



Evaluating Evaluation

Dan Roth

University of Pennsylvania & Amazon AWS AI

Eval4NLP @ EMNLP

November 2021

- Some thoughts on Evaluation in (some) NLP tasks
 - Evaluation is Essential
 - Can we believe what our evaluation methods tell us?
 - Not talking here about robustness, adaptation, transfer, etc.
 - Even within domain
 - Can we do better?

Outline



■ Text Correction

- I gave him a books

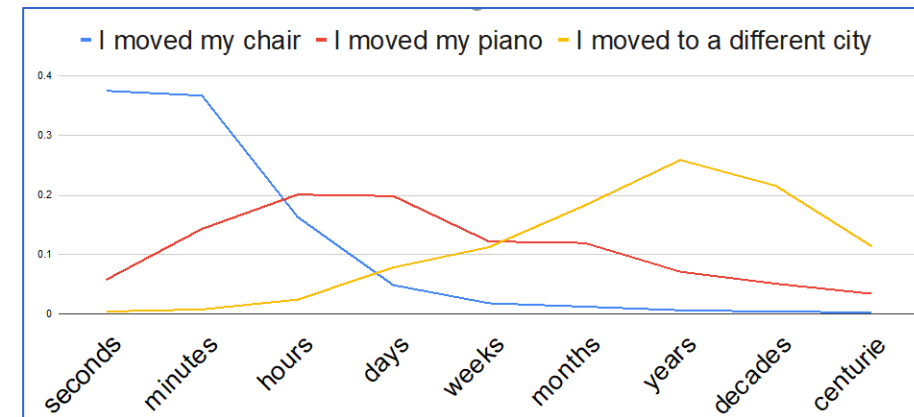


■ Summarization

- How to evaluate: a given document has multiple possible summaries
- How good is it: the statistical question



■ Evaluation in Commonsense Acquisition/Reasoning



Grammatical Error Correction (GEC)


How Good (really) are Grammatical Error Correction Systems?

Alla Rozovskaya and Dan Roth

EACL'21

When multiple golds are possible it is not clear how to evaluate the “standard” evaluation

Standard Reference-Based Evaluation for GEC



Source	The settings are very reallistic and the actors had a great performance .
Hypothesis 1	The settings are very <i>realistic</i> and the actors <i>had great</i> performance.
Reference Gold (RG)	The settings are very <u>realistic</u> and the actors <u>gave</u> a great performance.

- The set of possible golds for a given source sentence is extremely large
- Most GEC datasets contain 1 (or 2) golds for a source sentence
 - This (“**unique**”) gold is generated relative to the source sentence
 - And is **independent** of the system output

System edits: (1) reallistic -> realistic;
(2) had a great -> had great

Gold edits: (1) reallistic -> realistic;
(2) had -> gave


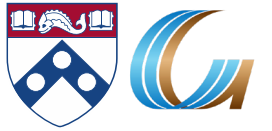
Correct edit: (1) reallistic -> realistic

Precision: $1/2=0.5$

Recall: $1/2=0.5$

- Impact:
 - **Evaluation:** Reference-based evaluation **underestimates** system performance
 - Also impacts training (which is done relative to the single reference gold)

Proposal: Evaluate Relative to Closest Gold



Source	The settings are very reallistic and the actors had a great performance .
Hypothesis 1	The settings are very <u>realistic</u> and the actors <u>had great</u> performance
Reference Gold (RG)	The settings are very <u>realistic</u> and the actors <u>gave</u> a great performance.
Closest Gold (CG) to Hypothesis 1	The settings are very <u>realistic</u> and the actors <u>had great performances</u> .

Reference Gold:

System edits: (1) reallistic -> realistic;
(2) had a great -> had great

Gold edits: (1) reallistic -> realistic;
(2) had -> gave

Correct edits: (1) reallistic -> realistic

Precision: $1/2=0.5$

Recall: $1/2=0.5$

Closest Gold:

System edits: (1) reallistic -> realistic;
(2) had a great -> had great

Gold edits: (1) reallistic -> realistic;
(2) had a great -> had great
(3) performance -> performances

Correct edits: (1) reallistic -> realistic
(2) had a great -> had great

Precision: $2/2=1.0$

Recall: $2/3=0.66$

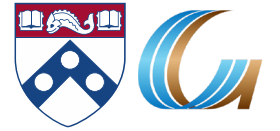
Proposal: Evaluation with Closest Golds



- Closest Golds (CGs) are generated relative to system hypotheses
 - Annotators generate a correct text that is the **closest to the system output**.
 - CGs are generated for the top hypothesis and hypotheses of lower ranks (2, 5, and 10)
- We use closest golds to evaluate system output of 4 GEC datasets
 - 2 English and 2 Russian
- We claim that evaluation relative to CGs gives true system performance

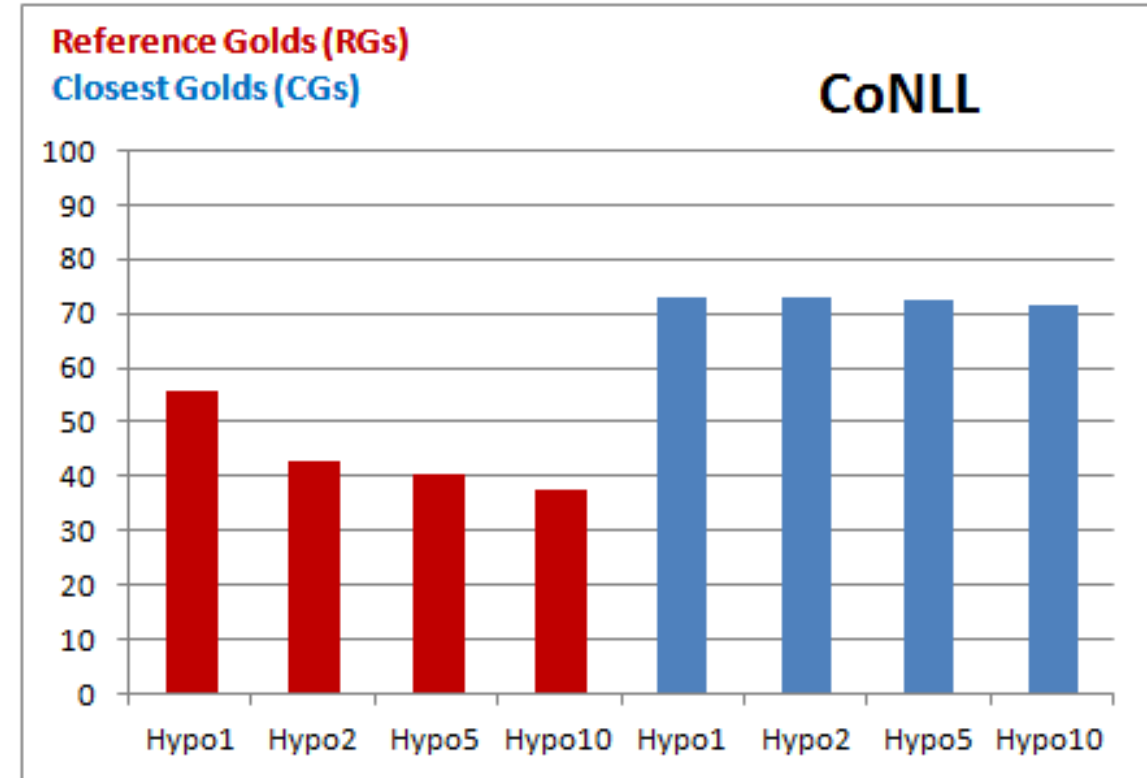
- Results:
 - The system performance, when evaluated relative to reference gold, is severely underestimated.
 - Lower rank hypotheses are often as good as top hypotheses (relative to their CGs)
 - And are more “interesting”

Key Findings



- Evaluation against CGs is significantly better than evaluation against RGs
 - This is **true** system performance
- Evaluation against RGs shows a large gap between top hypothesis and lower-ranked hypotheses.
 - Evaluation against CGs reveals very little degradation between top hypothesis and the rest
 - The reason is that lower-ranked hypotheses propose more diverse changes (e.g. lexical changes), that have lower chances of matching RGs

More analysis & Insights in the paper.
Rozovskaya & Roth EACL'21



Evaluating in multiple gold situations may not reveal the true power of a system.

Need more “Semantic Evaluation” rather than just distance from a **fixed gold**.

