Rethinking NLG Evaluation: From Metric Learning, Sample Variance, to Fact-Checking

William Wang

Mellichamp Chair in AI and Designs

Assistant Professor of Computer Science

UC SANTA BARBARA

Evaluation for Natural Language Generation

- It is one of the most challenging evaluation problems in NLP.
- Lexical overlap based evaluation is still dominating, and it was inappropriately used in many tasks.
- Diverse applications of NLG:
  - Data-to-text generation
  - Visually-grounded generation
  - Summarization
  - Dialogue
  - Machine Translation

Challenges of NLG Evaluation

- Lexical overlap based approaches (BLEU, ROUGE etc) do not always capture semantic matching.
- They are not suitable for long text generation and complex problems.
- Variance-free result reporting also amplifies issues.
- They fail on adversarial examples and factual generation problems on faithfulness.

This Talk

- **Metric Learning**: We show that modeling the **complex objective** and **diverse** answers is the key to better visually-grounded generation models.
- **Sample Variance**: We conduct a comprehensive evaluation on the sample variance for diverse problems in language-and-vision generation.
- **LogicNLG Evaluation**: We discuss semantic parsing and adversarial example based evaluations for faithful table-to-text NLG.

Outline

- Motivation
- Inverse RL for Visual Story Telling
- Understanding Sample Variance
- LogicNLG Evaluation
- Conclusion

No Metrics are Perfect:

From Optimizing End Metrics (e.g., BLEU/ROUGE) to Reward Learning

(Wang, Chen et al., ACL 2018)

Existing Automatic Evaluation Metrics for Language Generation

- Input: generation candidate and human reference(s).
- Output: a score.

- Metrics:
  - BLEU: precision-driven n-gram overlap.
  - ROUGE: recall-driven n-gram overlap.
  - METEOR: weighted f1 n-gram overlap.
  - CIDEr: TF-IDF + cosine similarity.

Pop Quiz: assuming reasonable references, what is the METEOR score of this sample output?

"We had a great time to have a lot of the. They were to be a of the. They were to be in the. The and it were to be the. The, and it were to be the."

Average METEOR score: 40.2

(SOTA model: 35.0)

How about this one?



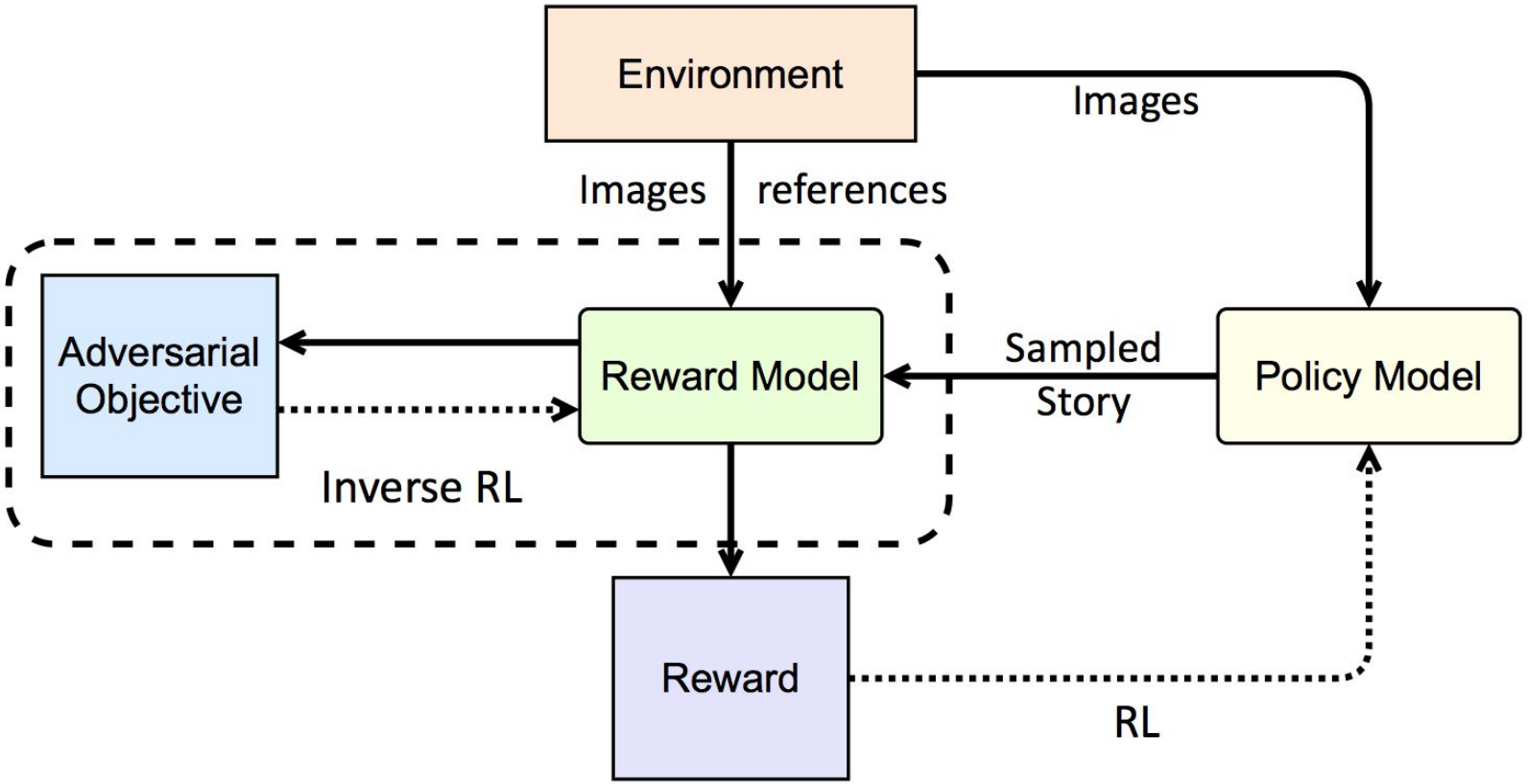
"I had a great time at the restaurant today. The food was delicious. I had a lot of food. I had a great time."

