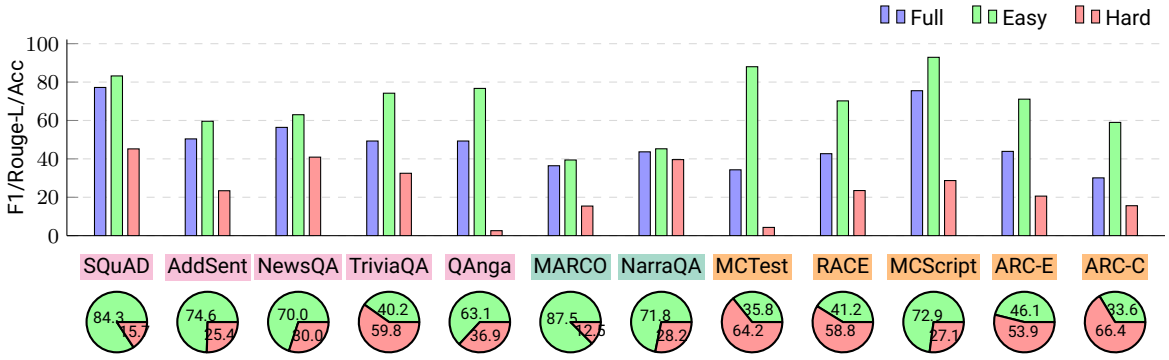


Easy and *Hard* Subsets



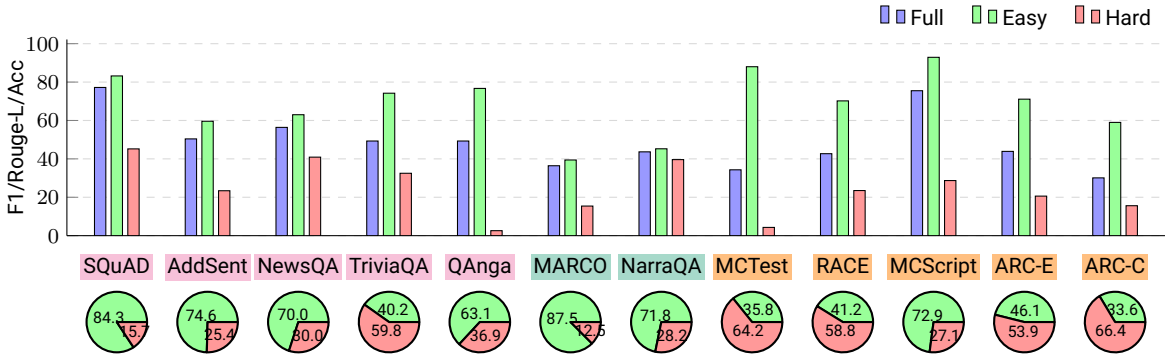
- Full → Hard : the degration ranges from 4 to 47% !

Easy and *Hard* Subsets



- Full → Hard : the degration ranges from 4 to 47% !
- QAngaroo and some multiple choice datasets: unbalanced

Easy and *Hard* Subsets



- Full → Hard : the degration ranges from 4 to 47% !
- QAngaroo and some multiple choice datasets: unbalanced
- Hard questions: multiple choice > answer extraction & description

Annotation: Motivations

Annotation: Motivations

1. How many questions are valid in each dataset?
   - E.g. Hard questions are really hard... or unanswerable?

Annotation: Motivations

1. How many questions are valid in each dataset?
   - E.g. Hard questions are really hard... or unanswerable?

2. What kinds of reasoning skills explain the Easy & Hard questions?

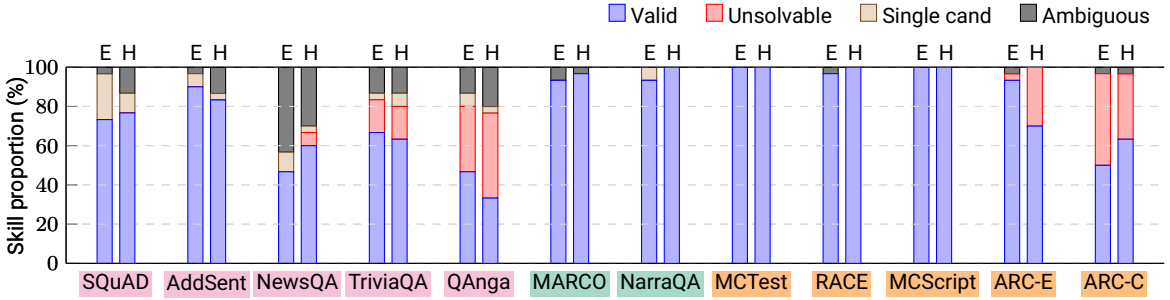
Annotation: Motivations

1. How many questions are valid in each dataset?
   - E.g. Hard questions are really hard... or unanswerable?
2. What kinds of reasoning skills explain the Easy & Hard questions?
3. Are there any differences among the question styles?