Summarization

1. How to evaluate?
2. Are we sure? (Statistical View)

Understanding the Extent to which Content Quality Metrics Measure the Information Quality of Summaries

Dan Deutsch, Dan Roth

CoNLL-2021

Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary

Dan Deutsch, Tanya Bedrax-Weiss, Dan Roth

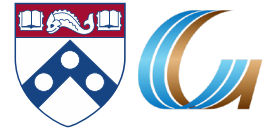
TACL-2021

A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods

Daniel Deutsch, Rotem Dror, Dan Roth

TACL-2021

Evaluating summaries

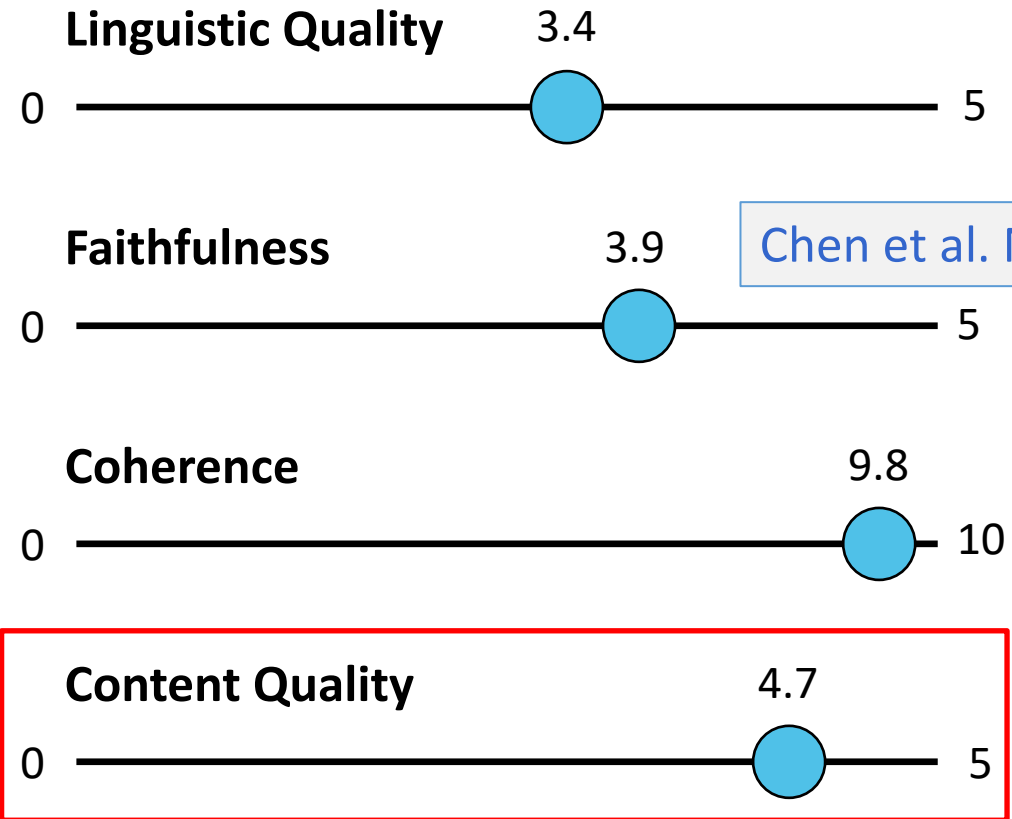


Document(s)

The Airbus A380 is a wide-body aircraft manufactured by Airbus. It is the world's largest passenger airliner. Airbus studies started in 1988 and the project was announced in 1990 to challenge the dominance of the Boeing 747 in the long haul market. The then-designated A3XX project was presented in 1994; Airbus launched the €9.5 billion (\$10.7 billion) A380 programme on 19 December 2000. The first prototype was unveiled in Toulouse on 18 January 2005, with its first flight on 27 April 2005. Difficulties in electrical wiring caused a two-year delay and the development cost ballooned to €18 billion. It obtained its type certificate from the European Aviation Safety Agency (EASA) and the US Federal Aviation Administration (FAA) on 12 December 2006. It was first delivered to Singapore Airlines on 15 October 2007 and entered service on 25 October. Production peaked at 30 per year in 2012 and 2014. However, Airbus concedes that its \$25 billion investment for the aircraft cannot be recouped. On 14 February 2019, after Emirates reduced its last orders.....

Summary

The super jumbo Airbus A380, the world's largest commercial airliner, took off Wednesday into cloudy skies over southwestern France for its second test flight. The European aircraft maker, based in the French city of Toulouse, said the second flight -- which came exactly a week after the A380's highly...



Chen et al. NAACL'21

There are different metrics to measure all of these, but we will focus on content quality

Evaluating summaries



Document(s)

The Airbus A380 is a wide-body aircraft manufactured by Airbus. It is the world's largest passenger airliner. Airbus studies started in 1988 and the project was announced in 1990 to challenge the dominance of the Boeing 747 in the long haul market. The then-designated A3XX project was presented in 1994; Airbus launched the €9.5 billion (\$10.7 billion) A380 programme on 19 December 2000. The first prototype was unveiled in Toulouse on 18 January 2005, with its first flight on 27 April 2005. Difficulties in electrical wiring caused a two-year delay and the development cost ballooned to €18 billion. It obtained its type certificate from the European Aviation Safety Agency (EASA) and the US Federal Aviation Administration (FAA) on 12 December 2006. It was first delivered to Singapore Airlines on 15 October 2007 and entered service on 25 October. Production peaked at 30 per year in 2012 and 2014. However, Airbus concedes that its \$25 billion investment for the aircraft cannot be recouped. On 14 February 2019, after Emirates reduced its last orders.....

Summary

The superjumbo Airbus A380, the world's largest commercial airliner, took off Wednesday into cloudy skies over southwestern France for its second test flight. The European aircraft maker, based in the French city of Toulouse, said the second flight -- which came exactly a week after the A380's highly...



Reference Summary

The European Airbus A380 flew its maiden test flight from France 10 years after design development started. The A380 super-jumbo passenger jet surpasses the Boeing 747 and breaks their monopoly. Airlines worldwide have placed orders but airports may need modification to accommodate the weight and width of the A380. U.S. airlines have not placed an order. Airbus has fallen behind in production and ...



Since there are **multiple golds** comparing a generated summary to a **single reference** limits our ability to evaluate the quality of a generated summary.

- Content quality: Does the summary say the “right” things?
- What is “right?”

Information Need: Describe developments in the production and launch of the Airbus A380

Launch

The European Airbus A380 flew its maiden test flight from France 10 years after design development started.

Production

The A380 super-jumbo passenger jet surpasses the Boeing 747 and breaks their monopoly. Airlines worldwide have placed orders but airports may need modification to accommodate the weight and width of the A380. U.S. airlines have not placed an order. Airbus has fallen behind in production and a backlog of orders has developed. Airbus must sell at least 250 planes to break even financially. The A380 is overweight and modifications to meet the weight requirements impacted the budget. Additional test flights are planned.

The European Aeronautic Defence and Space Co., which owns 80 percent of Airbus, says the A380 program will break even at about 250 sales. The A380 'superjumbo', which will be presented to the world in a lavish ceremony in southern France on Tuesday, will be profitable from 2008, its maker Airbus told the French financial newspaper La Tribune. Federal Express has ordered 10 of the planes. The A380 will take over from the Boeing 747 as the biggest jet in the skies. French President Jacques Chirac immediately hailed the "total success of the first test flight of the Airbus A380."

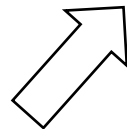
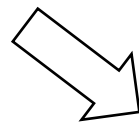
- Data-driven definition of “good” quality content
- Assumption: if a candidate summary is similar to a reference summary, it’s good
 - Reduced problem to comparing two summaries

Candidate Summary

The superjumbo Airbus A380, the world's largest commercial airliner, took off Wednesday into cloudy skies over southwestern France for its second test flight. The European aircraft maker, based in the French city of Toulouse, said the second flight -- which came exactly a week after the A380's highly...

Reference (Human) Summary

The European Airbus A380 flew its maiden test flight from France 10 years after design development started. The A380 super-jumbo passenger jet surpasses the Boeing 747 and breaks their monopoly. Airlines worldwide have placed orders but airports may need modification to ...