BLEU-4 score: 0

No Metrics Are Perfect: Adversarial Reward Learning (ACL 2018)

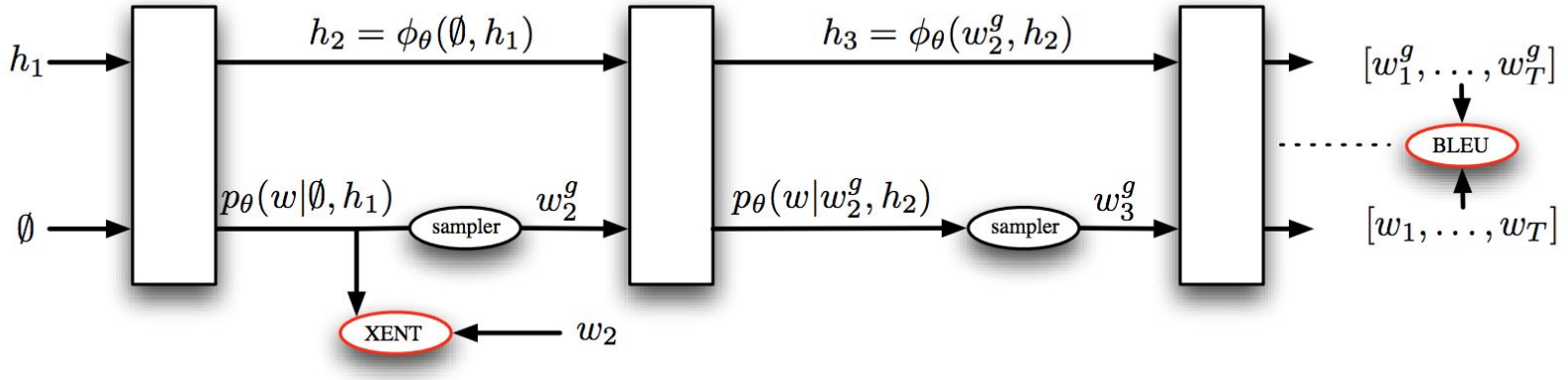
- Task: visual storytelling (generate a story from a sequence of images in a photo album).
- Difficulty: how to quantify a good story?
- Idea: given a policy, learn the reward function.

No Metrics Are Perfect: Adversarial Reward Learning (Wang, Chen et al., ACL 2018)



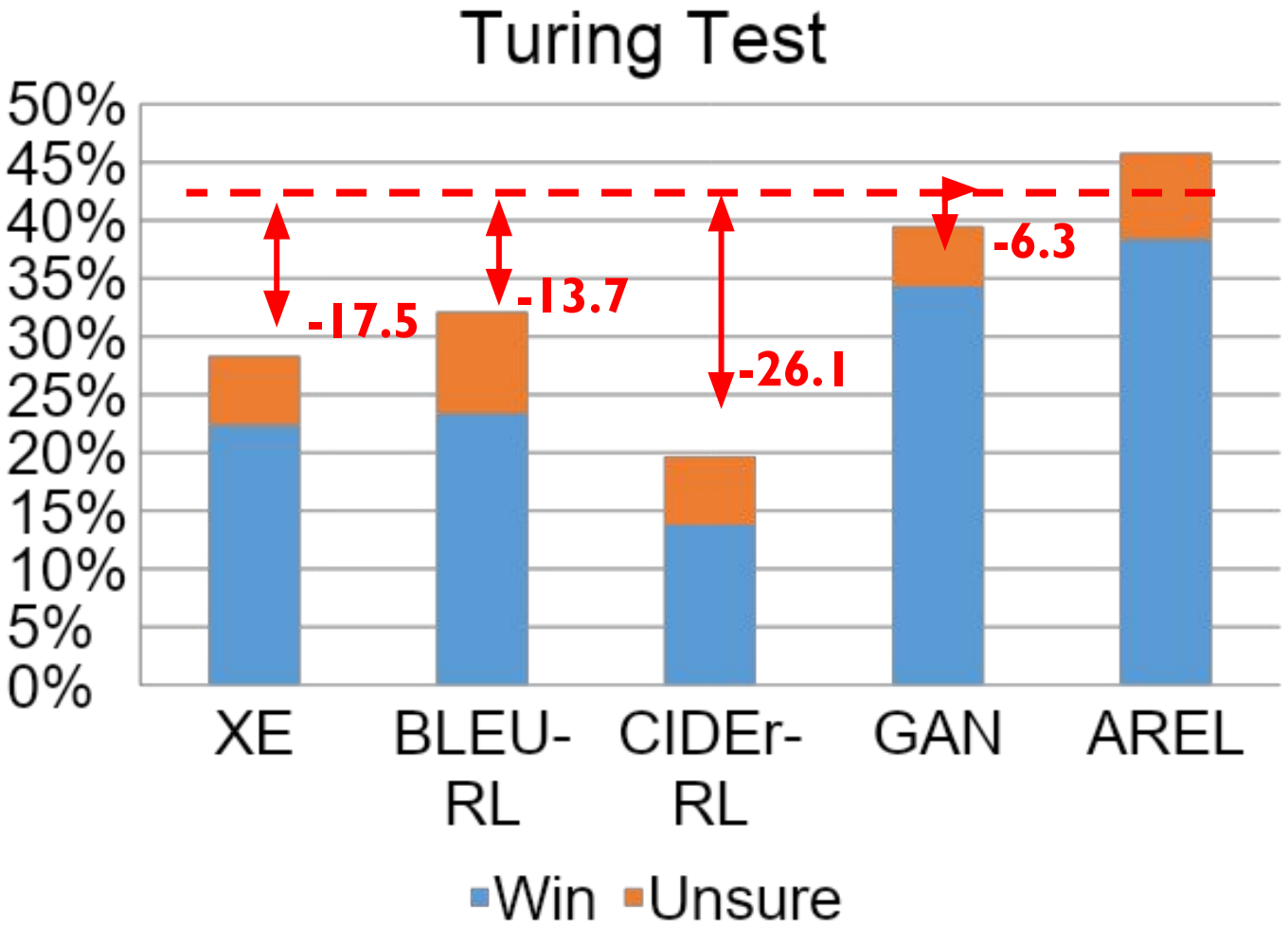
Baseline: MIXER (Ranzato et al., ICLR 2016)

- Optimize the cross-entropy loss and the BLEU score directly using REINFORCE (Williams, 1992).



