Evaluating Question Answering Evaluation

Anthony Chen
University of California, Irvine

MRQA@EMNLP 2019

WORK WITH

Gabriel Stanovsky
Allen Institute for Artificial Intelligence, Seattle
University of Washington

Sameer Singh
University of California, Irvine

Matt Gardner
Allen Institute for Artificial Intelligence, Irvine

QA is Important

- Question answering has received a huge amount of community attention with (at least) 6 QA datasets published at EMNLP.

QA is Important

- Question answering has received a huge amount of community attention with (at least) 6 QA datasets published at EMNLP.
- Designed to test a variety of reading comprehension skills.

QA Evaluation is Important

- Drives research focus!
  - metrics → leaderboard rankings → community attention

QA Evaluation is Important

- Drives research focus!
  - metrics → leaderboard rankings → community attention
- Drives dataset creation!
  - Most QA datasets rely on span extraction or are multiple choice.

QA Evaluation is Important

- Drives research focus!
  - metrics → leaderboard rankings → community attention
- Drives dataset creation!
  - Most QA datasets rely on span extraction or are multiple choice.
- Important that QA metrics correlate with human judgement.

Existing Metrics for Question Answering

- Largely based on *n*-gram similarity.
  - F1, METEOR, BLEU, ROUGE

Existing Metrics for Question Answering

- Largely based on *n*-gram similarity.
  - F1, METEOR, BLEU, ROUGE
- Number of known issues.
  - Weak to paraphrases.
  - Do not leverage context or question.

Context: . . . The next day, two men, **John and Jim** (who are **drug dealers**), arrive at the apartment to pick up the package. . .

Question: Who comes to pick up the package the next day?

Gold Answers: **drug dealers, the drug dealer**

Prediction: **John and Jim**

ROUGE-L: 0

METEOR: 0

Example from the generative **NarrativeQA** dataset.

Context: . . . David got two exercise tips from his personal trainer, **tip A** and **tip B** . **Tip A** involves weight lifting, but **tip B** does not involve weight lifting . . .

Question: In which tip the skeletal muscle would not be bigger, **tip A** or **tip B**?

Gold Answers: **tip B**

Prediction: **tip A**

F1: 0.5

Example from the span-based **ROPES** dataset.

How good are existing QA metrics?

Approach


Generate Candidate Answers
Humans Score Candidate Answers
Evaluate Metrics via Correlation to Human Scores

Approach

Generate Candidate Answers

Humans Score Candidate Answers

Evaluate Metrics via Correlation to Human Scores

Datasets

Datasets

Dataset	Dataset Type	Context	Question	Gold Answer
NarrativeQA	Generative			

Datasets

Dataset	Dataset Type	Context	Question	Gold Answer
NarrativeQA	Generative	. . . While seeking water for his cattle **Travis Fox enters a little known canyon in the Arizona desert and gets captured by three men**, one of whom he recognizes as Dr. Gordon Ashe . . .	What happen to Travis Fox in the canyon?	He was captured by three men.

Datasets

Dataset	Dataset Type	Context	Question	Gold Answer
NarrativeQA	Generative			
MCScript	Multiple Choice →Generative			

Datasets

Dataset	Dataset Type	Context	Question	Gold Answer
NarrativeQA	Generative			
MCScript	Multiple Choice →Generative	One evening, I noticed my alarm clock had stopped working . . . **I removed the old batteries by lifting them up**, then I placed the new batteries in the same position . . .	Why did they throw away the old batteries?	They were no longer useful

Datasets

Dataset	Dataset Type	Context	Question	Gold Answer
NarrativeQA	Generative			
MCScript	Multiple Choice →Generative			
ROPES	Span			

Datasets

Dataset	Dataset Type	Context	Question	Gold Answer
NarrativeQA	Generative			
MCScript	Multiple Choice →Generative			
ROPES	Span	...A catalyst is a chemical that speeds up chemical reactions ...[Mark] conducts two tests, test A and test B, on an organism. **In test A he reduces catalysts from the organism, but in test B he induces catalysts** in the organism ...	Which test would see reactions taking place slower, test A or test B?	test A

Models

Models

Dataset	Dataset Type	Model
NarrativeQA	Generative	Multi-Hop Pointer Generator (MHPG)
MCScript	Multiple Choice →Generative	
ROPES	Span	

Models

Dataset	Dataset Type	Model
NarrativeQA	Generative	Multi-Hop Pointer Generator (MHPG)
MCScript	Multiple Choice →Generative	Multi-Hop Pointer Generator (MHPG)
ROPES	Span	BERT

Approach


Generate Candidate Answers
Humans Score Candidate Answers
Evaluate Metrics via Correlation to Human Scores

Part 1. Read the following passage:

...The next day, two men, John and Jim (who are drug dealers), arrive at the apartment to pick up the package...