Reference-Based Evaluation Metric

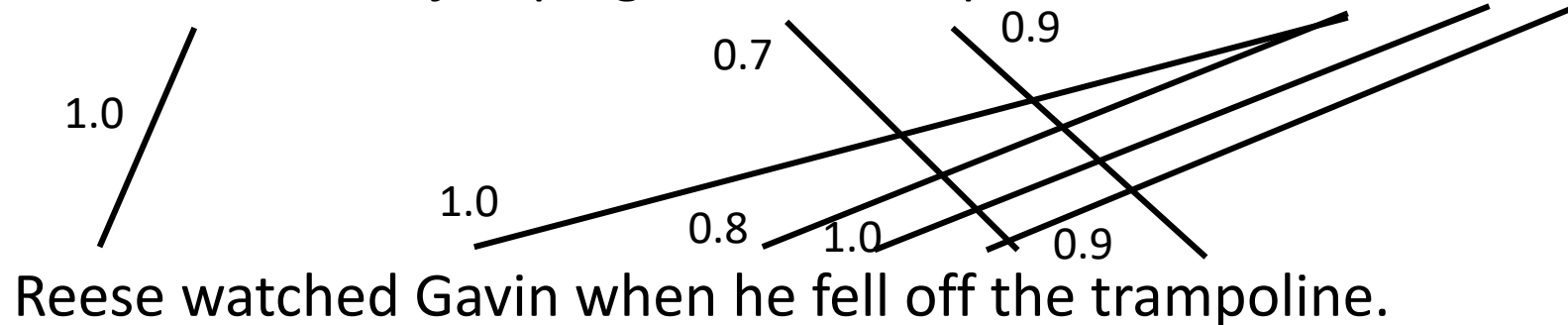


Current Methodology: Alignment-Based



- Both ROUGE/BERTScore can be cast as computing weighted alignments
 - Many-to-many, some tokens may be unaligned
 - Weight of the alignment = sum of the weight of the edges

Gavin and Reese were jumping on the trampoline when Gavin fell off.



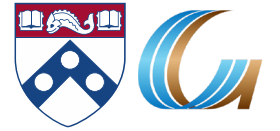
Weight of alignment = 6.3

What do ROUGE & BERTScore Measure?



- One way to understand it is by comparing to the Pyramid Method
 - The Pyramid Method is the gold-standard for manually comparing the content of two summaries (Nenkova and Passonneau, 2004)
 - Artifact of annotation: pairs of phrases that contain the same information

The Pyramid Method



Reference Summaries

The European Airbus A380 flew its maiden test flight from France 10 years after design development started. The A380 super-jumbo passenger jet surpasses the Boeing 747 and breaks their monopoly. Airlines worldwide have placed orders but airports may need modification to accommodate the weight and width of the A380. U.S. airlines have not placed an order. Airbus has fallen behind in production and a backlog of orders has developed. Airbus must sell at least 250 planes to break even financially. The A380 is overweight and modifications to meet the weight requirements impacted the budget. Additional test flights are planned.

The largest passenger airliner ever built, the Airbus 380(A380), took off on its maiden four-hour flight on April 27, 2005 in France. The European company, Airbus, is the newest competitor with the Boeing Company. The A380 is designed to carry 555 passengers, but can be expanded to 800 seats. Airbus stresses the plane's fuel efficiency. Its first test flight was successful. Orders for 149 aircraft from airlines and freight companies have been received. No US airline has ordered the jet yet. First commercial deliveries to Singapore Airlines are scheduled for 2006.

1. Exhaustively annotate SCUs (**summary content units**) in the reference
2. Identify occurrences of those SCUs in the candidate

Candidate Summary

The European Aeronautic Defence and Space Co., which owns 80 percent of Airbus, says the A380 program will break even at about 250 sales. The A380 'superjumbo', which will be presented to the world in a lavish ceremony in southern France on Tuesday, will be profitable from 2008, its maker Airbus told the French financial newspaper La Tribune. Federal Express has ordered 10 of the planes. The A380 will take over from the Boeing 747 as the biggest jet in the skies. French President Jacques Chirac immediately hailed the "total success of the first test flight of the Airbus A380."

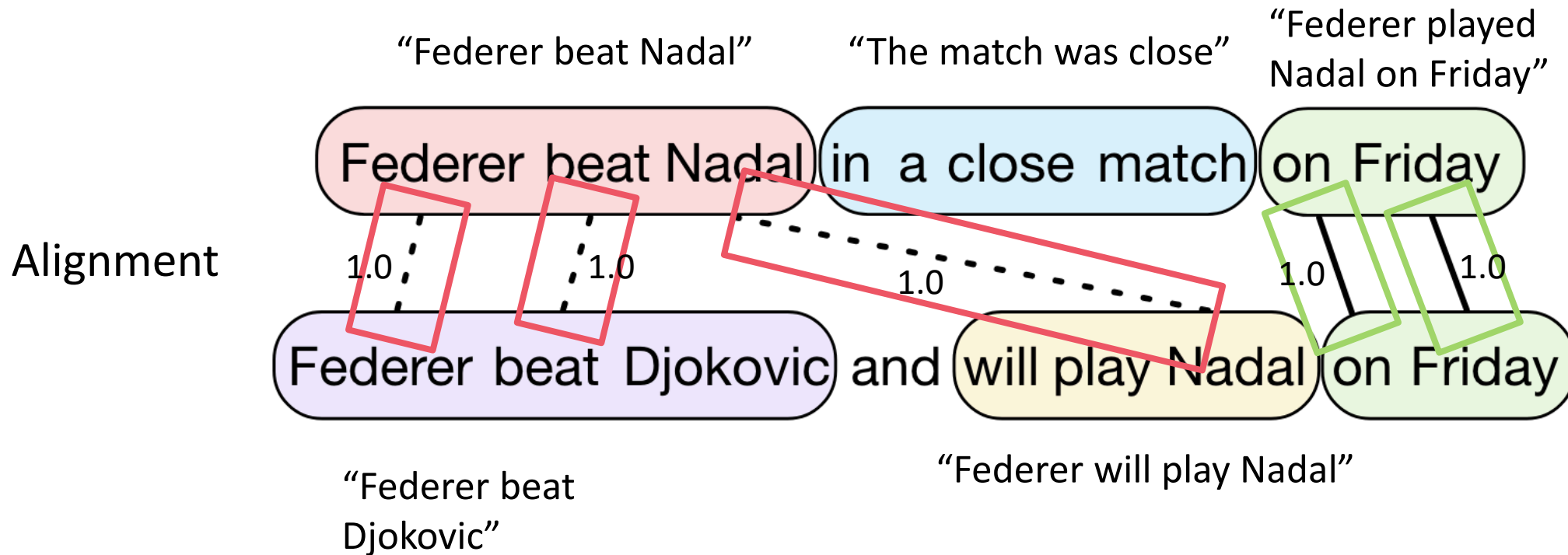
All identical information between a reference and a candidate is annotated

Exhaustive annotation => Anything not annotated **does not** have the same information

Comparison Alignment to SCUs



Colors = SCUs

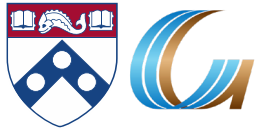


Count what portion of the weight of the alignment is between phrases that have the same information