AREL Storytelling Evaluation

- Dataset: VIST (Huang et al., 2016).



When will IRL work?

- When the optimization target is complex.
- There are no easy formulations of the reward.
- If you can clearly define the reward, don't use IRL and it will not work.

Outline

- Motivation
- Inverse RL for Visual Story Telling
- Understanding Sample Variance
- LogicNLG Evaluation
- Conclusion
- Other Research Interests and Goals

Sample Variance:

Towards Understanding Sample Variance in Visually Grounded Language Generation

(Zhu et al., EMNLP 2020)

Visually Grounded Language Generation

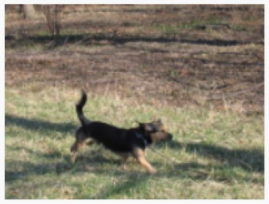
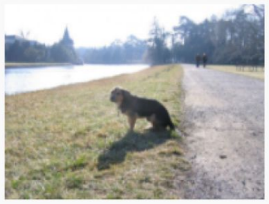
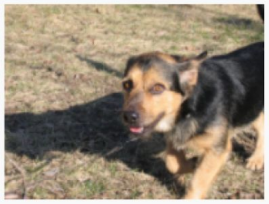
Image Captioning

A brown and white dog plays with blue ball in blue water-filled shell .

Video Captioning

A young man wearing a helmet riding a skateboard down the street in a neighborhood.

Visual Storytelling

1	2	3	4	5
				
The dog was ready to go.	He had a great time on the hike.	And was very happy to be in the field.	His mom was so proud of him.	It was a beautiful day for him.

Sample Variance

1. This group of folks comprising runners and bikers, some wearing identifying numbers, look like they are getting ready for a marathon.
2. A runner in yellow has a convoy of motorcycles following behind him on a highway as bystanders watch.
3. Marathon runners are running down a street with motorcyclists nearby.
4. A runner in the middle of a race running along side the road.
5. A man in a yellow shirt is running in a race.

Dataset & Metric

- 3 visually grounded language generation tasks
- 7 datasets
- 6 automatic metrics
  - BLEU
  - ROUGE
  - METEOR
  - CIDEr
  - SPICE
  - BERTScore

Task	Dataset	#Reference
Image Captioning	Flickr8k	5
	Flickr30k	5
	COCO	5
	PASCAL-50	50
Video Captioning	VATEX (English)	10
	VATEX (Chinese)	10
Visual Storytelling	VIST	5