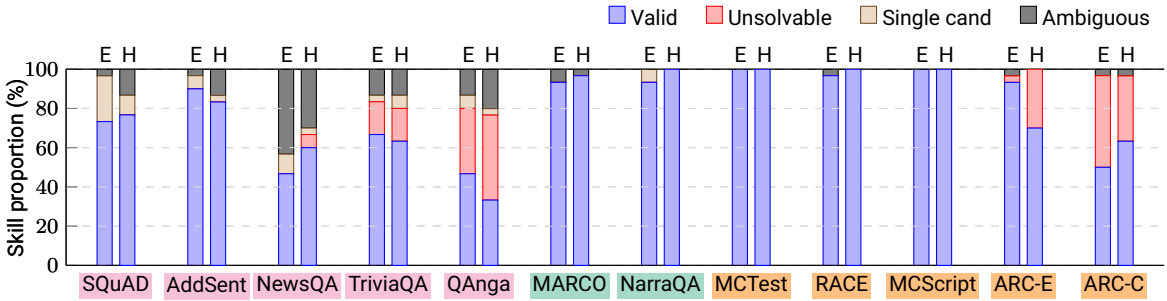
Anntation: Procedure

- Sampled 30 questions from each Easy & Hard subset
- Annotated with validity → requisite skills
- Used a skill classification in previous work [Trischler et al., 2017; Lai et al., 2017]

Anntation Results: Validity

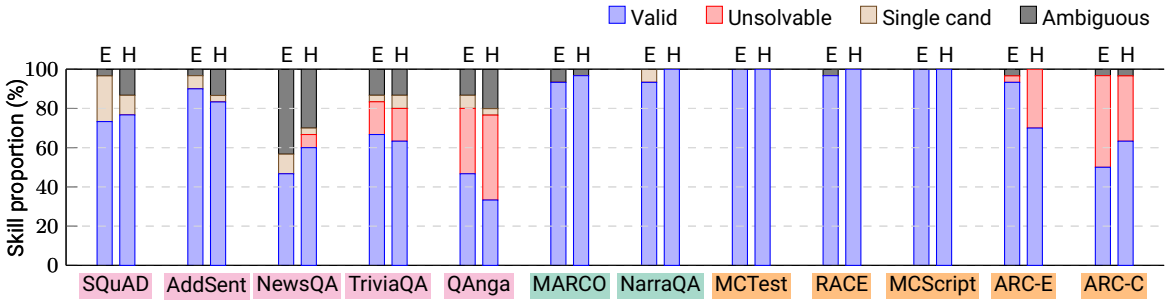


Anntation Results: Validity



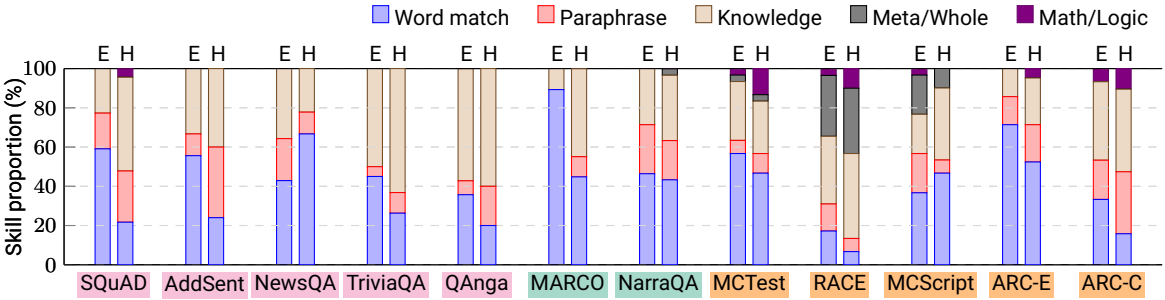
- TriviaQA, QAngaroo and ARCs: relatively high *unsolvability* by the inherent unrelatedness between the questions and their context.

