ACL 2022 #1566 [GitHub: nii-cl/qa_text_source_comparison]

What Makes Reading Comprehension Questions Difficult?

*saku@nii.ac.jp

*Saku Sugawara
National Institute of Informatics

Nikita Nangia Alex Warstadt Samuel R. Bowman
New York University

Takeaways

- Passage difficulty does not affect question difficulty in crowdsoucing MRC data.
- Selecting a diverse sef of passage can help ensure a diverse range of reasoning types.
- Adversarial data collection has a risk to encourage workers to focus on writing only a few specific types of questions (e.g., numerical reasoning).

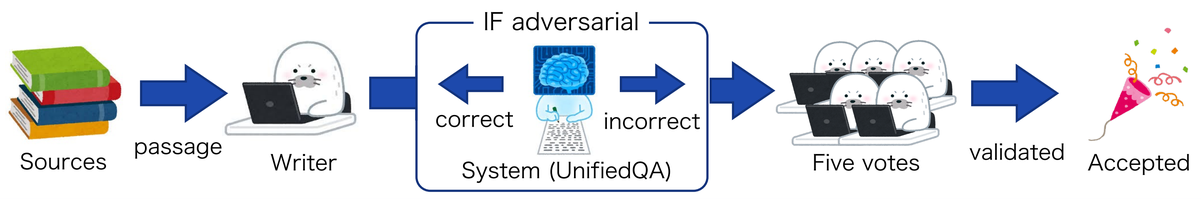
Motivation and Method

Motivation

- What aspects of text sources affect the difficulty and diversity of NLU examples?
- We analyze how question difficulty and type are affected by linguistic aspects of passages.

Data Collection

- Multiple-choice MRC
- Either standard or adversarial collection



Stats & Systems

- We collect 4,340 questions (310 Qs * 7 sources * standard or adversarial)
- About 90% of them are validated
- RoBERTa large * 4 systems
- DeBERTa large & xlarge * 4 systems
- Zero-short performance is reported

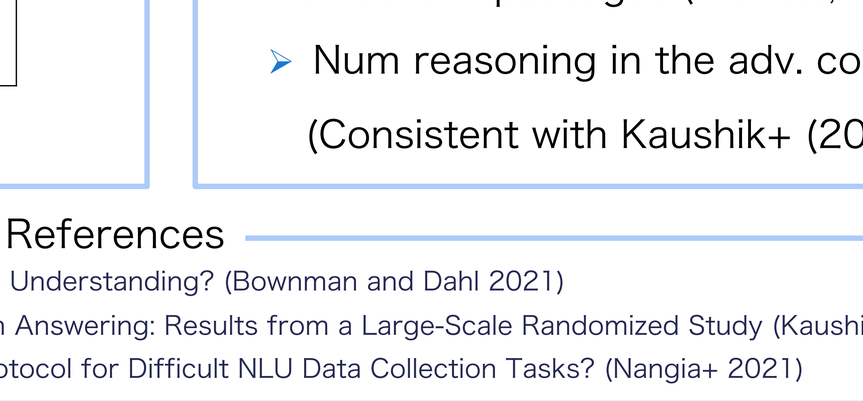
Results and Analyses

Question Difficulty vs. Linguistic Aspects

Source	Method	High-agreement portion Human	UniQA	DeBERTa	M-Avg.	Δ
MCTest	Dir.	95.0	71.5	88.2	81.5	**13.5**
	Adv.	**96.5**	27.9	78.6	68.2	**28.3**
	Total	**95.8**	49.3	83.3	74.7	**21.1**
Gutenberg	Dir.	92.8	75.0	88.5	83.4	9.4
	Adv.	87.5	28.3	82.6	72.9	14.6
	Total	90.3	53.1	85.7	78.4	11.9
Slate	Dir.	90.7	74.6	91.7	87.0	3.8
	Adv.	92.9	27.9	76.0	**73.8**	19.1
	Total	91.8	52.6	84.3	**80.8**	11.0
RACE	Dir.	95.4	74.8	90.4	84.6	10.8
	Adv.	94.3	31.0	73.8	67.3	27.0
	Total	94.9	53.3	82.2	76.1	18.8
ReClor	Dir.	**96.9**	79.6	91.1	84.4	12.5
	Adv.	88.8	32.4	74.5	71.3	17.5
	Total	93.2	58.1	83.5	78.5	14.8
Wiki. Sci.	Dir.	95.8	79.0	94.9	87.3	8.5
	Adv.	92.8	29.4	77.2	68.3	24.5
	Total	94.4	56.3	86.8	78.6	15.8
Wiki. Arts	Dir.	91.5	77.0	92.5	**88.1**	3.4
	Adv.	91.4	25.8	75.8	71.7	19.7
	Total	91.5	52.3	84.5	80.2	11.2
All sources	Dir.	94.0	75.9	91.0	85.2	8.8
	Adv.	92.0	29.0	76.9	70.5	21.5
	Total	93.1	53.6	84.3	78.2	14.9

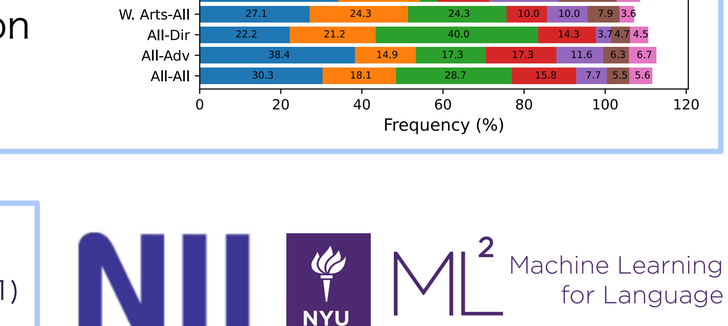
- Counter-intuitive results: easy passages (e.g., MCTest) yield difficult questions (i.e., larger human–model performance gap Δ)
- Human performance ≈ passage readability?
- No statistically significant correlations found between passage & question difficulty!

Δ = human acc – model avg acc, easy: Δ ≤ 20%, hard: Δ ≥ 40%



Question Types

- Hard questions seem to be generic, not specific to given passages (e.g., "which of the following is correct?")
- Many "how many" questions in Hard
- Questions in Easy are more balanced (because the standard Qs are?)
- The workers focus on writing specific Q types (i.e., generic and numeric) in the adversarial data collection.



Comprehension Types

- Numeric, spatial/temporal, and logical reasoning questions are relatively difficult

Comprehension Type vs. Text Sources

- Technical documents (ReClor & Slate)
  - ➢ Logical reasoning questions
- Subjective or argumentative topics (Gutenberg, Slate, & ReClor)
  - ➢ Gestalt/author's attitude questions
- Numbers in passages (MCTest, Wiki arts)
  - ➢ Num reasoning in the adv. collection (Consistent with Kaushik+ (2021))

References

- What Will it Take to Fix Benchmarking in Natural Language Understanding? (Bownman and Dahl 2021)
- On the Efficacy of Adversarial Data Collection for Question Answering: Results from a Large-Scale Randomized Study (Kaushik+ 2021)
- What Ingredients Make for an Effective Crowdsourcing Protocol for Difficult NLU Data Collection Tasks? (Nangia+ 2021)

NII NYU ML² Machine Learning for Language