Reference Variance within Datasets

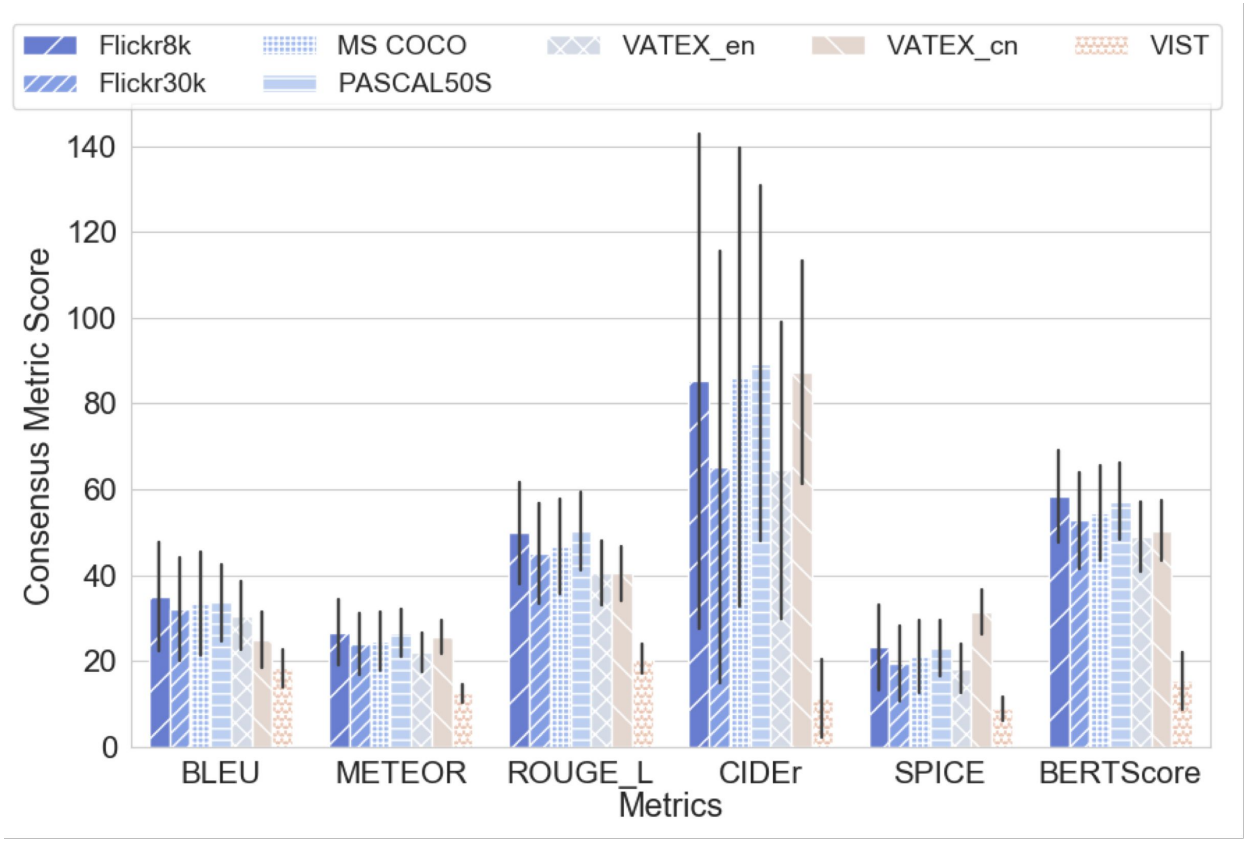
Experimental Setup

- Reference group $R = \{r_i\}_{i=1}^{n}$

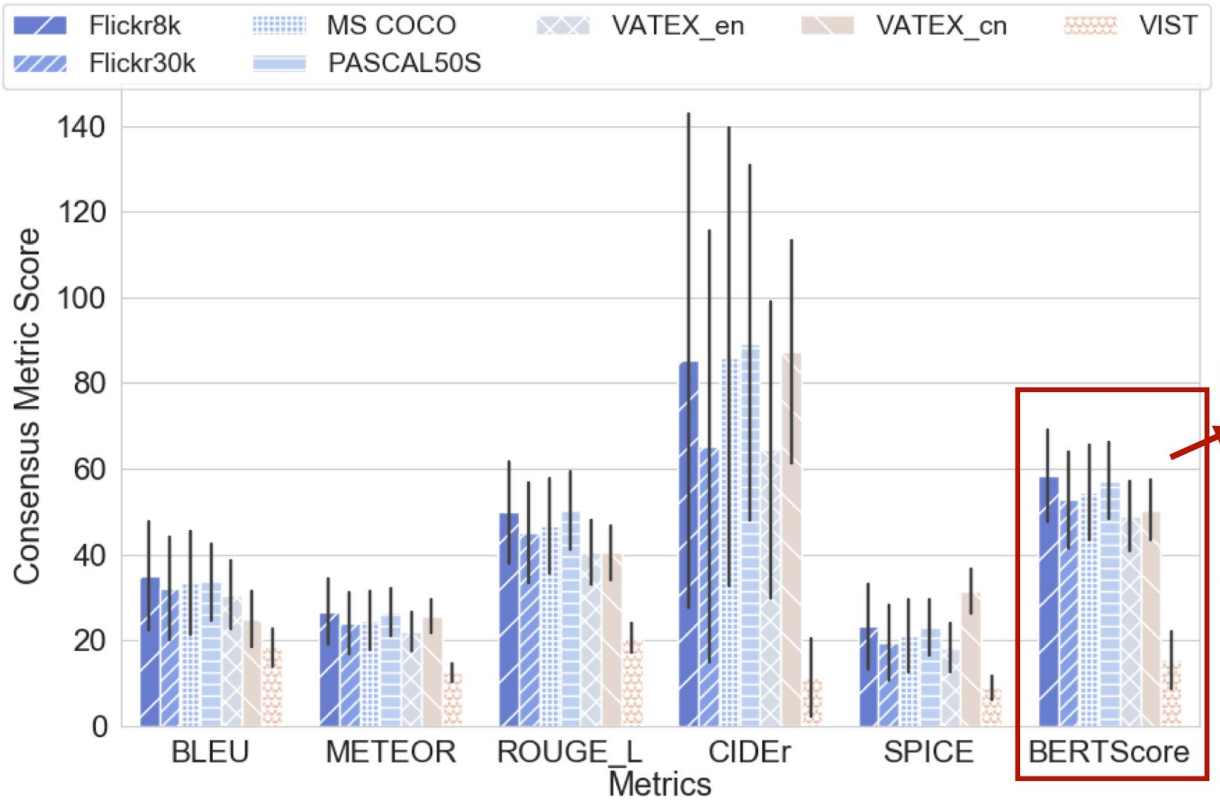
- Consensus metric score

$$c = \frac{1}{n} \sum_{i=1}^{n} metric(r_i, R \backslash \{r_i\})$$

Consensus Metric Score

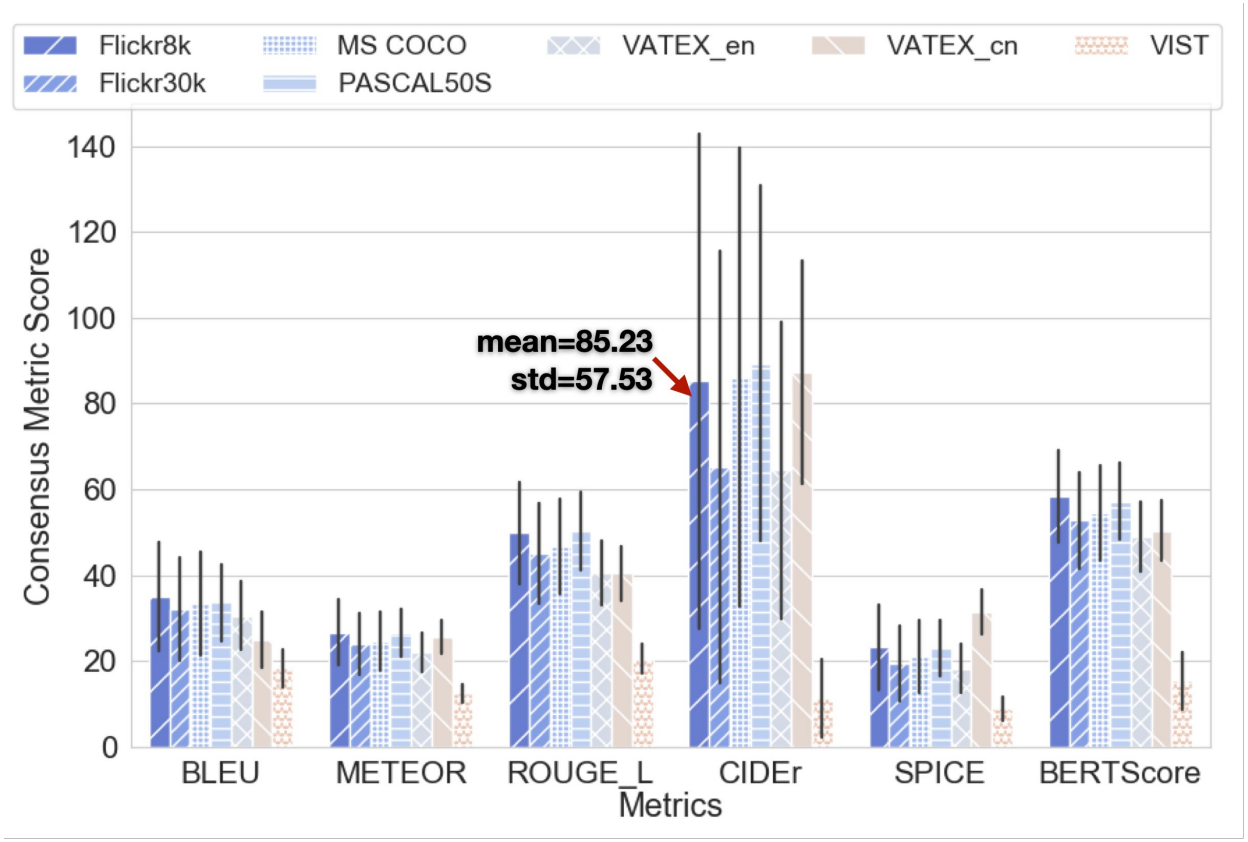


BERTScore order coincides with task difficulties.