Part 2. Read the following question, correct answer, and predicted answer:

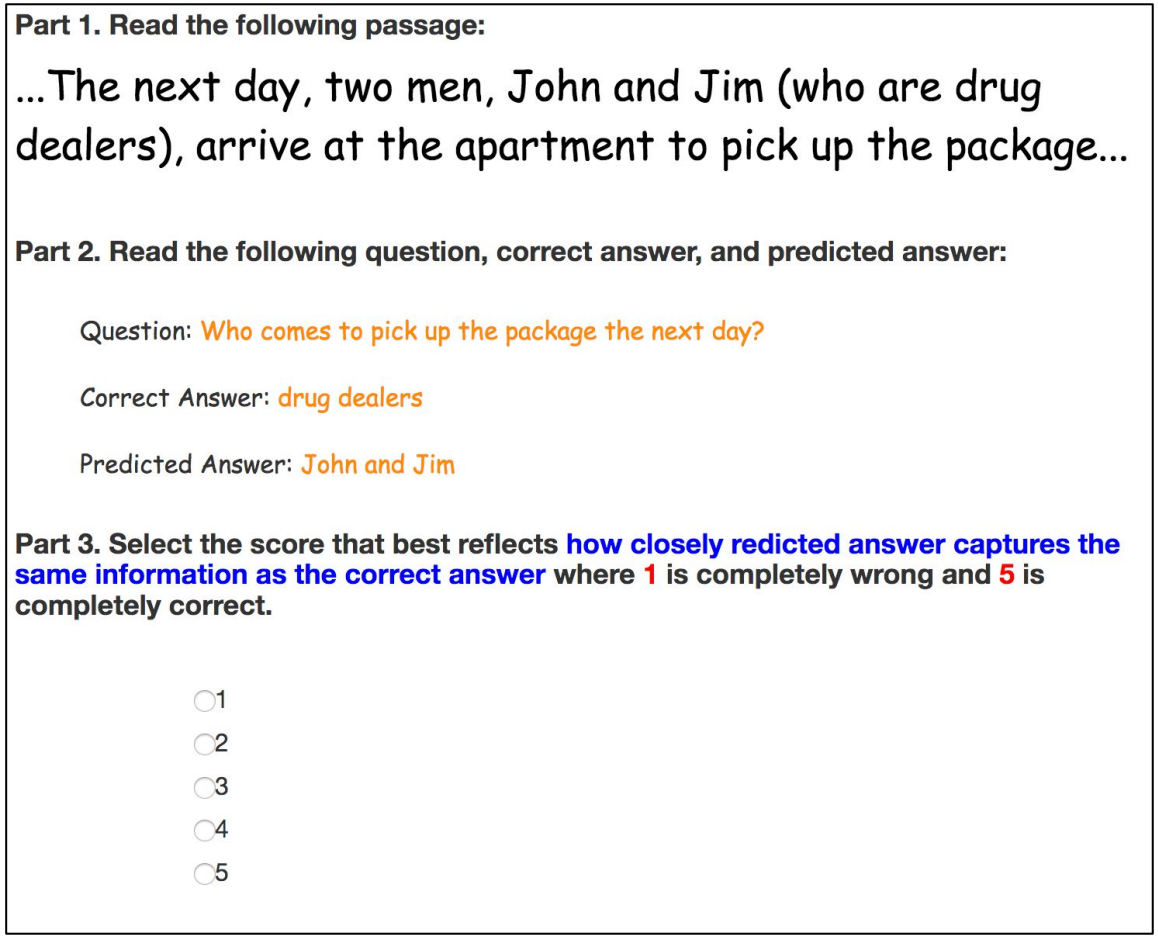
Question: Who comes to pick up the package the next day?

