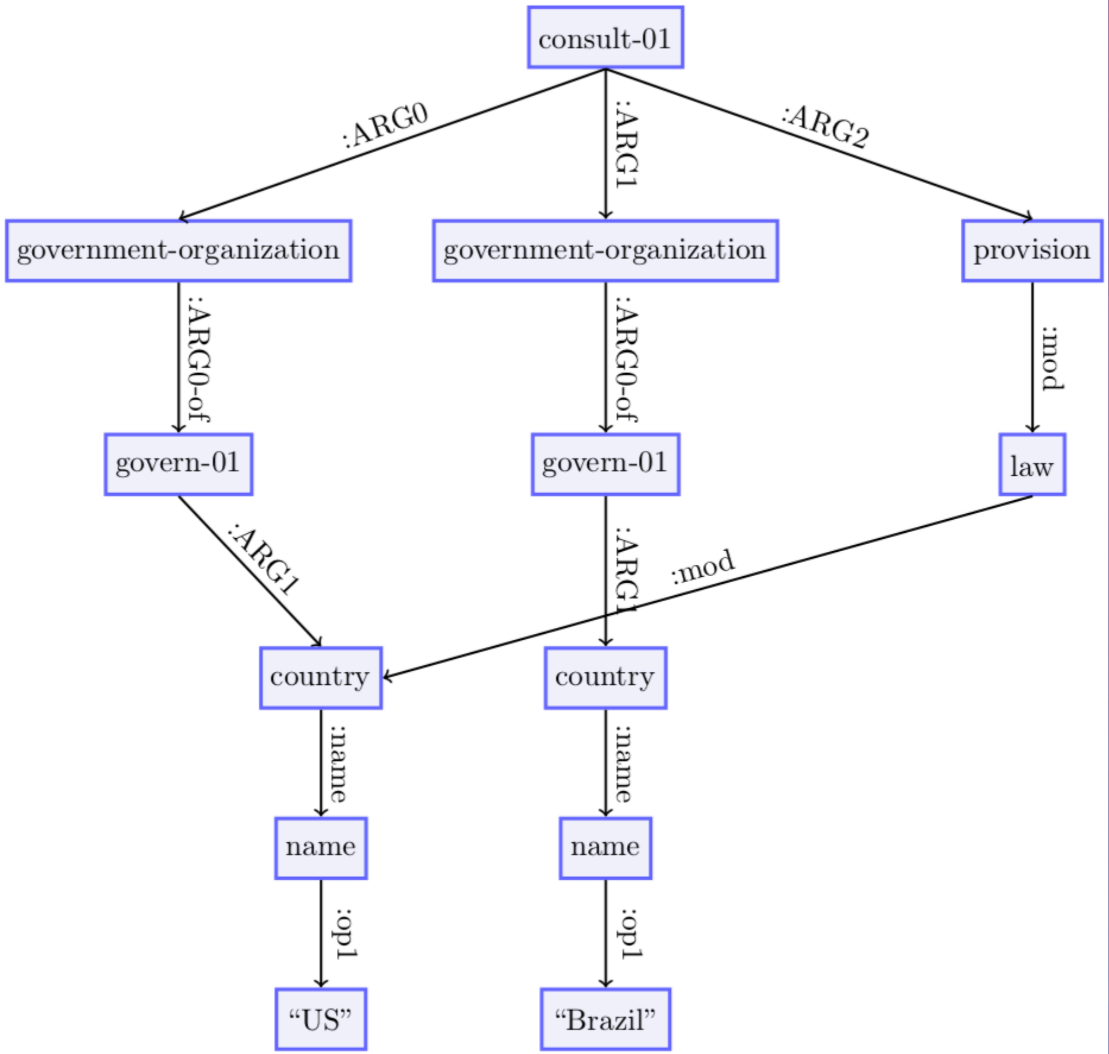
A HUMAN EVALUATION OF AMR-TO-ENGLISH GENERATION SYSTEMS

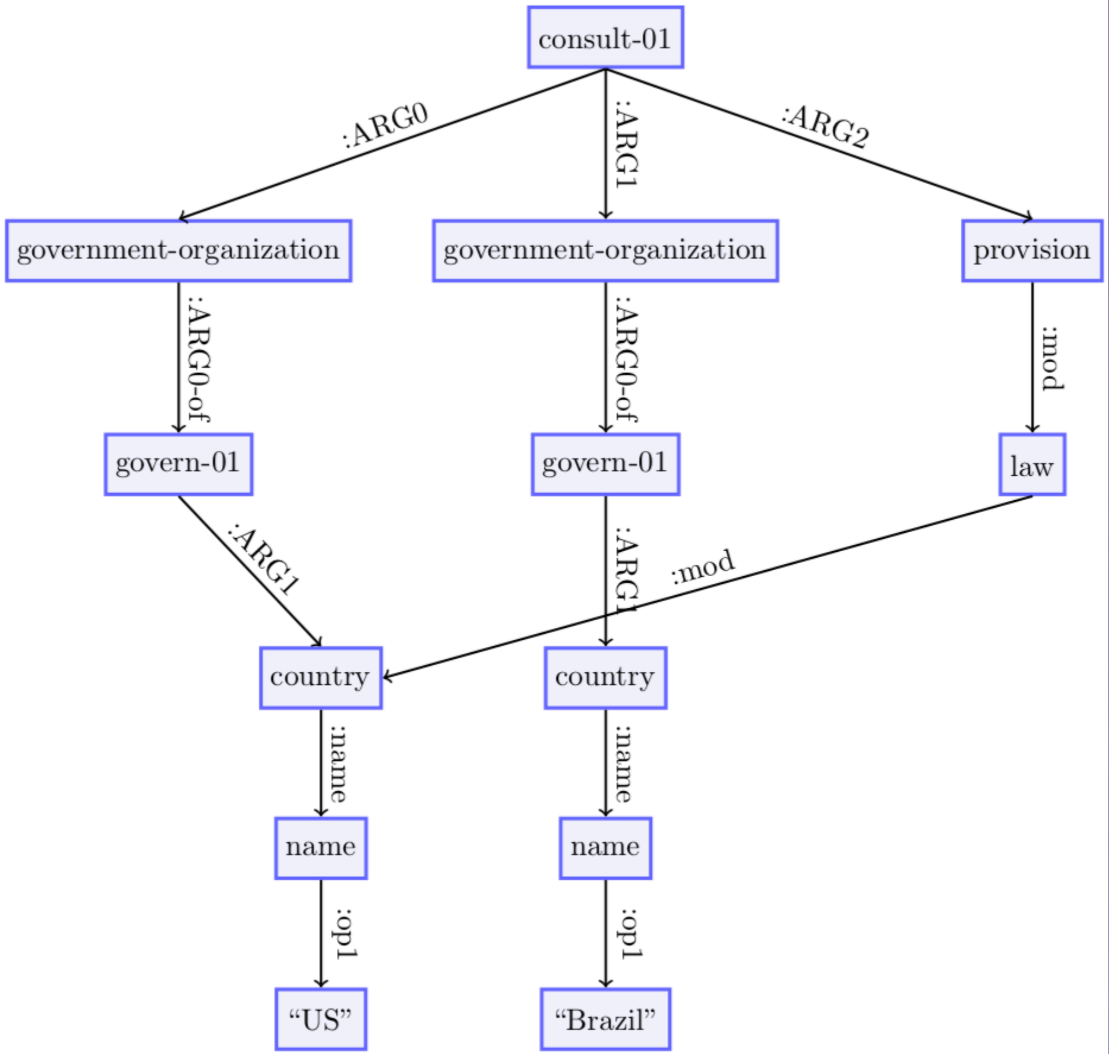
EMMA MANNING, SHIRA WEIN, NATHAN SCHNEIDER

GEORGETOWN UNIVERSITY

AMR (ABSTRACT MEANING REPRESENTATION)



The US government has consulted the Brazilian Government about the provisions of US law.



The US government has consulted the Brazilian Government about the provisions of US law.

AMR GENERATION

Sample generated sentences:

- the us government has consulted the brazilian government as a provision of brazilian law
- the us government has consulted with the brazil government for the provisions of the south korean law .
- the us government will consult the brazilian government with a canadian law provision .

LITERATURE: EVALUATING AMR GENERATION & NLG

For AMR: “We note that **BLEU**, which is often used as a generation metric, is **woefully inadequate** compared to human evaluation.” [May and Priyadarshi, 2017]

“State-of-the-art automatic evaluation metrics for NLG systems **do not sufficiently reflect human ratings**, which stresses the need for human evaluations” [Novikova et al., 2017]

“The evidence **does *not* support** using BLEU to evaluate other types of NLP systems (**outside of MT**) …. Also, BLEU **should not be the primary evaluation technique** in NLP papers.” [Reiter, 2018]

…and many more!