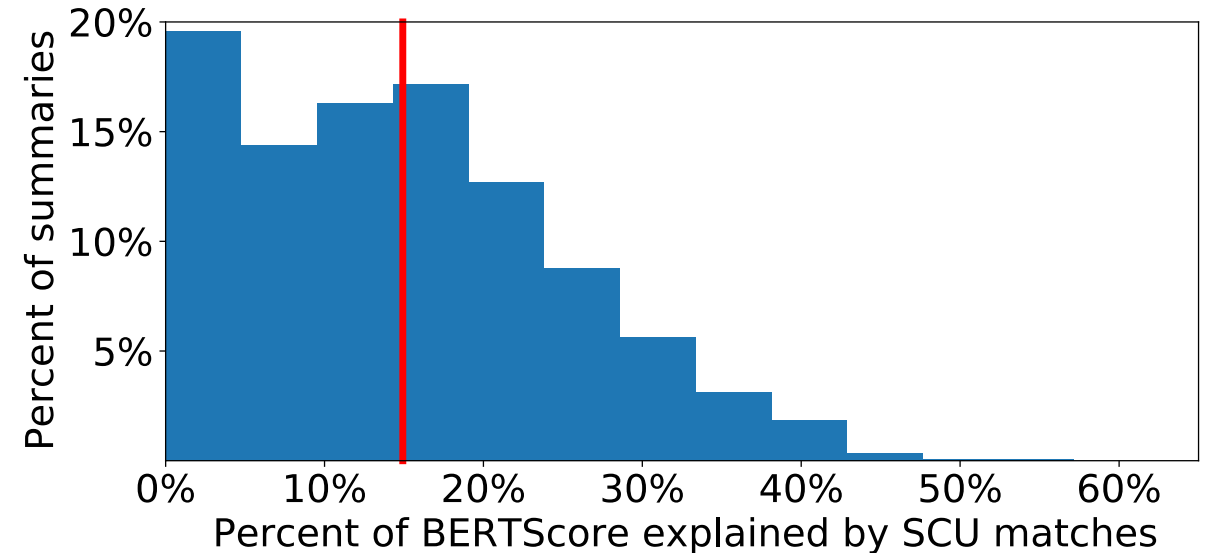
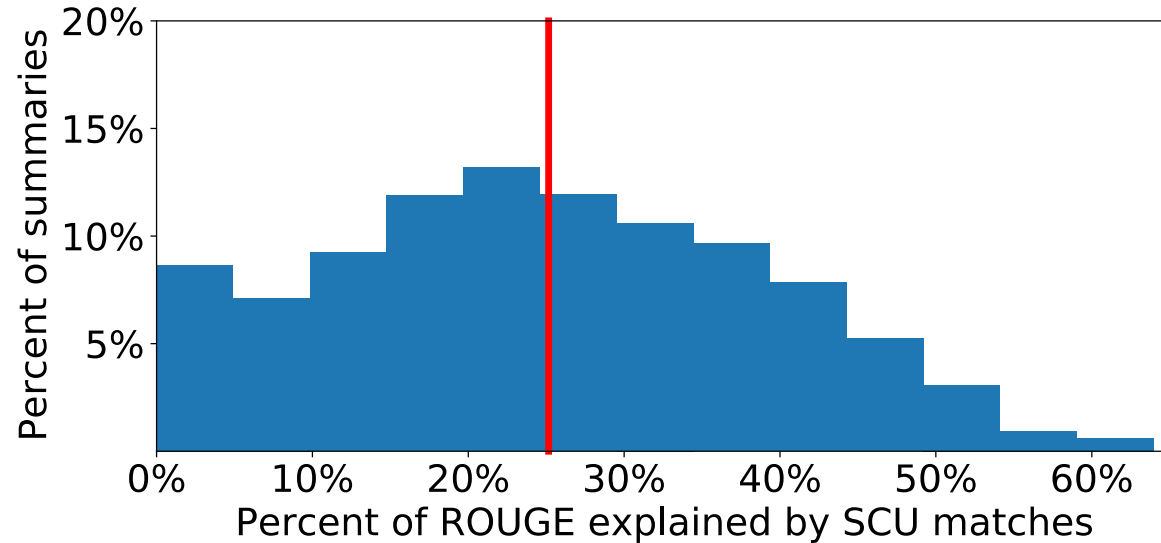
2 / 5

Most of the score is coming from phrases that don't have the same information

Comparison to SCUs



Calculate percent of metric scores explained by SCU matches on TAC 2008



On average, 25% of ROUGE and 15% of BERTScore values on TAC '08 come from phrases with the same information

⇒ 75% and 85% come from phrases that mean different things

Conclusion => ROUGE/BERTScore largely do not measure information overlap

[Deutsch & Roth'21] also analyzes the different categories of token matches.

So, how should we evaluate?

Towards

Question-Answering

as

an Automatic Metric for Evaluating the
Content Quality of a Summary



- ROUGE and BERTScore represented summaries as bags-of-words or BERT embeddings
- Two summaries compared via lexical overlap or cosine similarities
- Instead, could we represent a summary using QA pairs and compare content via answering questions against a summary?
 - Propose and analyze a QA-based evaluation metric called QAEval
 - Demonstrate state-of-the-art results on some evaluations
 - Identify performance bottlenecks = areas for future work
 - Estimate upper-bound performance if bottlenecks are addressed

These represent the information of the reference summary. We can throw away the reference summary text now.

Why do we need a reference at all?
We don't – if we can determine salience well



Steps

1. Answer Selection: select answers that will generate questions
2. Question Generation: generate wh-questions that target each answer using a learned model
3. Question Answering: answer questions against the candidate summary using a learned model
4. Answer verification: determine if the answers are correct or not
5. Final score: calculate portion of questions answered correctly

Discuss and evaluate each step, followed by the overall metric

Reference Summary

Yesterday, **Nadal** lost to **Federer**

QA Pairs

Who lost to Federer yesterday? **Nadal**

Who did Nadal lose to yesterday? **Federer**

Predictions

Who lost to Federer yesterday? Nadal

Who did Nadal lose to yesterday? Roger Federer

Answer Verification

Ground-truth: Nadal Prediction: Nadal EM = 1, F1 = 1.0

Ground-truth: Federer Prediction: Roger Federer EM = 0, F1 = 0.66

Final Scoring

QAEval-EM = $(1 + 0) / 2 = 0.5$

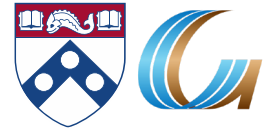
QAEval-F1 = $(1.0 + 0.66) / 2 = 0.83$

Candidate Summary

Roger Federer beat Nadal yesterday

EM = Exact match
F1 is ROUGE-1 F1

(1) Answer Selection



- A good answer selection method will end up generating questions that cover a large portion of the reference summary's information

Harry Potter is not a normal boy. Raised by his cruel Aunt and Uncle and tormented by his bully of a cousin, Dudley, he has resigned to a life of neglect. On his eleventh birthday, however, a half-giant called Hagrid comes crashing—quite literally—into his life, and announces that Harry is a wizard. Together they journey to London to get school supplies for Harry's first year at Hogwarts School of Witchcraft and Wizardry. On 1st September Harry takes a train from King's Cross station, Platform 9 $\frac{3}{4}$, to Hogwarts school, where he meets Ron Weasley and Hermione Granger. The three are sorted into the same House, Gryffindor, and although Harry and Ron find Hermione bossy and annoying at first, the three soon become best friends.

- Who is not a normal boy? Harry Potter
 - Harry potter is not a normal what? Boy
 - Harry potter is not a normal what? Boy
 - Who was he raised by? His cruel aunt and uncle
 - Who was he tormented by? His bully of a cousin, Dudley
- By comparing QA pairs' information to Pyramid Method SCUs
 - Determined that NP Chunk are the best units to use as answers

Single QA pair alone does not represent a ton of the information in the summary

More questions will cover more semantic information

(2) Question Generation



- Trained a question generation model

Input

<m> Federer </m> beat Nadal yesterday

Output

Who beat Nadal yesterday?

- Empirically, questions are good, but a bit verbose

Input: On Jan. 7, 2005, with inauguration scheduled for Jan. 12, [Rossi] filed a lawsuit seeking a new election.

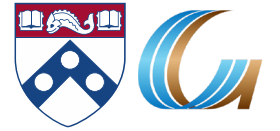
Expert: Who filed a lawsuit seeking a new election?

Model: On Jan. 7, 2005, with inauguration scheduled for Jan. 12, who filed a lawsuit seeking a new election?

- Nevertheless, the downstream performance is better when using a model versus an expert to write questions.

- Possibly since the questions are more verbose so they contain more keywords
- Summary-level is equal
- We conclude that the question generation model gives good performance

(3) Question Answering



- QA model is trained on SQuAD 2.0

Candidate Summary

Harry Potter is not a normal boy. Raised by his cruel Aunt and Uncle, and tormented by his bully of a cousin, Dudley, he has resigned to a life of neglect. On his eleventh birthday, however, a half-giant called Hagrid comes crashing—quite literally—into his life and announces that Harry is a wizard. Together they journey to London to get school supplies for Harry's first year at Hogwarts School of Witchcraft and Wizardry. On 1st September Harry takes a train from King's Cross station, Platform 9 3/4, to Hogwarts school, where he meets Ron Weasley and Hermione...

Question

Where does Harry go to school?

Output

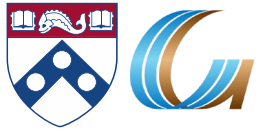
Probability question is not answerable: 0.2

Predicted answer: Hogwarts

Probability of predicted answer: 0.75

- How well does it work on the summaries and questions?
 - We manually answered ~3k QA pairs and verified if the answers were correct or not
 - Questions generated from 20 references over 10 input document sets
 - Answered against 4 systems outputs
 - Key area of concern:
 - Unanswerable questions (or not) [work in submission on that]
 - The fraction of unanswerable questions here is larger than in the training.
 - Even when the answer is there, the model does a poor job at picking the answer

(4) Answer Verification



- In typical QA data, the QA model is executed against the same text where the ground-truth was drawn from
 - Exact match always exists, F1 helps with some span correction
- Our ground-truth answers are selected from a different text than where the predictions come from
 - No guarantee an exact match will exist or be close, but the answer might be there

Summary: The killing of Lebanon's former PM Rafiq Hariri renewed calls for Syria to abide by UN Security Council Resolution 1559 and end its dominance of Lebanon...