Task	Dataset	BERTScore
Image Captioning	Flickr8k	58.40 ± 10.76
	Flickr30k	52.77 ± 11.14
	COCO	54.40 ± 10.98
	PASCAL-50	57.26 ± 9.00
Video Captioning	VATEX (English)	48.99 ± 8.06
	VATEX (Chinese)	50.40 ± 7.05
Visual Storytelling	VIST	15.46 ± 6.58

CIDEr has the largest std on consensus scores

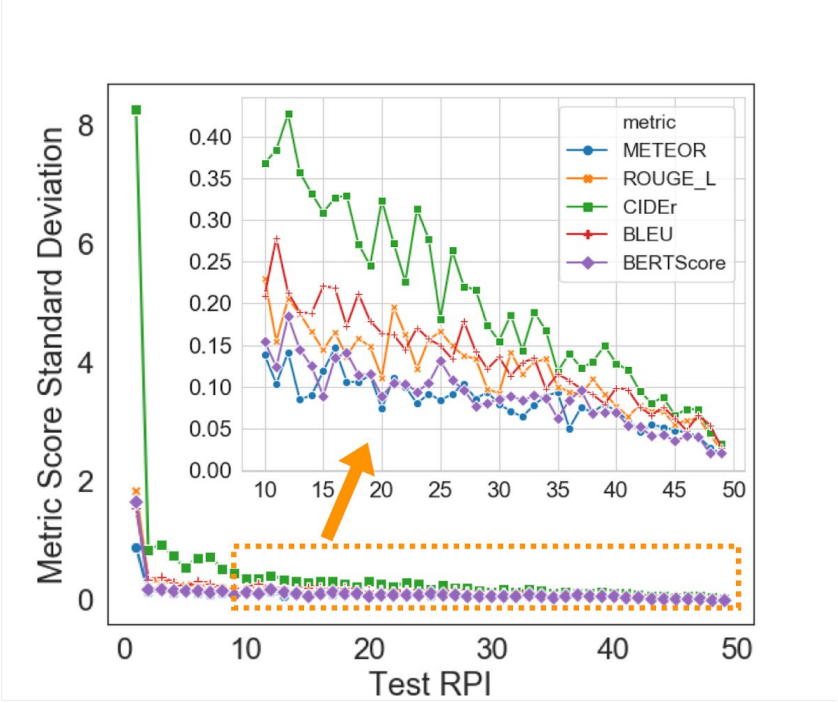


Effect of Sample Variance on Evaluation Performance

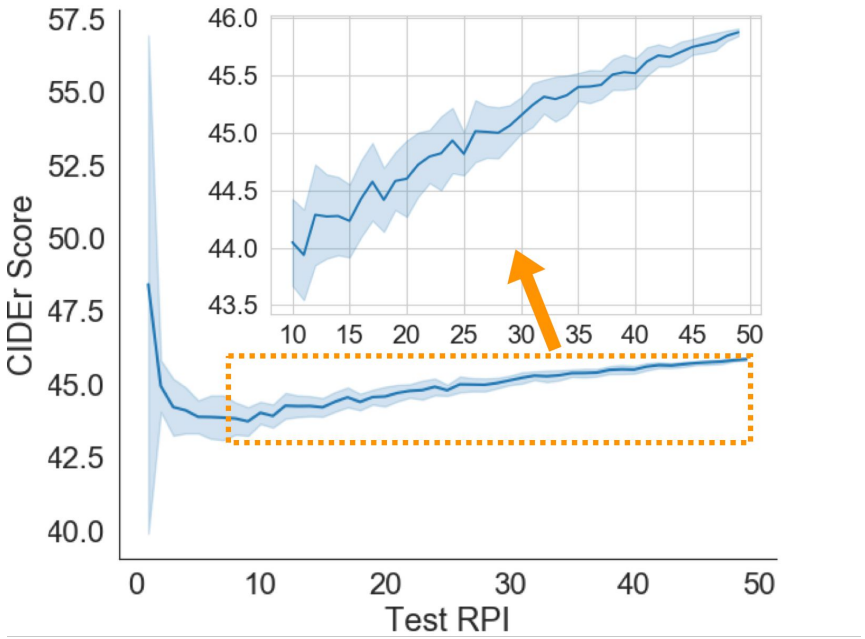
Experimental Setup

- The number of parallel References Per visual Instance → RPI
- The model is trained on the complete training set
- Incrementally set the **testing RPI** as 1, 2, ..., n − 1

Evaluation score deviation is salient with only using a few parallel testing reference



Score deviation on PASCAL50S



CIDEr score on PASCAL50S

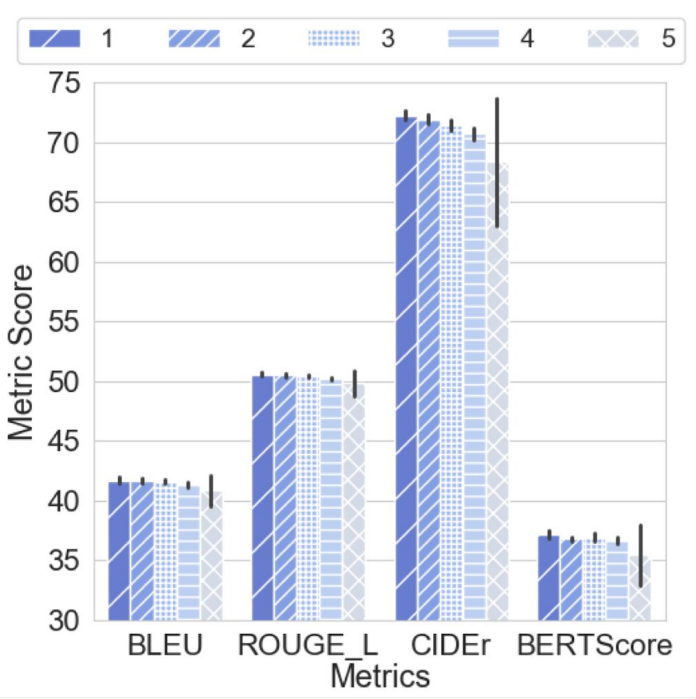
More visuals or more references?