RESEARCH QUESTIONS

- How do recent AMR generation systems compare to each other?
  - Which is best overall?
  - What are their relative strengths and weaknesses?
- How well do automatic metrics capture human judgments of generation quality?
- What are common problems in the output of AMR generation systems?

SYSTEMS INCLUDED

- Seq2seq:
  - Konstas et al. (2017) – augmented with silver data
  - Zhu et al. (2019) – transformer-based
- Graph2seq:
  - Guo et al. (2019) – densely-connected graph convolutional network
  - Ribeiro et al. (2019) – dual graph representation
- Non-neural:
  - Manning (2019) – handwritten rules + ngram language model

DATA

- Standard LDC AMR dataset (LDC2017T10)
  - mix of news, blogs, forums, etc.
- Sampled 100 AMRs from test set
  - See paper for data sampling details!
- For each of those 100 AMRs, evaluated 6 sentences:
  - 1 reference + output from each of the 5 systems

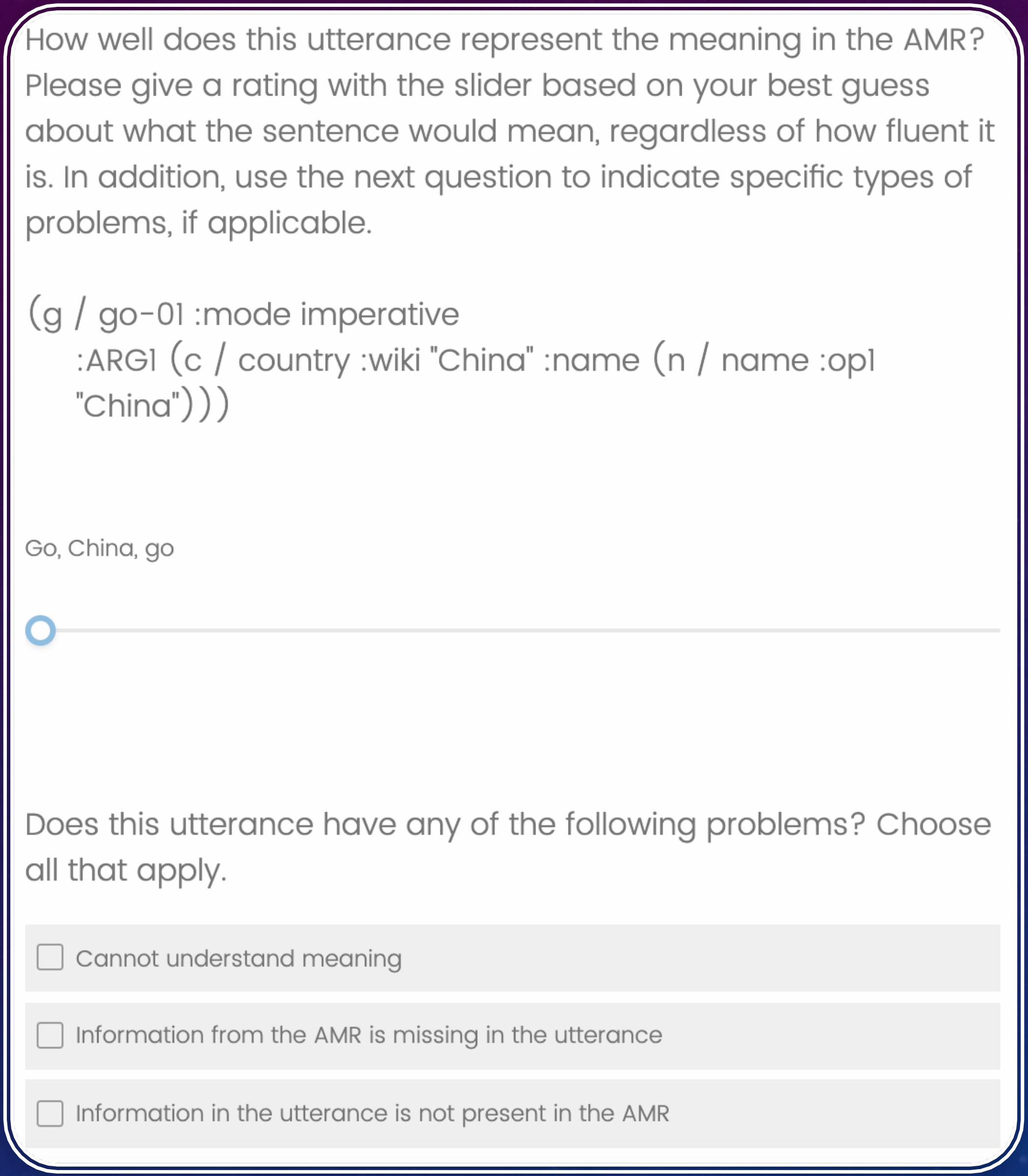
ANNOTATION

- 9 Annotators
  - Mostly PhD students
  - All trained in AMR
- All data double-annotated

ANNOTATION INTERFACE: FLUENCY

*“Please use the slider to indicate how well each [utterance] represents **fluent English**, like you might expect a person who is a native speaker of English to use.*

*Some of these may be **sentences fragments** rather than complete sentences, but can still be considered **fluent utterances**."*