Question: What event put Syria under renewed pressure from the international community to abide by UN Security Council Resolution 1559 and withdraw its troops from Lebanon?

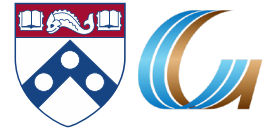
Answer: The February assassination

Prediction: The killing of Lebanon's former PM Rafiq Hariri

Same event referred to in different ways, but EM and F1 near 0

This problem could introduce a lot of noise into the evaluation metric

(5) Overall Evaluation



■ Correlations of fully automatic metric

Metric	TAC 2008 System-Level			TAC 2008 Summary-Level		
	r	ρ	τ	r	ρ	τ
Pyramid Score	.90	.88	.70	.59	.59	.50
ROUGE-1	.79	.80	.60	.49	.48	.39
ROUGE-2	.83	.87	.67	.48	.48	.39
PyrEval	.81	.79	.59	.31	.31	.25
MoverScore	.83	.82	.61	.50	.49	.40
QAEval-EM	.93	.91	.76	.33	.33	.27
QAEval-F ₁	.90	.88	.71	.46	.45	.36

Metric	TAC 2009 System-Level			TAC 2009 Summary-Level		
	r	ρ	τ	r	ρ	τ
Pyramid Score	.90	.87	.70	.59	.57	.48
ROUGE-1	.83	.78	.60	.54	.47	.38
ROUGE-2	.76	.84	.67	.50	.50	.40
PyrEval	.86	.82	.64	.39	.35	.28
MoverScore	.82	.80	.63	.51	.52	.42
QAEval-EM	.70	.87	.69	.42	.38	.30
QAEval-F ₁	.81	.89	.72	.50	.45	.36

- The results are even better on CNN/Dailymail
- QAEval gets best system-level correlation, > all other metrics
 - Consistently better than the Pyramid Method
 - Quite surprising because of the noisy QA and answer verification
- But, summary-level results are lower or competitive
 - Averaging over a small # of answers isn't enough to deal with the noise introduced by the QA components.

- The QA model effectively does the job of the human who marks common SCUs between summaries
- QA pairs represent more information than SCUs
 - Humans can't annotate the data as exhaustively => QA can have better coverage

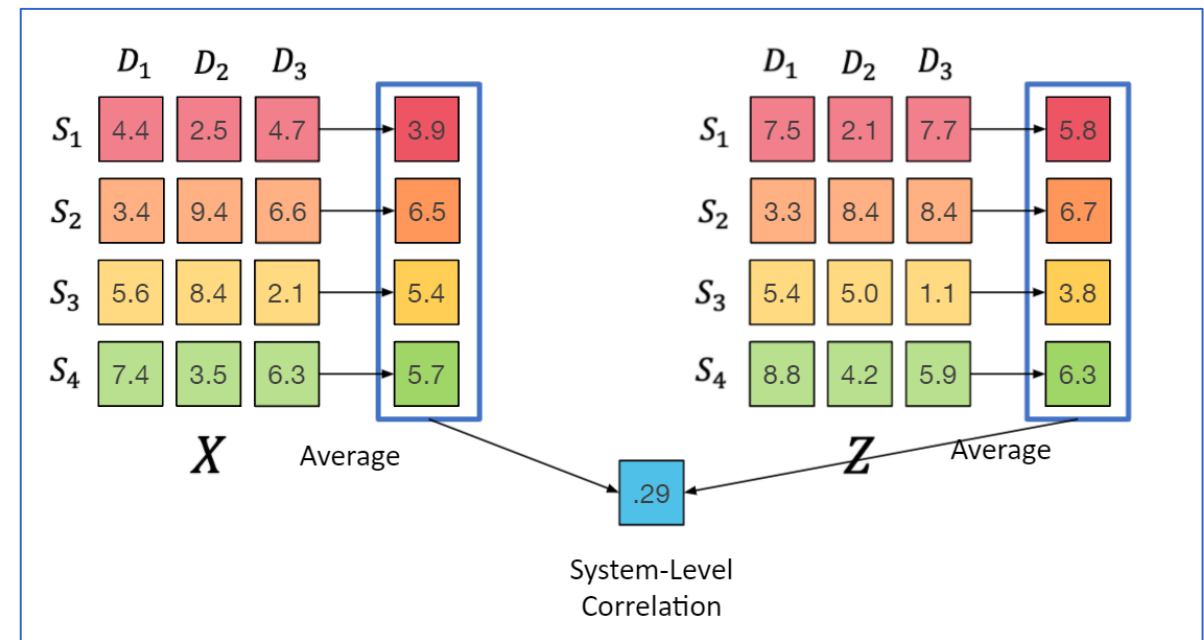
- Automatic metrics are evaluated by considering their correlation to human judgments of summary quality.
 - We should care about system-level correlation (not summary level) calculated with Kendall's tau.
 - Since we compare systems (not individual summaries).

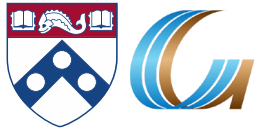
- Ranking Comparison:

- How frequently do ROUGE and Humans agree that: **system1** is better than **system2**?

- Do this for multiple metrics, and see which correlates better with humans

- For **N systems**, Kendall's tau enumerates the (N choose 2) possible pairs of systems then calculates the proportion of pairs for which ROUGE and humans agree on the ranking.

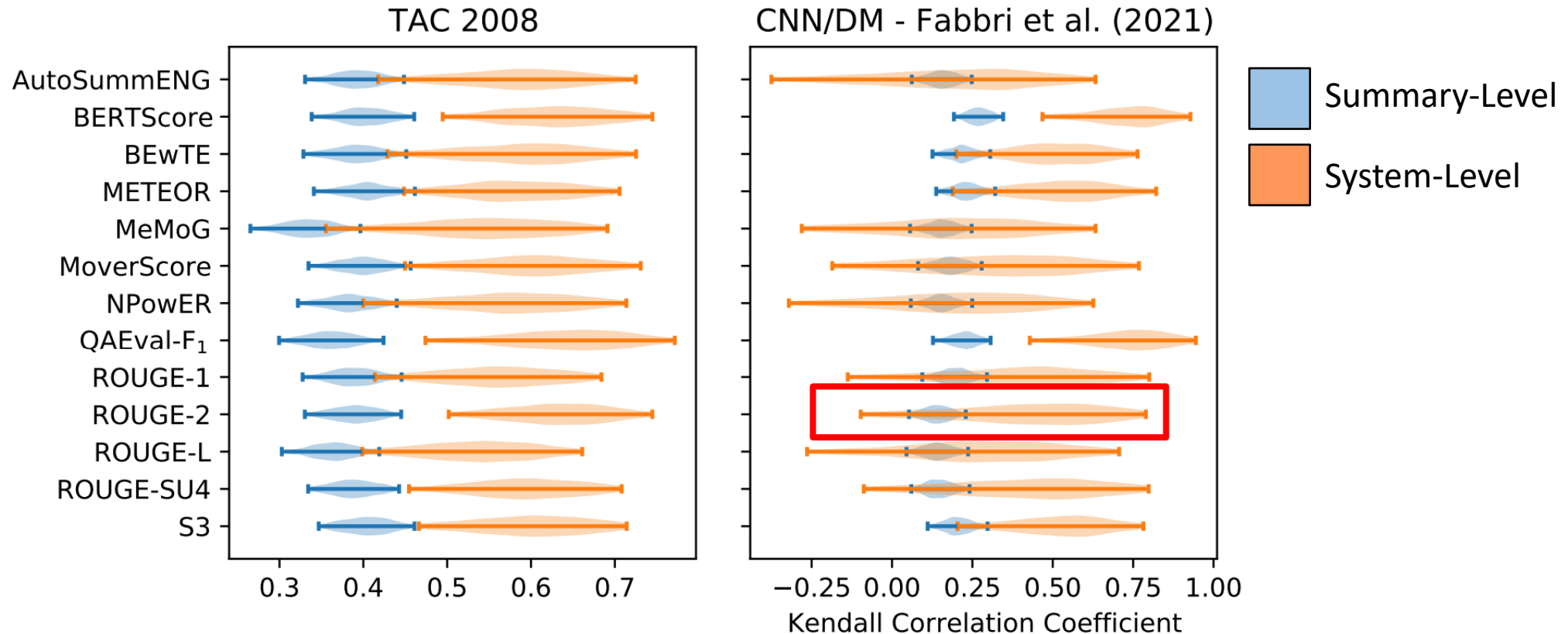




- When scoring the quality of a System – its correlation with human performance – we need to attend to the confidence of these scores.
- But, new evaluation metrics rarely come with **confidence intervals** or any other statistical test that shows the significance of the correlation improvements.
- We propose bootstrapping/permutation tests to do exactly that.
- When done the right way, the confidence intervals are shown to be very wide.