Experimental Setup

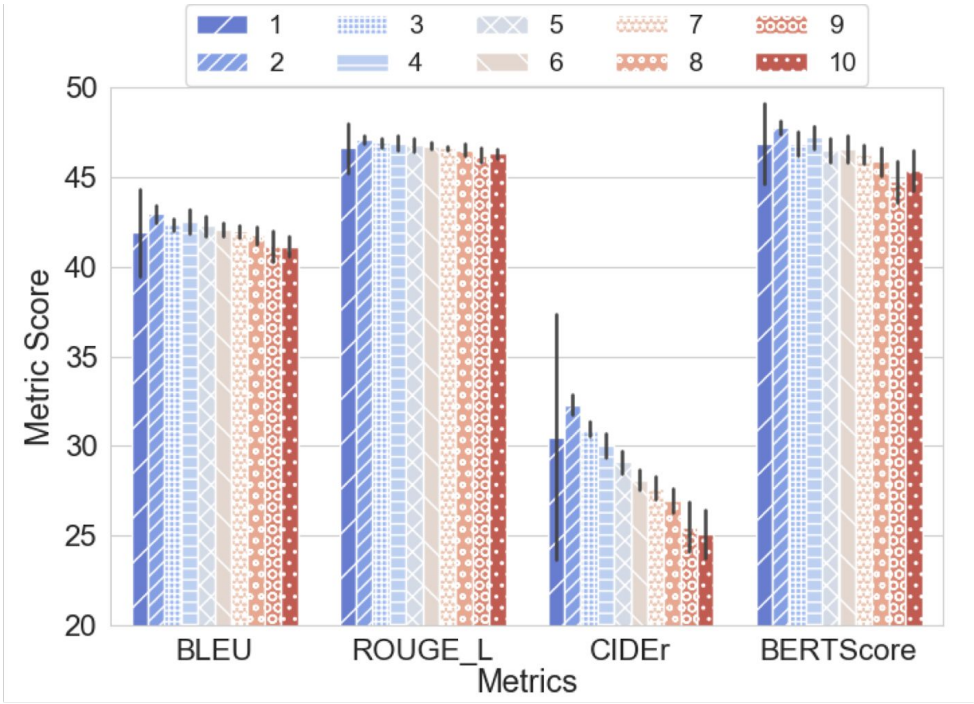
- Fix the total number of training data samples
- set the t**raining RPI** as 1, 2, ..., n − 1

$$\#sample = \#visual\_instance * RPI$$

Introducing more visual instances during training is beneficial for the captioning tasks

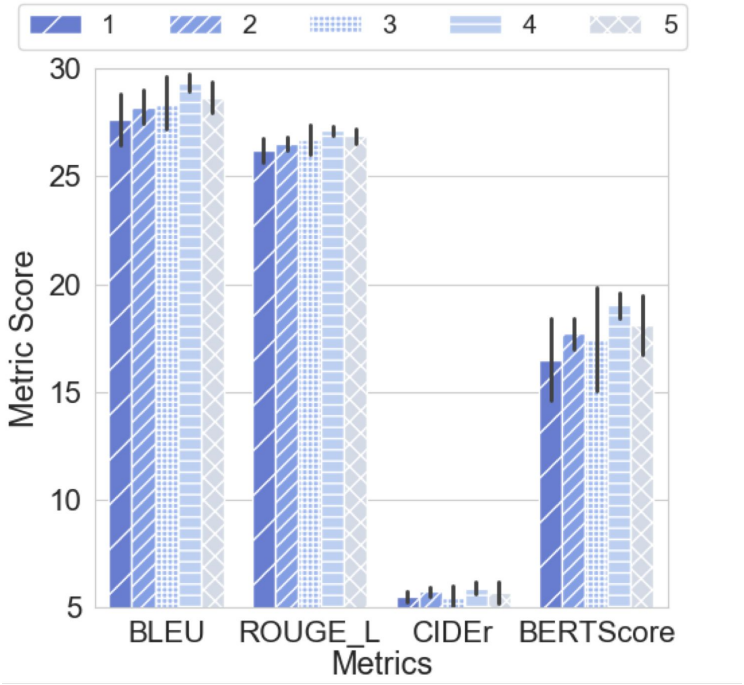


MS COCO



VATEX (English)

More parallel text references can help train a more stable and better-performing storytelling model



VIST

Practical Takeaways

- **Please report metric score variance in addition to the metric scores when comparing models' performance**

Practical Takeaways

- Please report sample variance in addition to the metric scores when comparing models' performance
- **When collecting a new dataset …**
  - the **test set** should include more parallel references for fair evaluation

Practical Takeaways

- Please report sample variance in addition to the metric scores when comparing models' performance
- **When collecting a new dataset …**
  - the test set should include more parallel references for fair evaluation
  - the **training set** should collect …
    - more parallel references if the text generations are expected to be distinctive and complicated
    - otherwise, a larger variety of visual appearances is more favorable

Outline

- Motivation
- Inverse RL for Visual Story Telling
- Understanding Sample Variance
- LogicNLG Evaluation
- Conclusion
- Other Research Interests and Goals

Logical Natural Language Generation

- The existing NLG paradigm
  - Straightforwardly convert the data into surface form.
  - No logical reasoning or inference is involved in generation.

- We propose the Logical Natural Language Generation
  - Generate content not explicitly represented in the data.
  - The content is derived from logical inference.

LogicNLG:

Logical Natural Language Generation from Open-Domain Tables

(Chen et al., ACL 2020)

Logical Natural Language Generation

- Example: "1 more gold" is not explicitly encoded in the table

Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Logical Natural Language Generation

- Example: "1 more gold" is not explicitly encoded in the table

Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Surface-level Generation
Sentence: Canada has got 3 gold medals in the tournament. **Sentence:** Mexico got 3 silver medals and 1 bronze medal.