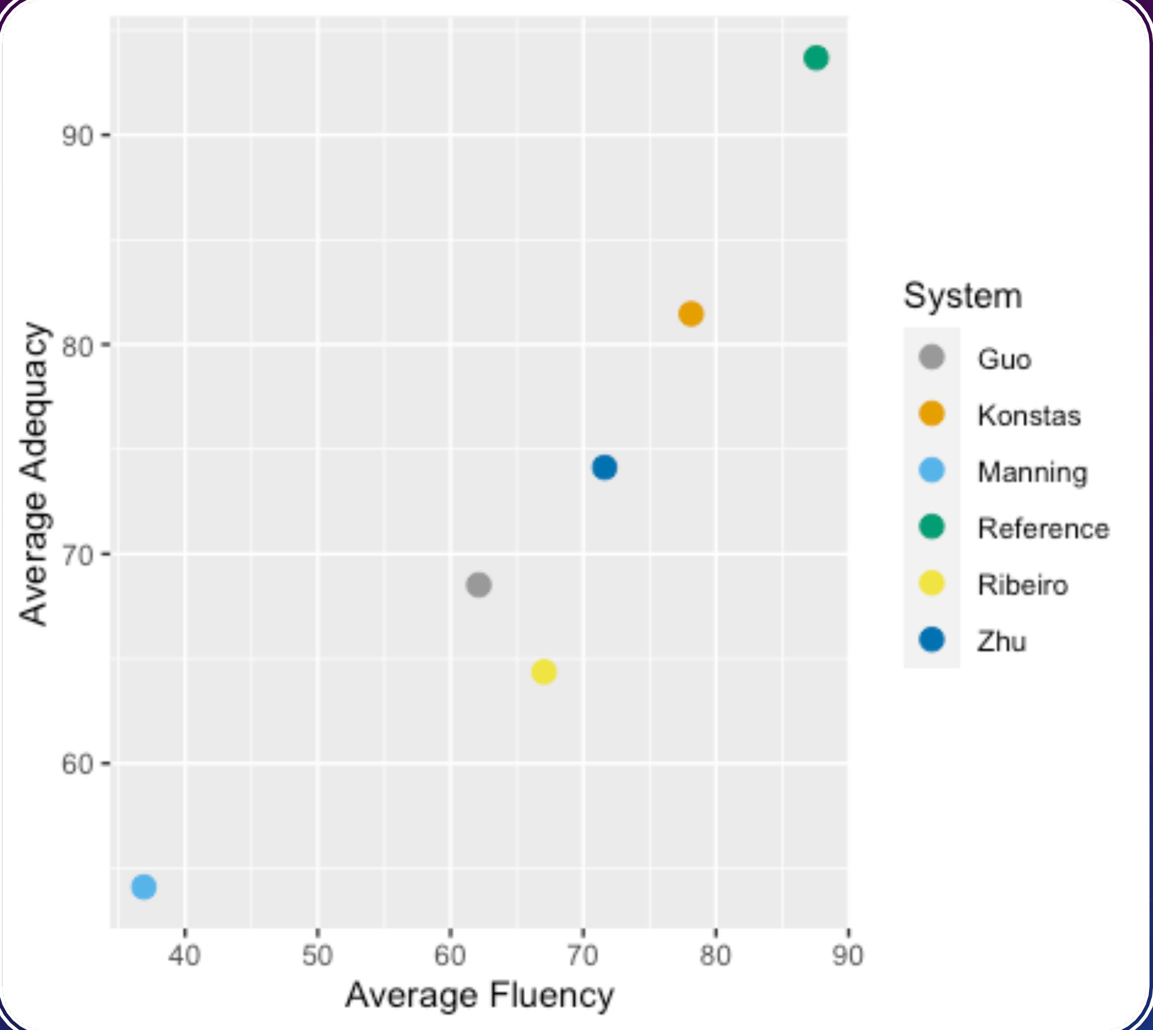


ANNOTATION INTERFACE: ADEQUACY

*"Your task is to determine how **accurately** the sentence expresses the **meaning** in the AMR."*

Also: Checkboxes for 3 error types

RESULTS & ANALYSIS



COMPARISON OF SYSTEMS: SCALAR

- Konstas best, followed by Zhu
- Manning scores lowest, especially for Fluency
- Ribeiro & Guo very close
  - Ribeiro slightly better for Fluency, Guo for Adequacy

COMPARISON TO AUTOMATIC METRICS: SYSTEM-LEVEL

- Score on the 100 sentences in our evaluation
    - See paper for full test set
- Similar ranking to humans, but not perfect

System	BLEU$_\uparrow$	METEOR$_\uparrow$	TER$_\downarrow$	CHRF++$_\uparrow$	BERTScore$_\uparrow$
Konstas	**38.1**	**39.2**	45.1	**64.3**	**95.0**
Zhu	**38.1**	38.7	**44.2**	56.3	92.7
Ribeiro	31.9	35.8	53.8	52.1	92.4
Guo	28.1	35.0	56.7	50.2	92.4
Manning	10.6	28.1	67.6	48.5	89.8

COMPARISON TO AUTOMATIC METRICS: SYSTEM-LEVEL

- Score on the 100 sentences in our evaluation
  - See paper for full test set
- Similar ranking to humans, but not perfect

System	BLEU↑	METEOR↑	TER↓	CHRF++↑	BERTScore↑
Konstas	**38.1**	**39.2**	45.1	**64.3**	**95.0**
Zhu	**38.1**	38.7	**44.2**	56.3	92.7
Ribeiro	31.9	35.8	53.8	52.1	92.4
Guo	28.1	35.0	56.7	50.2	92.4
Manning	10.6	28.1	67.6	48.5	89.8

COMPARISON TO AUTOMATIC METRICS: SYSTEM-LEVEL

- Score on the 100 sentences in our evaluation
  - See paper for full test set
- Similar ranking to humans, but not perfect