We have a very poor idea how well automatic metrics agree with human judgments of summary quality

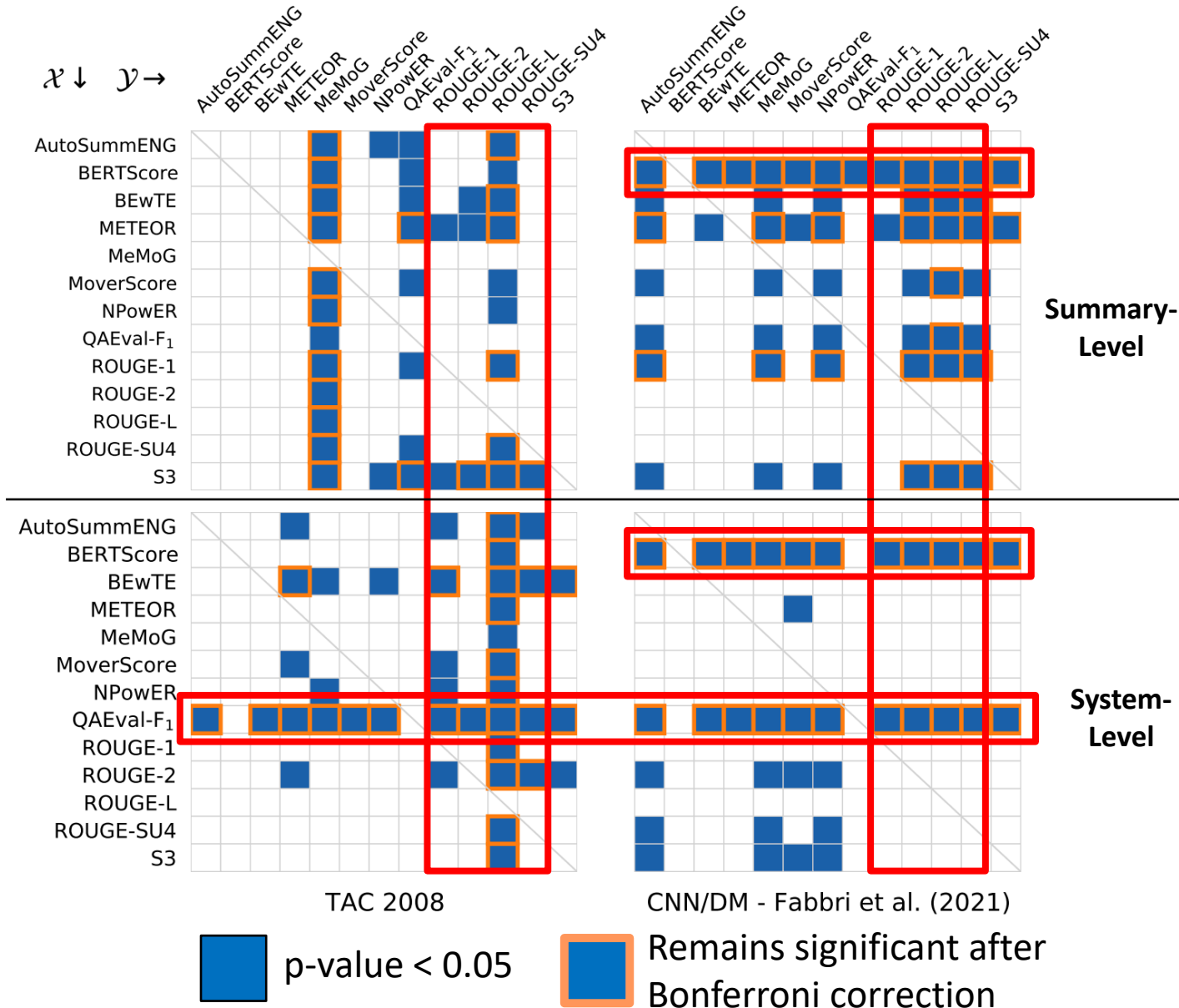
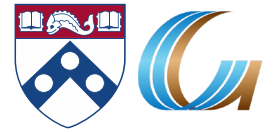
Confidence Intervals



- Confidence intervals are wide
 - High-level of uncertainty about their true values

- Wide CIs have serious implications
 - ROUGE-2 incorrectly ranks systems 9-54% of the time with respect to human rankings

Hypothesis test results



- Orange in rows = row metric is good
- Orange in columns = column metric is bad
- BERTScore (Zhang et al., 2020) does well on SummEval
- QAEval (Deutsch et al., 2021) does well at the system-level
- Many metrics don't do better than ROUGE

Evaluation of Commonsense Acquisition/Reasoning Systems

Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema

Yanai Elazar, Hongming Zhang, Yoav Goldberg, Dan Roth,
EMNLP-2021

“Going on a vacation” takes longer than “Going for a walk”: A Study of Temporal Commonsense Understanding

Ben Zhou, Daniel Khashabi, Qiang Ning, Dan Roth,
EMNLP-2019

Have we Solved Common Sense?



- Some of the evaluations we do today to commonsense task show performance that is on-par with human performance.
- Are we done?

- We are not;
- These are artifacts of how we evaluate
 - Winograd Schemas
 - Temporal Commonsense
 - Multiple choice QA

 - Introduced in 2011 as an alternative to the Turing Test by Hector J. Levesque
- The purpose is to test for common sense
- “ Moreover, the test is arranged in such a way that having full access to a large

The Winograd Schema

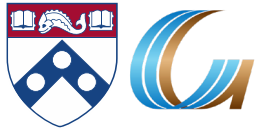


- Every question involves:

Joan made sure to thank Susan for all the help she had given.

- Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;

The Winograd Schema

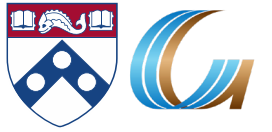


- Every question involves:

***Joan** made sure to thank **Susan** for all the help she had given.*

- Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;

The Winograd Schema



- Every question involves:

Joan made sure to thank Susan for all the help she had given.

- Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
- A pronoun is used in the example to refer to one of the entities

The Winograd Schema

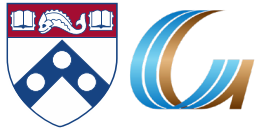


- Every question involves:

*Joan made sure to thank Susan for all the help **she** had given.*

- Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
- A pronoun is used in the example to refer to one of the entities

The Winograd Schema

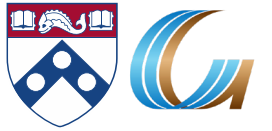


- Every question involves:

Joan made sure to thank Susan for all the help she had given.

- Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
- A pronoun is used in the example to refer to one of the entities
- The task is to determine which of the two entities is referred to by the pronoun (coreference)

The Winograd Schema



- Every question involves:

*Joan made sure to thank **Susan** for all the help **she** had given.*

- Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
- A pronoun is used in the example to refer to one of the entities
- The task is to determine which of the two entities is referred to by the pronoun (coreference)

The Winograd Schema

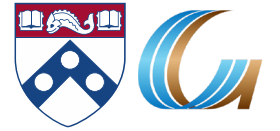


- Every question involves:

Joan made sure to thank Susan for all the help she had given.

- Two entities are mentioned in each sentence, and they can be two males, two females, two inanimate objects, or two groups of people or objects;
- A pronoun is used in the example to refer to one of the entities
- The task is to determine which of the two entities is referred to by the pronoun (coreference)
- Each sentence contains a special word which, when replaced, the answer changes.

The Winograd Schema



- Every question involves:

Joan made sure to thank Susan for all the help she had given.

- **Two entities** are mentioned in each sentence

- two males, two females, two inanimate objects, or two groups of people or objects;

- A **pronoun** is used in the example to refer to one of the entities

- The task: coreference: which of the two entities is referred to by the pronoun

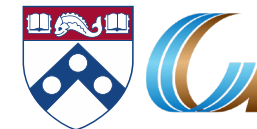
- Each sentence contains a **special word** which, when replaced, the answer changes.