Correct Answer: drug dealers

Predicted Answer: John and Jim

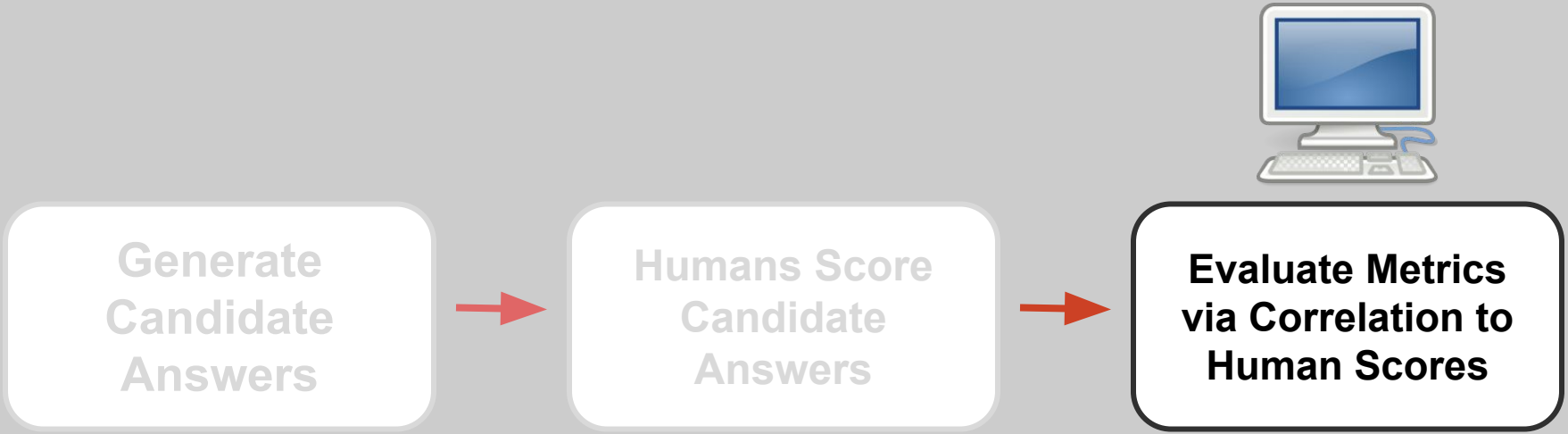
Part 3. Select the score that best reflects how closely redicted answer captures the same information as the correct answer where 1 is completely wrong and 5 is completely correct.

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5



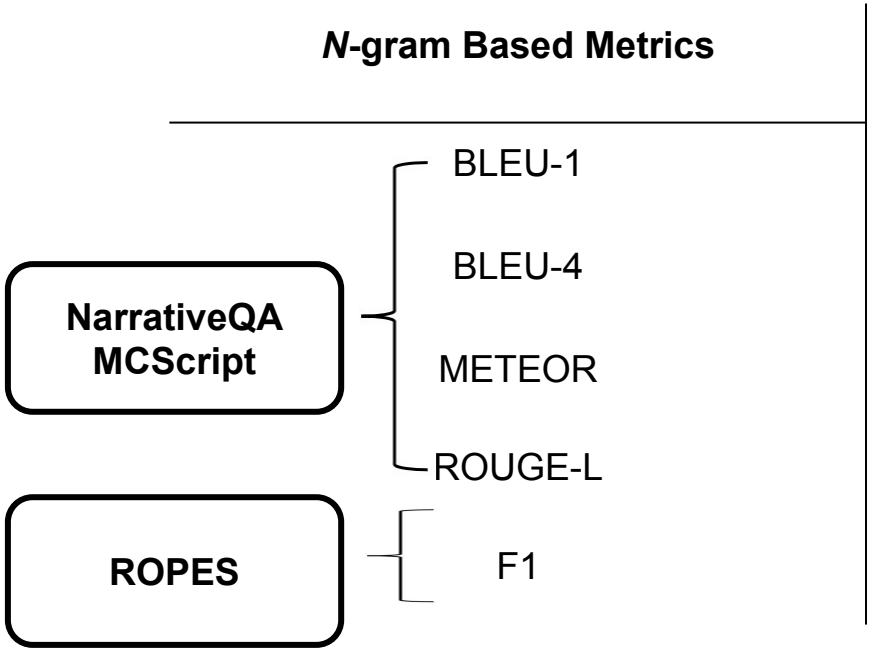
- Collect ~500 annotations per dataset.
- Each candidate answer scored by two annotators and their scores are merged.