Logical Natural Language Generation

- Example: "1 more gold" is not explicitly encoded in the table

Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

Surface-level Generation
Sentence: Canada has got 3 gold medals in the tournament. **Sentence:** Mexico got 3 silver medals and 1 bronze medal.

Logical Natural Language Generation
Sentence: Canada obtained 1 more gold medal than Mexico. **Sentence**: Canada obtained the most gold medals in the game.

Logical Natural Language Generation

- Collect a table-to-text dataset for this specific problem
  - Open-domain Tables
  - Rich Inference over Numbers/Dates/Text.

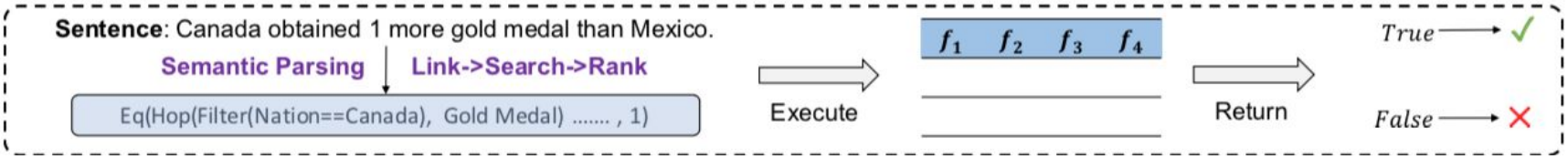
	Vocab	Examples	Vocab/Sent	Tables	Domain	Source	Inference	Schema
WEATHERGOV	394	22.1K	0.01	22.1K	Weather	Crawled	No	Known
WikiBIO	400K	728K	0.54	728K	Biography	Crawled	No	Limited
ROTOWIRE	11.3K	4.9K	0.72	4.9K	NBA	Annotated	Few	Known
LOGICNLG	122K	37.0K	**3.31**	7.3K	**Open**	Annotated	**Rich**	**Unlimited**

Logical Natural Language Generation

- Weakness of Fidelity Metric
  - IE-based evaluation mechanism miserably fails

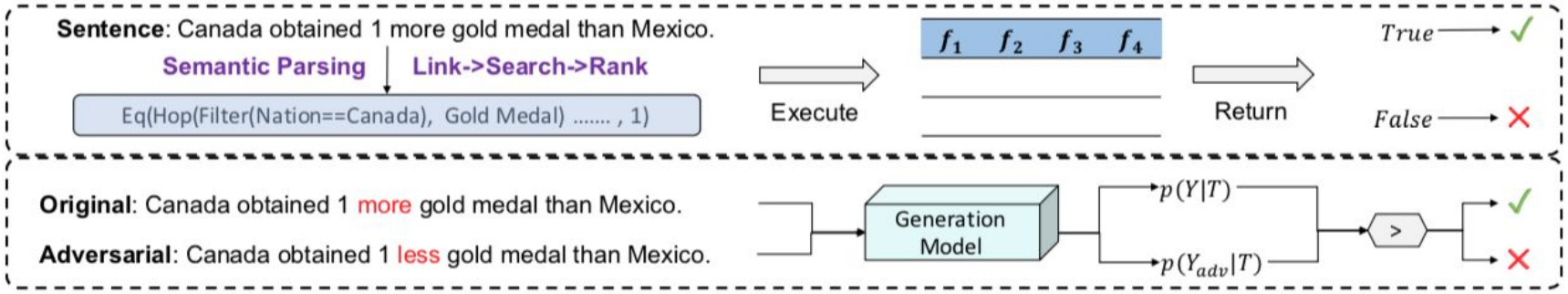
Logical Natural Language Generation

- Proposed fidelity metrics
  - Semantic-parsing Evaluation



Logical Natural Language Generation

- Proposed fidelity metrics
  - Semantic-parsing Evaluation
  - Adversarial Evaluation

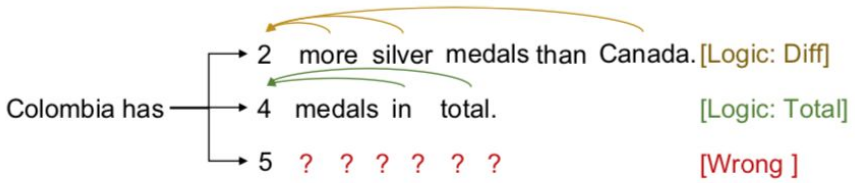


Logical Natural Language Generation

- Challenge for left-to-right generation paradigm:

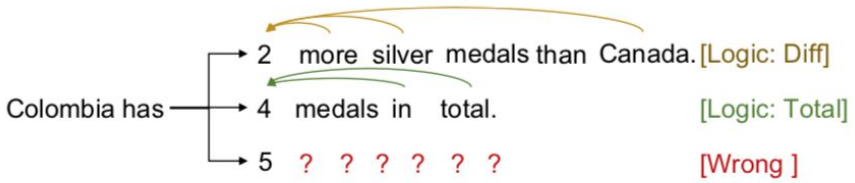
Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating



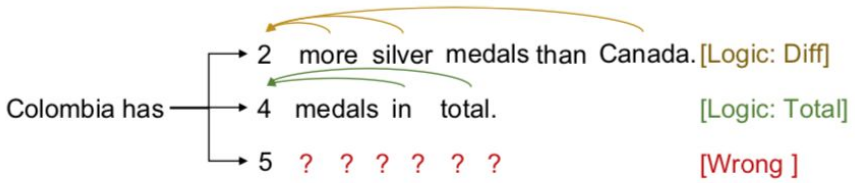
Logical Natural Language Generation

- Challenge for left-to-right generation
  - At 3rd step, the model needs to make a difficult decision.
  - Once decision made, it reaches a point of no return.
  - For example, once "5" is generated, there is no way to amend.