Joan made sure to thank Susan for all the help she had received.

The trophy doesn't fit in the brown suitcase because it was too large.

The trophy doesn't fit in the brown suitcase because it was too small.

The Winograd Schema



- Initial dataset of 273 examples et al., 2012


- Written by experts

- 2 years ago: Winogrande with 44K examples et al., 2019

- Written by crowd workers

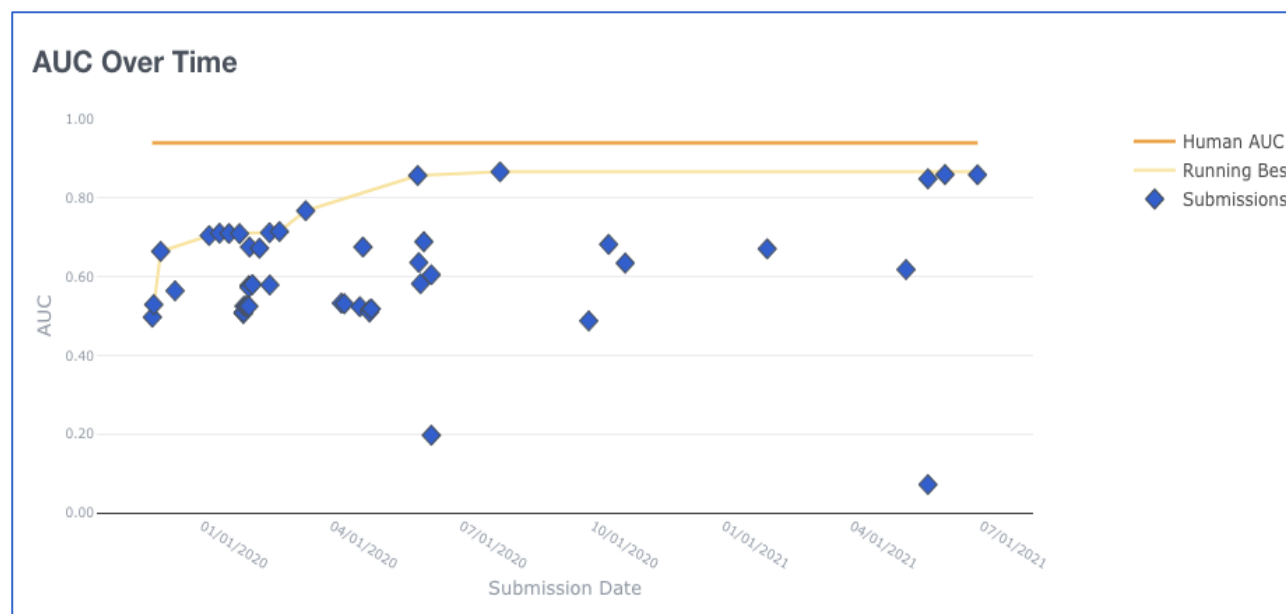
- Today:

Three reasons the results are inflated:
[Elazar et al. EMNLP'21]

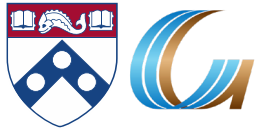
1. Artifacts in the data (bias)
2. Evaluation 
3. Limited Generalization

--Levesque

--Sakaguchi



Standard Evaluation



- We get a set of inputs, and report accuracy

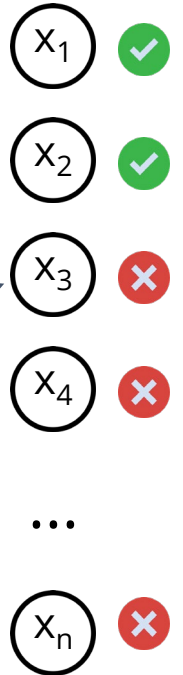
- this is fine, when the data is sampled i.i.d

- But this is not the case in the winograd schema!

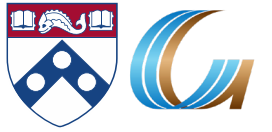
- Recall the pairs:

- The trophy doesn't fit into the brown suitcase because it is too large
- The trophy doesn't fit into the brown suitcase because it is too small

- If a model got only one item of a pair right, did it really understand the question?



Paired Evaluation

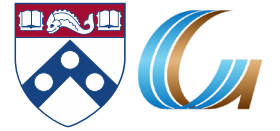


- Instead, let's assign a point to a pair
 - only if a model gets both right
- This way, the risk of giving away points is reduced...
 - and this evaluation becomes more robust and meaningful
- The results on Winogrande go down 71.49 → 58.45
- The paired setting can be generalized to larger groups:

p_1		p_2	
x_1	✓	x'_1	✓
x_2	✓	x'_2	✗
x_3	✗	x'_3	✓
...		...	
x_m	✗	x'_m	✗

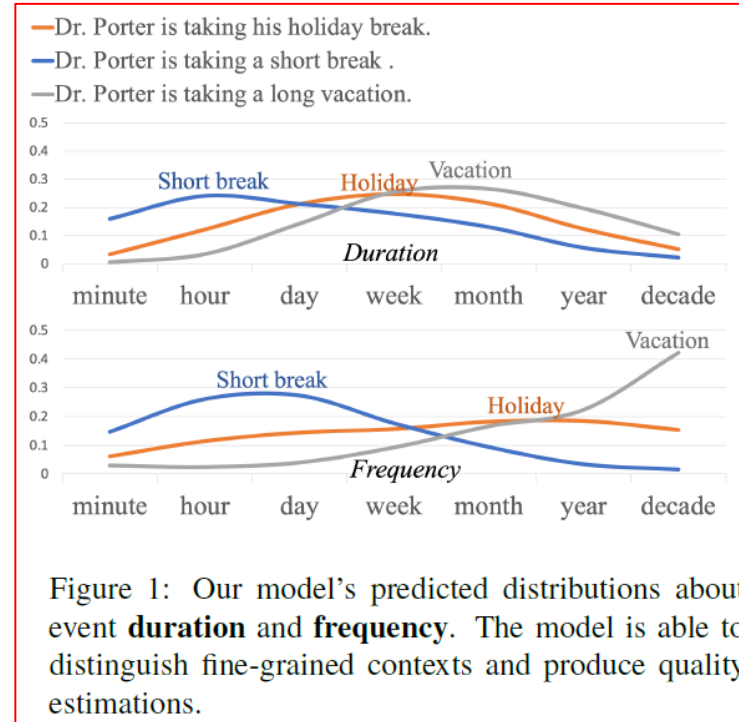
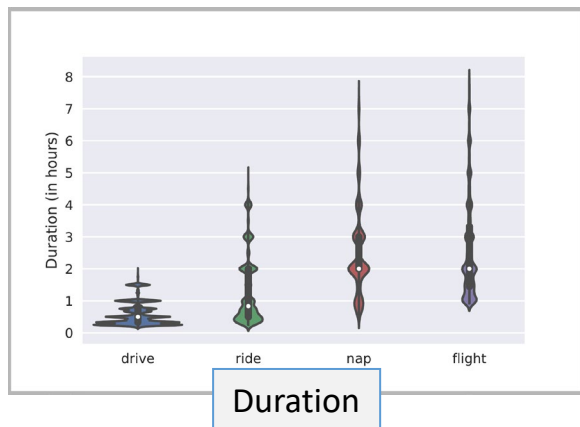
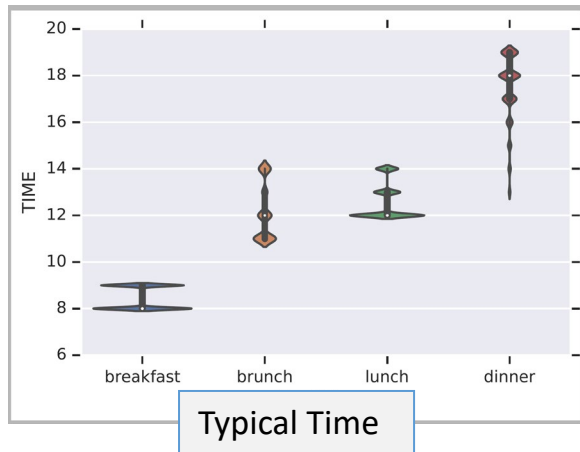
$$groupScore(x_i) = \min_j f(x_{ij})$$

