Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets

Sugawara et al. 2019 (aaai 2020)

Minbyul Jeong

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

1 Introduction

Motivation: Investigate to what extent a dataset allows unintended solutions that do not need requisite skills

(1) Solvable Questions (after removing features associated with requisite skills)
→ do not need requisite skills (coreference resolution, commonsense reasoning)
(2) Unsolvable Questions
→ cannot deal (limitations)

Goal: Suggesting analysis methodology (semi-automated, ablation-based) for the benchmarking capacity of datasets

Experiments: 12 requisite skills & 10 machine reading comprehension datasets

1 Introduction

Original context

Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared ***to*** Saint Bernadette Soubirous ***in 1858***. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.

Anonymized context

@adv1 @prep5 @other0 @noun17 @verb2 @other0 @noun20 @punct0 @other1 @adj3 @noun21 @prep1 @noun22 @other2 @noun23 @period0 @other3 @verb2 @other1 @noun24 @prep1 @other0 @noun20 @prep6 @noun25 @punct0 @noun26 @wh0 @other0 @noun7 @noun8 @adv3 @verb4 ***@prep4*** @noun27 @noun28 @noun29 ***@prep2*** ***@num0*** @period0 @prep6 @other0 @noun30 @prep1 @other0 @adj4 @noun31 @punct3 @other2 @prep2 @other1 @adj5 @noun32 @wh1 @verb5 @prep7 @num1 @noun6 @other2 @other0 @noun4 @noun5 @punct4 @punct0 @verb2 @other1 @adj6 @punct0 @adj7 @noun33 @noun6 @prep1 @noun8 @period0

Question

To whom did the Virgin Mary allegedly appear ***in 1858*** in Lourdes France?

Anonymized question

@prep4 @wh2 @verb6 @other0 @noun7 @noun8 @adv4 @verb4 ***@prep2*** ***@num0*** @prep2 @noun25 @noun26 @period1

Baseline model's prediction before / after anonymization

Saint Bernadette Soubirous / noun27 @noun28 @noun29

Figure 1: Example of an ablation test that anonymizes context and question words, applied to a question from SQuAD v1.1 (Rajpurkar et al. 2016) with the correct answer in underscored. We found that the baseline model can achieve 61.2% F1 on SQuAD v1.1 even after the anonymization.

Original context

[...] By now you have probably heard about Chris Ulmer, the 26-year-old teacher in Jacksonville, Florida, who starts his special education 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. Ulmer recently shared a video of his teaching experience. All I could think was: how lucky these students are to have such inspirational teachers. [...]

Context with shuffled context words

[...] their with and to kids combined , t always of (has) mean problems the palsy five cerebral that communication , her standard " assess (. teacher a a now gesture Florida admiration and , much calling Ulmer to individually (of class his heard Jacksonville year special you up Chris greeting five) congratulation by give education who , them or about probably the in by each - student high , old - - have starts 26 . I s she similar reminded be ' each t and in help ' kid teacher [...]

Question

What can we learn about Chris Ulmer?

Options (the answer is in bold)

(A) He praises his students one by one. (B) He is Syona's favorite teacher. (C) He use videos to teach his students. (D) He asks his students to help each other.

Figure 2: Example of questions with shuffled context words from RACE. Although the question appears unsolvable for humans, the baseline model predicts the correct answer.

2 Problem Definition

F(X) = Y : X is **input**, Y is **output**, F is **model**, sigma is **ablation method**

Without comprehension skill s_i, F(sigma(X)) = Y is possible?

1. F(X) = Y and F(sigma(X)) ≠ Y

→ F라는 모델로 comprehension skill s 없이는 풀 수 없는 문제

: 적어도 하나의 모델이 comprehension skill s_i 없이 문제를 풀 수 있다면

문제 자체에 comprehension skill s_i를 무시하고 문제를 해결하는 unintended way가 존재한다.

2 Problem Definition

	Comprehension skill s_i	Ablation method σ_i
Reading-class	1. Recognizing question words excluding 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 question 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 in terms of unigram overlap and drop the other sentences.
	6. Recognizing the word order	Randomly shuffle all words in the context.
Reasoning-class	7. Grasping sentence-level compositionality	Randomly shuffle the words in all the sentences except the last token.
	8. Understanding of discourse relations	Randomly shuffle the order of the sentences in the context.
	9. Performing basic arithmetic operations	Replace numerical expressions (CD tag) 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 (PRP and PRP$ tags).
	12. Reasoning about explicit causality	Drop causal terms/clauses such as *because* and *therefore*.

Table 1: Example set of requisite skills $\{s_i\}$ and corresponding ablation methods $\{\sigma_i\}$. f is a model and (x, y) is a pair consisting of an input instance and its gold-standard answer. We interpret that for x s.t. $f(x) = y$, if $f(\sigma_i(x)) = y$, then x is solvable without s_i.