Anntation Results: Validity

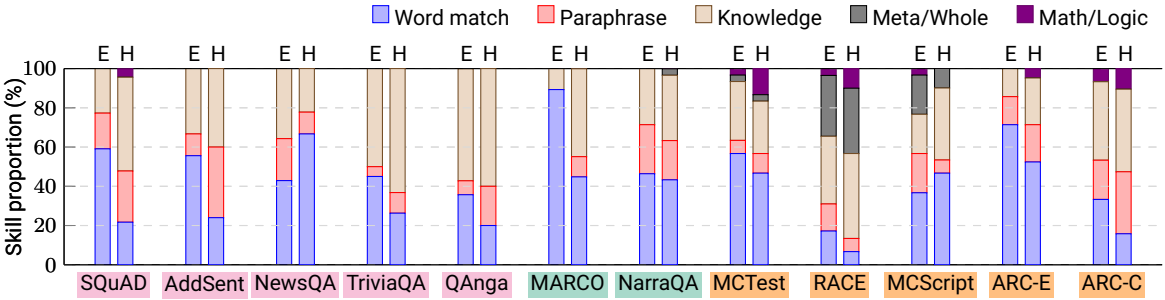


- TriviaQA, QAngaroo and ARCs: relatively high *unsolvability* by the inherent unrelatedness between the questions and their context.
- Other multiple choice datasets: high validity!

