AAAI-20 #7755

[GitHub: Alab-NII/mrc-ablation]

Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets

*saku@nii.ac.jp

*Saku Sugawara
Univ. of Tokyo

Pontus Stenetorp
UCL

Kentaro Inui
Tohoku Univ. / RIKEN

Akiko Aizawa
NII

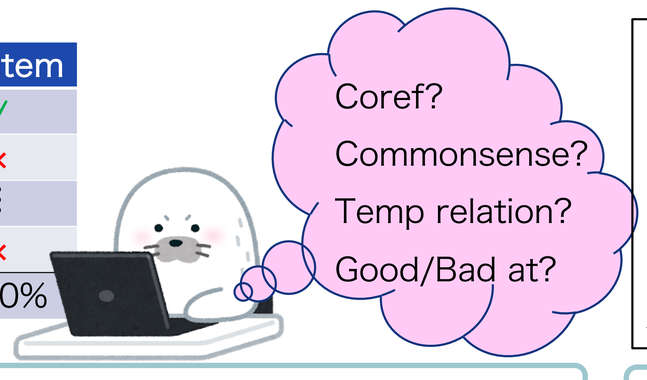
Abstract: Evaluation methods for MRC datasets

- Background: Machine reading comprehension (MRC) is a useful benchmarking task for natural language understanding.
- Issue: MRC shows the low explainability because we cannot specify what is required for answering questions.
- Solution: We propose ablation-based methods that evaluate to what degree the questions do not necessitate requisite skills.
- Results: Most of the questions correctly answered by a baseline model do not necessarily require complex understanding.
- Conclusion: MRC datasets should be carefully designed to ensure that questions can correctly evaluate the intended skills.

Background and Motivation

Issue 1: Simple Evaluation Metrics

Dataset	System
Q1	✓
Q2	×
⋮	⋮
Q1000	×
Acc.	75.0%



Because requisite skills are not identified, the questions lack explainability for NLU.

Issue 2: Low Quality Questions

Context: In *November 2014*, Sony Pictures Entertainment was targeted by hackers who released details of confidential e-mails [...]. Included within these were several memos relating to the production [...]. Eon Productions later issued a statement [...].

Question: When did hackers get into the Sony Pictures e-mail system?

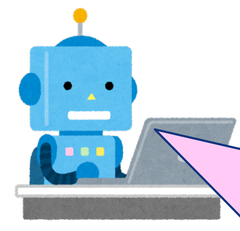
Answer: *November 2014*

too easy...

Low-quality questions prevent us from evaluating deeper language understanding.

Goal: Detailed Evaluation of Datasets

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



Intuition

If a question is correctly answered even after removing features associated with a given skill, the question does not require the skill.

Methods and Results

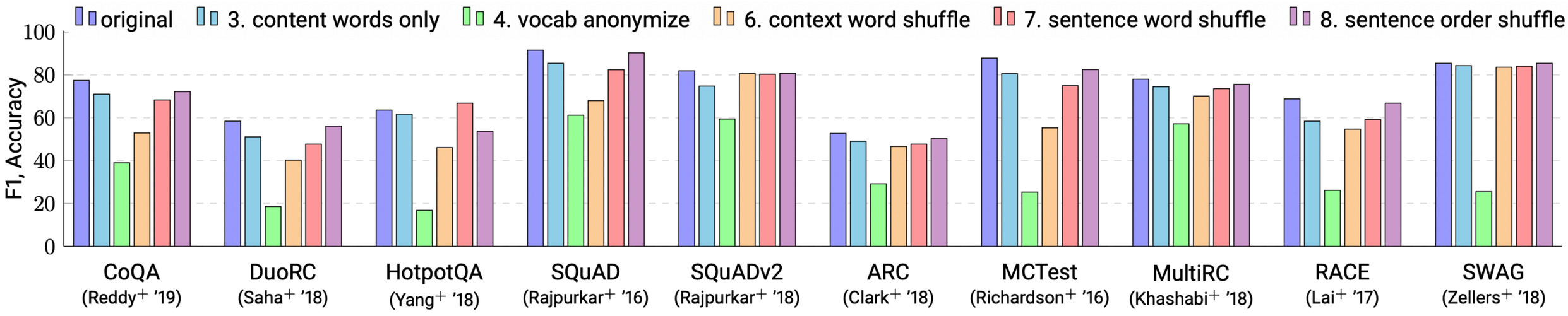
Methods and Requisite Skills

- We defined requisite skills and corresponding ablation methods.
- We used BERT-large (Devlin+ 2018) and evaluated on the ablated inputs.

1. Recognizing excluding interrogatives	7. Sentence-level compositionality
2. Recognizing content words	8. Understanding of discourse relations
3. Recognizing function words	9. Basic arithmetic operations
4. Recognizing vocabulary	10. Explicit logical reasoning
5. Attending similar context sentences	11. Resolving pronoun coreference
6. Recognizing the word order	12. Explicit causal relasions

Observations

- The baseline model exhibits remarkably high performance on some of the ablation tests: especially on 3 & 7–12 (≧90% of the original).
- When we train models on ablated inputs, the scores improved (3, 6, & 7 below).
- Ablated features are “reconstructable”?
  → Human evaluation ensured that ablated features are not required in any case.



3. Content Words Only

C: there are many persons who cannot afford to buy books, but who usually go to libraries and spend hours reading something that interests them a lot. From my point of view, literature is very important in our life. For example, reading is a means of gaining culture and enriching our knowledge in different areas .
Q: People who are fond of literature are those that ____ .
A: have much interest in reading (multiple choice)

C: [redacted] many persons [redacted] cannot afford [redacted] buy books, [redacted] usually go [redacted] libraries and spend hours reading something [redacted] interests [redacted] lot. [redacted] point [redacted] view, literature [redacted] important [redacted] life. [redacted] example, reading [redacted] means [redacted] gaining culture [redacted] enriching [redacted] knowledge [redacted] different areas .
Q: People who are fond of literature are those that ____ .
A: have much interest in reading (multiple choice)

4. Vocabulary Anonymization

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 […] @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

6. Context Words Shuffle

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 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. Jackson-ville 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
Q: What can we learn about Chris Ulmer?
A: He praises his students one by one (multiple choice)

References

- What Makes Reading Comprehension Questions Easier? (Sugawara+, EMNLP 2018)
- Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability (Sugawara+, ACL 2017)
- Prerequisite Skills for Reading Comprehension: Multi-perspective Analysis of MCTest Datasets and Systems (Sugawara+, AAAI 2017)