3 Evaluation

Ablation method \ Dataset	CoQA	DuoRC	Hotpot-QA	SQuAD v1.1	SQuAD v2.0	ARC	MCTest	Multi-RC	RACE	SWAG	Rel. avg.
Answering style	answer extraction (F1)					multiple choice (accuracy)					
Original dataset	77.4 0.0	58.4 0.0	63.6 0.0	91.5 0.0	81.9 0.0	52.7 0.0	87.8 0.0	78.0 0.0	68.8 0.0	85.4 0.0	0.0
1. Q interrogatives only	20.1 -74.0	14.2 -75.8	15.0 -76.4	15.2 -83.4	50.1 -38.9	35.6 -32.5	64.1 -27.0	52.6 -32.6	56.7 -17.5	77.1 -9.7	-46.8
2. Function words only	53.0 -31.5	5.8 -90.1	7.8 -87.8	17.4 -81.0	50.2 -38.7	44.0 -16.6	32.2 -63.3	61.9 -20.6	43.2 -37.3	68.9 -19.4	-48.6
3. Content words only	60.9 -21.3	47.9 -18.0	56.2 -11.6	80.7 -11.8	73.5 -10.3	48.0 -8.9	80.3 -8.5	74.5 -4.5	62.0 -9.8	82.6 -3.3	-10.8
4. Vocab. anonymization	39.0 -49.6	18.6 -68.2	16.8 -73.6	61.2 -33.1	59.4 -27.0	29.2 -44.6	25.3 -71.2	57.2 -26.7	26.1 -62.1	25.5 -70.1	-52.6
5. Most sim. sent. only	32.6 -57.9	35.8 -38.7	16.9 -73.4	68.5 -25.1	72.8 -11.2	43.6 -17.2	50.3 -42.7	67.9 -12.9	52.1 -24.3	85.4 -0.1	-30.4
6. Context words shuff.	29.8 -61.5	25.4 -56.6	23.6 -62.9	35.9 -60.7	52.4 -36.1	47.4 -9.9	47.2 -46.3	64.3 -17.6	51.7 -24.9	78.6 -8.0	-38.4
7. Sentence words shuff.	53.0 -31.6	35.9 -38.6	43.1 -32.2	62.1 -32.1	64.4 -21.4	46.4 -11.8	70.6 -19.6	71.4 -8.5	59.7 -13.3	80.3 -6.0	-21.5
8. Sentence order shuff.	72.2 -6.8	56.1 -4.0	53.7 -15.6	90.3 -1.3	80.7 -1.5	50.3 -4.5	82.5 -6.0	75.6 -3.0	66.8 -2.9	85.4 -0.0	-4.6
9. Dummy numerics	75.9 -1.9	57.8 -1.0	60.0 -5.6	89.5 -2.2	78.7 -3.9	49.7 -5.7	85.0 -3.2	76.2 -2.3	67.8 -1.5	85.3 -0.1	-2.8
10. Logical words dropped	76.7 -0.9	58.0 -0.7	62.1 -2.3	91.0 -0.5	80.6 -1.6	52.0 -1.3	85.3 -2.8	77.3 -1.0	67.7 -1.5	85.4 0.0	-1.3
11. Pronoun words dropped	76.5 -1.2	57.0 -2.5	63.4 -0.3	91.2 -0.2	81.8 -0.2	52.0 -1.3	86.6 -1.4	77.4 -0.8	68.3 -0.7	84.8 -0.8	-0.9
12. Causal words dropped	77.3 -0.1	58.3 -0.3	63.3 -0.5	91.2 -0.3	81.8 -0.2	52.0 -1.3	87.5 -0.4	77.6 -0.6	68.2 -0.8	85.5 0.0	-0.4

Table 2: The performances (%) of the baseline model with the ablation tests on the development set. Values in smaller font are changes (%) relative to the original baseline performance, and the rightmost column ("Rel. avg.") shows their averages.

3 Evaluation

Ablation method \ Dataset	CoQA	DuoRC	Hotpot-QA	SQuAD v1.1	SQuAD v2.0	ARC	MCTest	MultiRC	RACE	SWAG	Rel. avg.
Original dataset	77.4 0.0	58.4 0.0	63.6 0.0	91.5 0.0	81.9 0.0	52.7 0.0	87.8 0.0	78.0 0.0	68.8 0.0	85.4 0.0	0.0
Drop all Q words	6.7 -91.3	10.8 -81.6	10.0 -84.2	12.0 -86.9	50.1 -38.9	36.6 -30.6	61.6 -29.9	53.2 -31.8	55.4 -19.5	76.9 -10.0	-50.5
Drop all C words	-	-	-	-	-	40.3 -23.6	32.5 -63.0	61.7 -20.9	41.0 -40.4	71.7 -16.0	-32.8
Drop all C&Q words	-	-	-	-	-	29.9 -43.3	35.3 -59.8	57.2 -26.7	34.9 -49.3	62.1 -27.3	-41.3
Trained & evaluated on											
3′. Content words only	71.0 -8.3	51.1 -12.6	61.7 -3.0	85.4 -6.6	74.8 -8.7	49.0 -7.0	80.6 -8.2	74.5 -4.4	58.4 -15.2	84.3 -1.4	-7.5
6′. Context word shuff.	52.9 -31.7	40.2 -31.2	46.1 -27.4	68.0 -25.7	80.6 -1.7	46.6 -11.5	55.3 -37.0	70.1 -10.2	54.7 -20.5	83.6 -2.1	-19.9
7′. Sentence word shuff.	68.3 -11.8	47.7 -18.4	66.8 5.0	82.4 -9.9	80.3 -2.0	47.7 -9.6	75.0 -14.6	73.6 -5.6	59.2 -14.0	84.0 -1.6	-8.2

Table 3: Results of further analyses: the performance (%) after dropping all question ("Q") and/or context ("C") words, and that of the baseline model both trained and evaluated on the modified inputs.