Approach

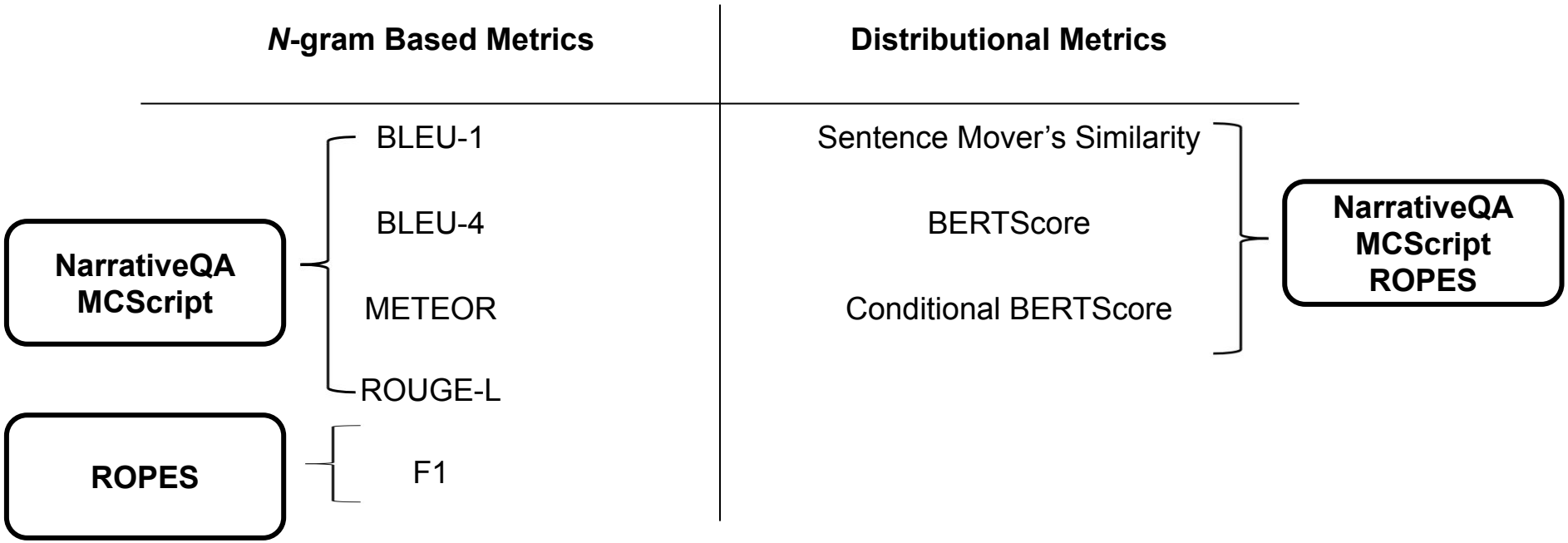

Generate Candidate Answers
Humans Score Candidate Answers
Evaluate Metrics via Correlation to Human Scores

Metrics

Metrics


N-gram Based Metrics
NarrativeQA
MCScript
BLEU-1
BLEU-4
METEOR
ROUGE-L
ROPES
F1

Metrics


N-gram Based Metrics
Distributional Metrics
BLEU-1
BLEU-4
METEOR
ROUGE-L
F1
Sentence Mover's Similarity
BERTScore
Conditional BERTScore
NarrativeQA
MCScript
ROPES
NarrativeQA
MCScript
ROPES

Metrics	NarrativeQA	MCScript	ROPES
BLEU-1			
BLEU-4			
METEOR			
ROUGE-L			
Sentence Mover's Similarity			
BERTScore			
Conditional BERTScore			
F1			

Results presented are Spearman correlations. "-" indicates the metric is not used for the dataset.

Metrics	NarrativeQA	MCScript	ROPES
BLEU-1	0.61		
BLEU-4	0.56		
METEOR	**0.75**		
ROUGE-L	0.70		
Sentence Mover's Similarity			
BERTScore			
Conditional BERTScore			
F1	-		

Results presented are Spearman correlations. "-" indicates the metric is not used for the dataset.

Metrics	NarrativeQA	MCScript	ROPES
BLEU-1	0.61	0.44	
BLEU-4	0.56	0.43	
METEOR	**0.75**	**0.64**	
ROUGE-L	0.70	0.57	
Sentence Mover's Similarity			
BERTScore			
Conditional BERTScore			
F1	-	-	

Results presented are Spearman correlations. "-" indicates the metric is not used for the dataset.

- Results for *n*-gram based metrics are consistent across datasets.
- For generative QA evaluation, use METEOR.
- "More generative" MCScript is harder for all metrics.