  - Doesn't capture fluency vs. adequacy

System	BLEU↑	METEOR↑	TER↓	CHRF++↑	BERTScore↑
Konstas	**38.1**	**39.2**	45.1	**64.3**	**95.0**
Zhu	**38.1**	38.7	**44.2**	56.3	92.7
Ribeiro	31.9	35.8	53.8	52.1	92.4
Guo	28.1	35.0	56.7	50.2	92.4
Manning	10.6	28.1	67.6	48.5	89.8

COMPARISON TO AUTOMATIC METRICS: SENTENCE-LEVEL

- Metrics correlate more strongly with **adequacy** than **fluency** (Spearman's Rho)
  - IAA was also better for adequacy
- **BERTScore** does best of these
- METEOR is also slightly better than BLEU

	Fluency	Adequacy
BLEU$_\uparrow$	0.40	0.52
METEOR$_\uparrow$	0.41	0.57
TER$_\downarrow$	-0.33	-0.43
CHRF++$_\uparrow$	0.32	0.47
BERTScore$_\uparrow$	**0.47**	**0.60**

QUALITATIVE ERROR ANALYSIS

System	# low F	# low A
Konstas	5	9
Zhu	9	16
Ribeiro	21	34
Guo	21	28
Manning	60	51
Reference	0	1
Total	116	139

- Identified sentences that received **low scores** on Fluency or Adequacy from **both annotators**
- **Manually inspected** low-scoring sentences for common issues

ERROR ANALYSIS: ADEQUACY

- All sentences with low adequacy scores were marked with at least one error type by at least one annotator
- Added information particularly concerning for real-world applications

Hallucination Example:

REFERENCE: A high-security Russian laboratory complex storing anthrax, plague and other deadly bacteria faces loosing electricity for lack of payment to the mosenergo electric utility.