Temporal Common Sense



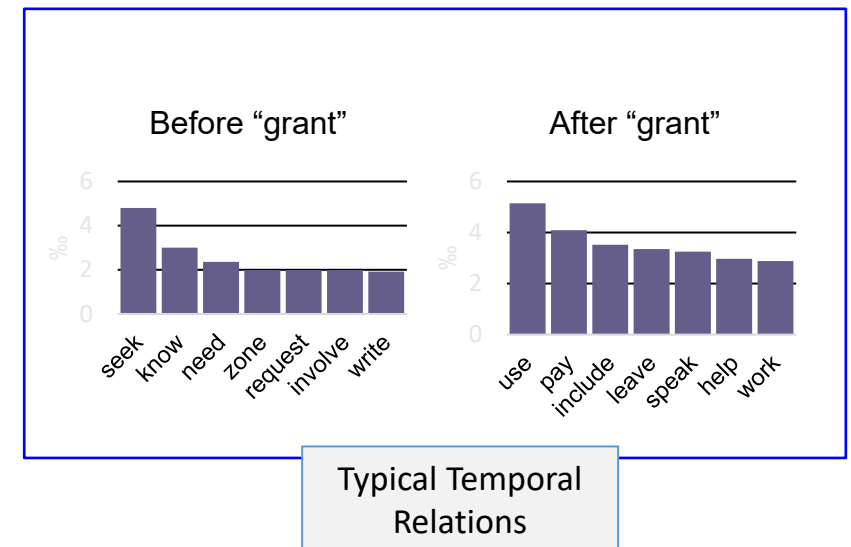
■ Two efforts:

- A dataset MC-TACO [Zhou et al. EMNLP'19] ←
- Acquisition + Representation [Zhou et al. ACL'20]: Duration, typical time, frequency.



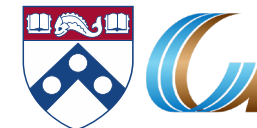
[Elazar et al. ACL'19]

[Zhou et al. ACL'20]



Ning et al. NAACL'18

Defining the Temporal Commonsense Challenge



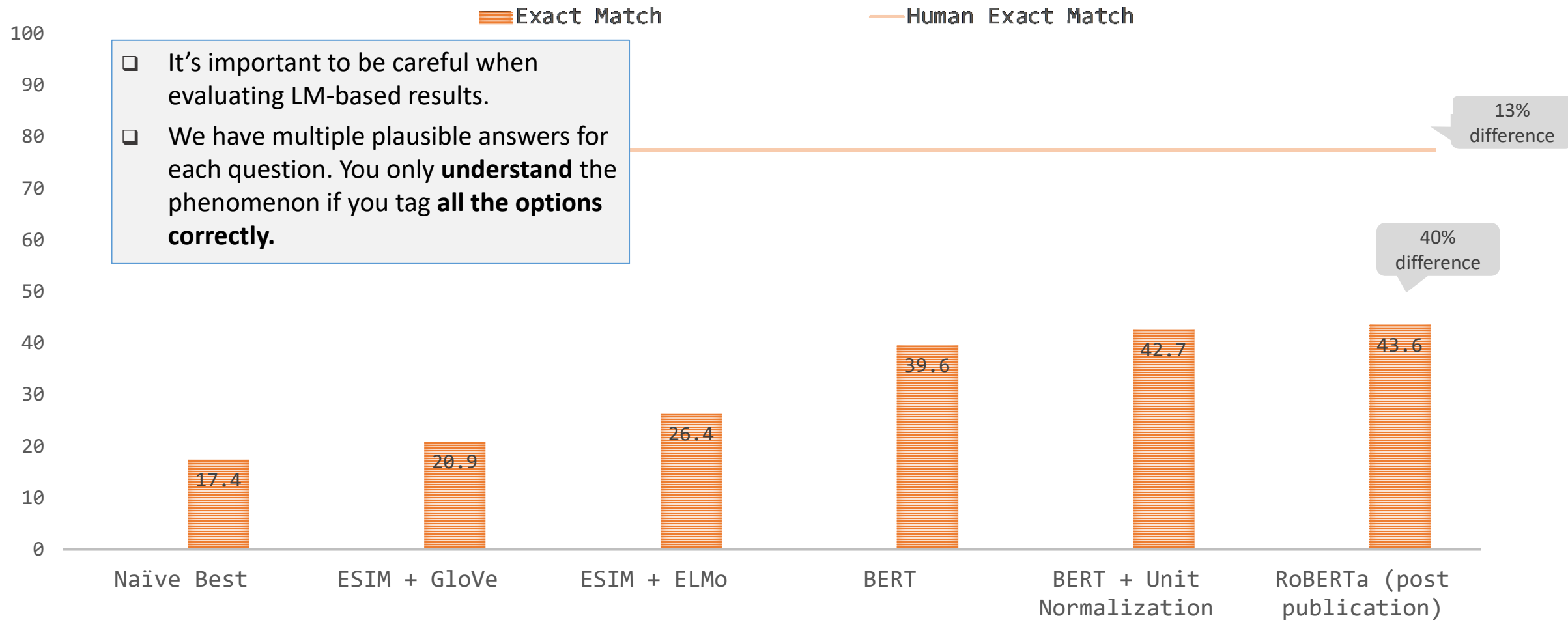
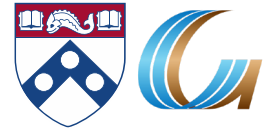
■ MC-TACO [Zhou et al. EMNLP 2019]

- **M**ultiple **C**hoice **T**empor**A**I **C**ommon-sense
- 1,893 questions; 13,225 question-answer pairs
- Querying at least one of the five dimensions:
 - Duration
 - Frequency
 - Typical Occurring Time
 - Stationarity
 - Ordering

			Gold	Prediction	
He went to Duke University.	How long did it take him to graduate?	4 years	■	■	✓
		10 days	■	■	✓
		3.5 years	■	■	✗
		16 hours	■	■	✓

- **Exact Match**: the percentage of questions of which **all** candidates are predicted correctly (here: 0.0)
- **F1**: Gives partial credit (credits “accidental” correct perditions (here: 66.7%))

Results: We are Far (from where we want to be)



It's important to be careful when evaluating LM-based results.

We have multiple plausible answers for each question. You only **understand** the phenomenon if you tag **all the options correctly**.

ESIM: Enhanced LSTM for Natural Language Inference (Chen et al., 2016)
GloVe: Global Vectors for Word Representation (Pennington et al., 2014)
ELMo: Deep contextualized word representations (Peters et al., 2018)
BERT: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2019)
RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)



- Evaluation is important, tricky, and we are not so good at it.
 - Getting results from our evaluation metrics does not mean that we know how good we are

- True even for seemingly simple tasks such as Grammatical Error Correction

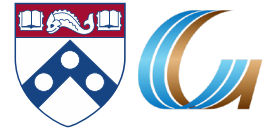
- Summarization:
 - We proposed QAEval, a QA-based metric for evaluating the content quality of summaries
 - QA as an evaluation methodology is fundamentally different and better than text overlap methods
 - Demonstrated state-of-the-art system-level performance
 - Identified **QA** and **answer verification** as bottlenecks

- Common Sense Reasoning
 - Evaluations that are too relaxed credit accidentally correct predictions
 - Provide misleading evaluation of where we are

END

- FEQA (Durmus et al., 2020) and
- QAGS (Wang et al., 2020)
 - Focus on faithfulness: is the information consistent with the input
 - Compare summaries to the input documents; we compare summaries to references.
- APES (Eyal et al. 2019), is the most relevant.
 - Create fill-in-the-blank questions by removing named entities from the reference summary and use a reading comprehension model to predict which entity was removed using the candidate summary.
 - We argue that QA as an evaluation methodology is fundamentally different and better than text overlap methods,
 - Our proposed metric QAEval is more widely applicable than APES because QAEval asks and answers questions about noun phrases; APES is restricted to named entities.
 - Our evaluation of QAEval is more comprehensive:

QA Metric Comparison



	This Work	Eyal et al., 2019	Wang et al., 2020	Durmus et al., 2020
Question Source	Reference Summary	Reference Summary	Candidate Summary	Candidate Summary
Answer Type	NP Chunks	NER	10 NER + NP	NER + NP
Question Type	Wh	Fill-in-the-blank	Wh	Wh
QA Model	SQuAD	CNN/DailyMail	SQuAD	SQuAD
Prediction Source	Candidate Summary	Candidate Summary	Input Document	Input Document
Answer Verification	EM/F1	EM	F1	F1