Annotation Results: Requisite Skills

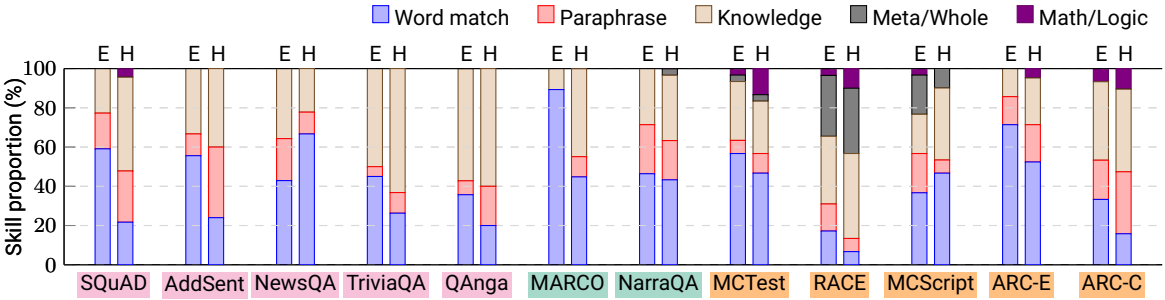


Annotation Results: Requisite Skills



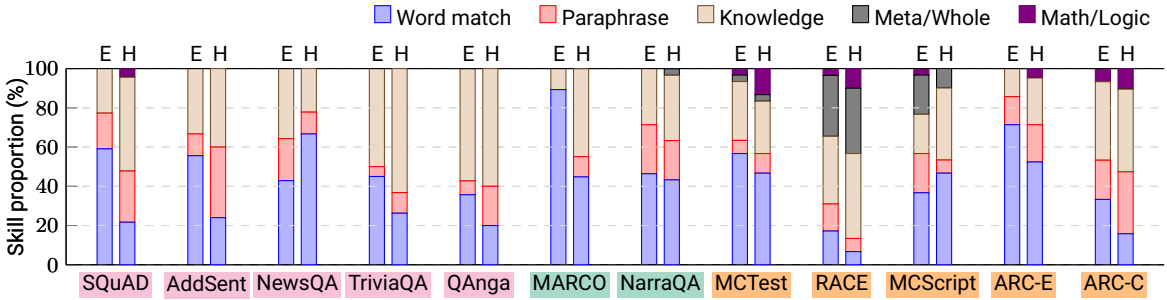
- Word matching: Easy > Hard

Annotation Results: Requisite Skills



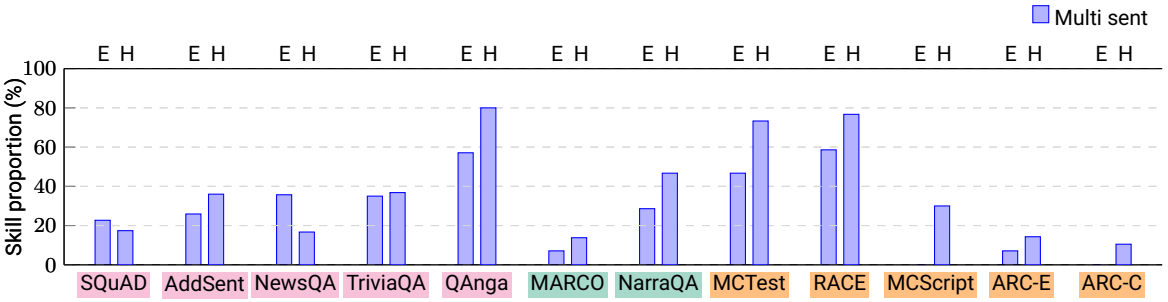
- Word matching: Easy > Hard
- Knowledge reasoning: Hard > Easy

Annotation Results: Requisite Skills



- Word matching: Easy > Hard
- Knowledge reasoning: Hard > Easy
- Meta/whole & math/logic:
  multiple choice > answer extraction & description

Annotation Results: Multiple sentence reasoning



- 10 datasets: Hard > Easy

Related Work

Related Work

Macro perspective:

- **How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks** [Kaushik and Lipton, 2018] (EMNLP)
  - Context & question ablations
  - On 5 datasets: CBT, CNN, Who-did-What, bAbI, SQuAD

Related Work

Macro perspective:

- **How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks** [Kaushik and Lipton, 2018] (EMNLP)
  - Context & question ablations
  - On 5 datasets: CBT, CNN, Who-did-What, bAbI, SQuAD

Micro perspective:

- **A Systematic Classification of Knowledge, Reasoning, and Context within the ARC Dataset** [Boratko et al., 2018] (MRQA Workshop)
  - Detailed classifications of knowledge and reasoning
  - On the ARC dataset [Clark et al., 2018]

Related Work

Macro perspective:

- **How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks** [Kaushik and Lipton, 2018] (EMNLP)
  - Context & question ablations
  - On 5 datasets: CBT, CNN, Who-did-What, bAbI, SQuAD

Micro perspective:

- **A Systematic Classification of Knowledge, Reasoning, and Context within the ARC Dataset** [Boratko et al., 2018] (MRQA Workshop)
  - Detailed classifications of knowledge and reasoning
  - On the ARC dataset [Clark et al., 2018]
- **Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability** [Sugawara et al., 2017b,a] (AAAI & ACL)
  - A set of requisite skills as evaluation metrics
  - On 6 datasets: QA4MRE, MCTest, SQuAD, Who-did-What, MARCO, NewsQA

Summary and Conclusions

Summary and Conclusions

Proposing two heuristics to identify *easy/hard* questions

- The baseline performances: Easy $>$ Hard

Summary and Conclusions

Proposing two heuristics to identify *easy/hard* questions

- The baseline performances: Easy > Hard

Annotating with varidity and skills

- Knowledge reasoning & multi sentence reasoning: Hard > Easy
- Kinds of reasoning skills: multiple choice > answer extraction & description

Summary and Conclusions

Proposing two heuristics to identify *easy/hard* questions

- The baseline performances: Easy > Hard

Annotating with varidity and skills

- Knowledge reasoning & multi sentence reasoning: Hard > Easy
- Kinds of reasoning skills: multiple choice > answer extraction & description

→ Let's focus on Hard questions!

`https://github.com/Alab-NII/mrc-heuristics`

Looking for next summer internship position & full-time research position from 2020!

Thank you 😊😊😊

References I

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A systematic classification of knowledge, reasoning, and context within the arc dataset. In Proceedings of the Workshop on Machine Reading for Question Answering, pages 60–70, Melbourne, Australia. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. CoRR, abs/1803.05457.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1832–1846. 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.

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, Copenhagen, Denmark. Association for Computational Linguistics.

References II

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611. 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, Brussels, Belgium. Association for Computational Linguistics.

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.

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, Copenhagen, Denmark. 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.

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.

References III

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA).

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.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 193–203.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In International Conference on Learning Representations.

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), pages 806–817. Association for Computational Linguistics.

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, pages 3089–3096.

References IV

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 191–200. 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.