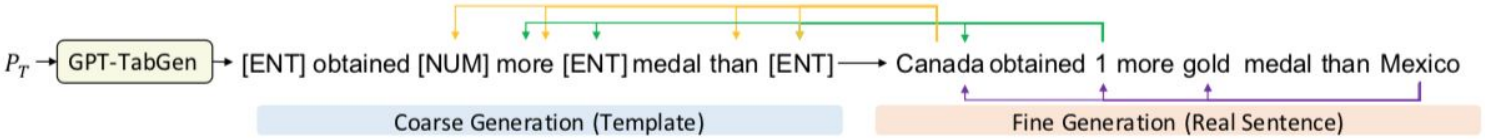
Logical Natural Language Generation

- Root Cause of the Challenge:
  - Mismatch of linguistic order vs. logical order
  - **"Canada":** logical order=1, linguistic order=8
  - **"more":** logical order=2, linguistic order=4
  - **"2":** logical order=3, linguistic order=3



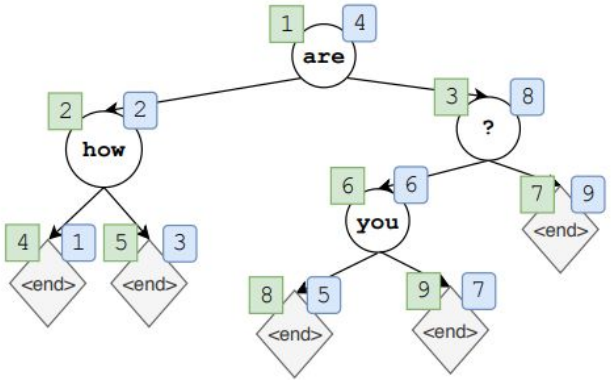
Logical Natural Language Generation

- Imperfect Cure: Coarse-to-fine Generation (Template->Realization)
  - The templates are preprocessed by masking entities and numbers.
  - The GPT-2 needs to first generate a template, and then



Logical Natural Language Generation

- Non-monotonic Generation
  - Optimal Case: is to generate the sentence according to logical order.
  - Challenge: No annotation of the logical order for the sentences.
  - We need to learn it in a weakly supervised manner.



Logical Natural Language Generation

- Conclusion
  - LogicNLG is a perfect testbed for studying non-monotonic generation.
  - How can we use unsupervised algorithm to induce the logical order of a given sentence?
  - How to leverage semantic parsing into the generation process to help it perform reasoning and inference?
  - Introduces new robustness and fact-checking requirements in evaluation.

Conclusion

- We show that inverse reinforcement learning is a secret weapon for addressing the diverse and complex nature of language generation problems.
- We conduct comprehensive evaluations with sample variance to show important observations, and make practical recommendations.
- We describe semantic parsing and adversarial examples as alternatives to evaluate complex logic based generation.

Open Challenges

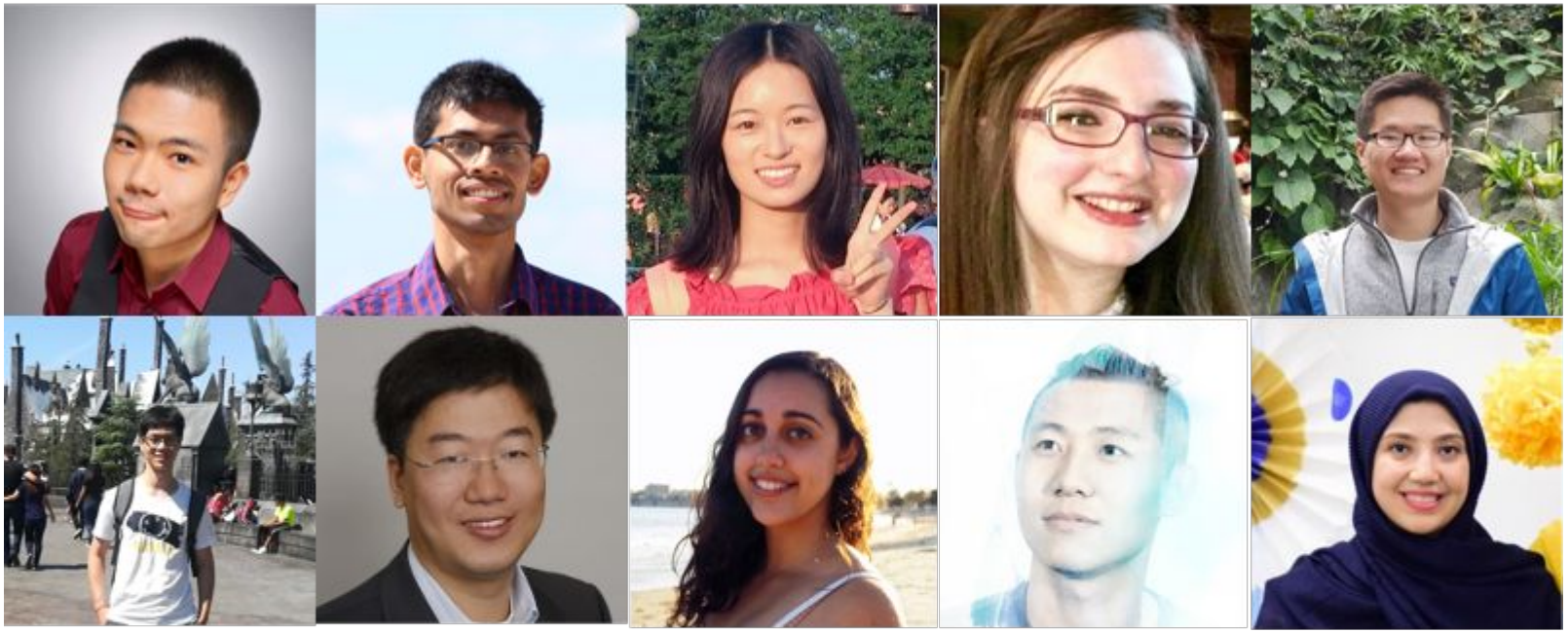
- TabFact (ICLR 2020); LogicNLG (ACL 2020); HybridQA (EMNLP 2020).

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

John E. Moss and Phillip Burton are both re-elected in the house of representative election in 1972.

Acknowledgment



Sponsors: Adobe, Amazon, ByteDance, DARPA, Facebook, Google, IBM, Intel, LogMeIn, NVIDIA, and Tencent.

Thank you!

- UCSB NLP Group: nlp.cs.ucsb.edu

- OpenSource Github repos:
  - https://github.com/wangwilliamyang?tab=stars