A HUMAN EVALUATION OF AMR-TO-ENGLISH GENERATION SYSTEMS

EMMA MANNING, SHIRA WEIN, NATHAN SCHNEIDER
GEORGETOWN UNIVERSITY
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.

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.
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
“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.”
How well does this utterance represent the meaning in the AMR?
Please give a rating with the slider based on your best guess about what the sentence would mean, regardless of how fluent it is. In addition, use the next question to indicate specific types of problems, if applicable.

(g / go-01 :mode imperative
 :ARG1 (c / country :wiki "China" :name (n / name :op1 "China")))

Go, China, go

Does this utterance have any of the following problems? Choose all that apply.

☐ Cannot understand meaning
☐ Information from the AMR is missing in the utterance
☐ Information in the utterance is not present in the AMR

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

Also: Checkboxes for 3 error types
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

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU↑</th>
<th>METEOR↑</th>
<th>TER↓</th>
<th>CHRF++↑</th>
<th>BERTScore↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konstas</td>
<td>38.1</td>
<td>39.2</td>
<td>45.1</td>
<td>64.3</td>
<td>95.0</td>
</tr>
<tr>
<td>Zhu</td>
<td>38.1</td>
<td>38.7</td>
<td>44.2</td>
<td>56.3</td>
<td>92.7</td>
</tr>
<tr>
<td>Ribeiro</td>
<td>31.9</td>
<td>35.8</td>
<td>53.8</td>
<td>52.1</td>
<td>92.4</td>
</tr>
<tr>
<td>Guo</td>
<td>28.1</td>
<td>35.0</td>
<td>56.7</td>
<td>50.2</td>
<td>92.4</td>
</tr>
<tr>
<td>Manning</td>
<td>10.6</td>
<td>28.1</td>
<td>67.6</td>
<td>48.5</td>
<td>89.8</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU↑</th>
<th>METEOR↑</th>
<th>TER↓</th>
<th>CHRF++↑</th>
<th>BERTScore↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konstas Zhu</td>
<td>38.1</td>
<td>39.2</td>
<td>45.1</td>
<td>64.3</td>
<td>95.0</td>
</tr>
<tr>
<td>Ribeiro Guo</td>
<td>31.9</td>
<td>35.8</td>
<td>53.8</td>
<td>52.1</td>
<td>92.4</td>
</tr>
<tr>
<td>Manning</td>
<td>10.6</td>
<td>28.1</td>
<td>67.6</td>
<td>48.5</td>
<td>89.8</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU↑</th>
<th>METEOR↑</th>
<th>TER↓</th>
<th>CHRF++↑</th>
<th>BERTScore↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konstas Zhu</td>
<td>38.1</td>
<td>39.2</td>
<td>45.1</td>
<td>64.3</td>
<td>95.0</td>
</tr>
<tr>
<td>Ribeiro Guo</td>
<td>31.9</td>
<td>35.8</td>
<td>53.8</td>
<td>52.1</td>
<td>92.4</td>
</tr>
<tr>
<td>Manning</td>
<td>10.6</td>
<td>28.1</td>
<td>67.6</td>
<td>48.5</td>
<td>89.8</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Fluency</th>
<th>Adequacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU↑</td>
<td>0.40</td>
<td>0.52</td>
</tr>
<tr>
<td>METEOR↑</td>
<td>0.41</td>
<td>0.57</td>
</tr>
<tr>
<td>TER↓</td>
<td>-0.33</td>
<td>-0.43</td>
</tr>
<tr>
<td>CHRF++↑</td>
<td>0.32</td>
<td>0.47</td>
</tr>
<tr>
<td>BERTScore↑</td>
<td>0.47</td>
<td>0.60</td>
</tr>
</tbody>
</table>
QUALITATIVE ERROR ANALYSIS

- 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