Metrics	NarrativeQA	MCScript	ROPES
BLEU-1			
BLEU-4			
METEOR	**0.75**	**0.64**	
ROUGE-L			
Sentence Mover's Similarity	0.47	0.48	
BERTScore	0.73	0.40	
Conditional BERTScore	0.74	0.41	
F1	-	-	

Results presented are Spearman correlations. "-" indicates the metric is not used for the dataset.

- Sentence Mover's Similarity does worse across the board.
- BERTScore significantly worse on MCScript.
  - Difficulty handling answers of significantly different lengths.

Metrics	NarrativeQA	MCScript	ROPES
BLEU-1			
BLEU-4			
METEOR	**0.75**	**0.64**	
ROUGE-L			
Sentence Mover's Similarity	0.47	0.48	
BERTScore	0.73	0.40	
Conditional BERTScore	0.74	0.41	
F1	-	-	

Results presented are Spearman correlations. "-" indicates the metric is not used for the dataset.

- Sentence Mover's Similarity does worse across the board.
- BERTScore significantly worse on MCScript.
  - Difficulty handling answers of significantly different lengths.

Question: Is John tired after running?

Reference: Yes, John is tired after running.

Candidate: Yes.

Metrics	NarrativeQA	MCScript	ROPES
BLEU-1			-
BLEU-4			-
METEOR			-
ROUGE-L			-
Sentence Mover's Similarity			0.37
BERTScore			0.44
Conditional BERTScore			0.43
F1			**0.591**

Results presented are Spearman correlations. "-" indicates the metric is not used for the dataset.

- F1 is the best metric for ROPES, but far from perfect.
  - Need to consider the answer-type.

Metrics	NarrativeQA	MCScript	ROPES
BLEU-1	0.61	0.44	-
BLEU-4	0.56	0.43	-
METEOR	**0.75**	**0.64**	-
ROUGE-L	0.70	0.57	-
Sentence Mover's Similarity	0.47	0.48	0.37
BERTScore	0.73	0.40	0.44
Conditional BERTScore	0.74	0.41	0.43
F1	-	-	**0.591**

Results presented are Spearman correlations. "-" indicates the metric is not used for the dataset.

- BERTScore and Sentence Mover's Similarity are SOTA for translation and summarization respectively but fall behind for QA.
  - Metrics don't necessarily transfer across NLP tasks!!!
- There exists a significant gap between the best metrics and human judgement!

What’s next?

- We need a larger and more comprehensive evaluation dataset.

What’s next?

- We need a larger and more comprehensive evaluation dataset.
  - Candidate answers from more datasets.
  - Candidate answers from more models.
  - Lots of annotations will be needed.

What’s next?

- We need a larger and more comprehensive evaluation dataset.
    - Candidate answers from more datasets.
    - Candidate answers from more models.
    - Lots of annotations will be needed.
- We need a new QA metric.

Towards a Better Metric

- Should incorporate the context and question.

Towards a Better Metric

- Should incorporate the context and question.
- Needs to generalize *across* datasets.

Towards a Better Metric

- Should incorporate the context and question.
- Needs to generalize *across* datasets.
- Will likely need to be ***learned.***
    - Precedent in image captioning Cui et al (2018).

Towards a Better Metric

- Should incorporate the context and question.
- Needs to generalize *across* datasets.
- Will likely need to be ***learned.***
    - Precedent in image captioning Cui et al (2018).
    - Will be difficult because we do not have much positive training data.

Conclusions

- Current metrics struggle with free-form answer evaluation and certain span evaluation.
  - This limits the complexity of (generative) QA datasets that can be created.
- We will need a better metric for QA.
  - Metrics that work on other NLP tasks don’t necessarily transfer.
- We will need a larger evaluation dataset for studying metrics.

Conclusions

- Current metrics struggle with free-form answer evaluation and certain span evaluation.
  - This limits the complexity of (generative) QA datasets that can be created.
- We will need a better metric for QA.
  - Metrics that work on other NLP tasks don't necessarily transfer.
- We will need a larger evaluation dataset for studying metrics.

Thanks for listening!

UCI nlp

Conclusions

- Current metrics struggle with free-form answer evaluation and certain span evaluation.
  - This limits the complexity of (generative) QA datasets that can be created.
- We will need a better metric for QA.
  - Metrics that work on other NLP tasks don't necessarily transfer.
- We will need a larger evaluation dataset for studying metrics.

anthony.chen@uci.edu
@_anthonychen
anthonywchen.github.io

Thanks for listening!

UCI nlp

AI2