RIBEIRO: the russian laboratory complex as a high - security complex will be faced with anthrax , prostitution , and and other killing bacterium losing electricity as it is lack of paying for mosenergo .

ERROR ANALYSIS: FLUENCY

- Common issues in neural systems:
  - Anonymization of named entities, quantities, and low-frequency/OOV items
  - Repetition of words and phrases

Anonymization example:

REFERENCE: Georgia labeled Russia's support an act of annexation

GUO: georgia labels russia 's support for the <unk> act .

Repetition example:

REFERENCE: and I happen to LIKE large lot development .

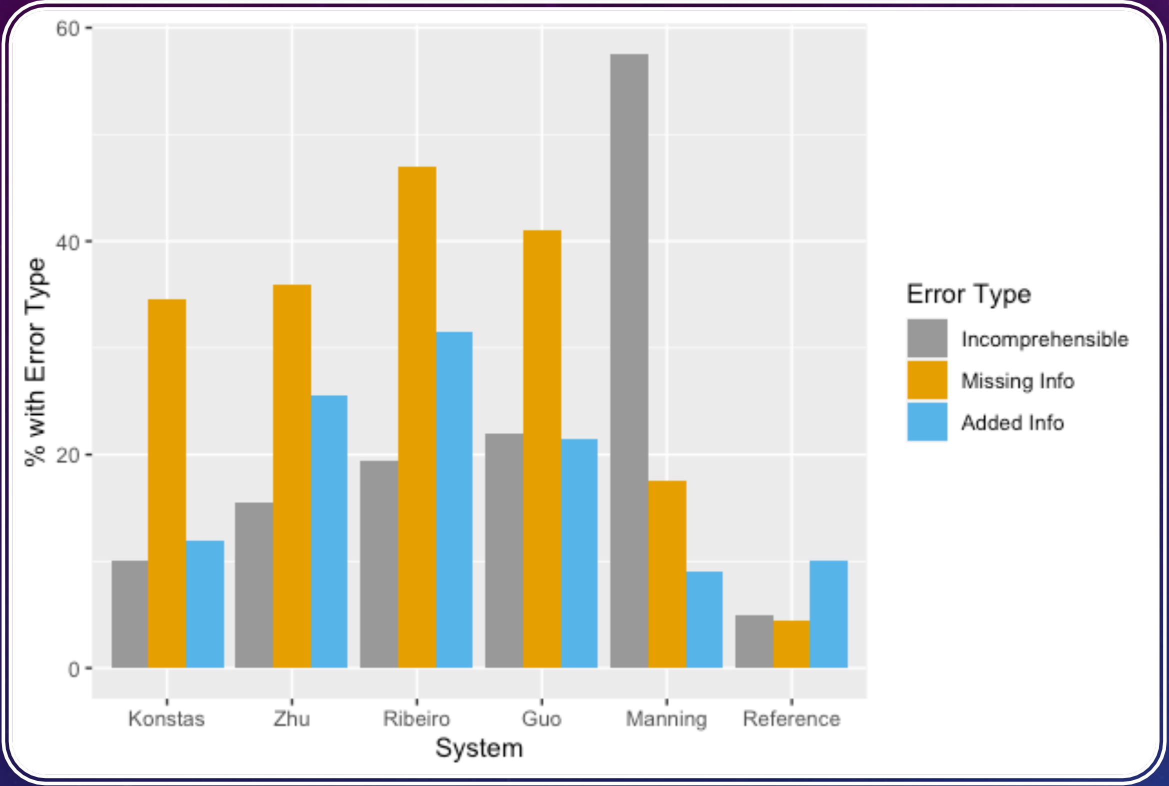
RIBEIRO: and i happen to like a large lot of a lot .

MAIN TAKEAWAYS

- Automatic evaluation can't replace human evaluation
  - BERTScore looks like the best existing metric
    - Need more human evaluation studies for this task to validate metrics!
  - We learn much more from multi-dimensional evaluation and manual inspection of output
- Major frontiers for improvement from neural systems:
  - Anonymization
  - Hallucination
  - Repetition

ADDITIONAL SLIDES

COMPARISON OF SYSTEMS: ERROR TYPES



- Incomprehensibility corresponds with Fluency rankings
- Missing and Added Information *mostly* correspond with Adequacy rankings
  - Notable exception: Manning has **lowest** rates of these errors