REFERENCELESS PARSING-BASED EVALUATION OF AMR-TO-ENGLISH GENERATION

EMMA MANNING & NATHAN SCHNEIDER
GEORGETOWN UNIVERSITY
INTRODUCTION

• Evaluating NLG is notoriously difficult: one-to-many problem
• Reference-based metrics are popular but flawed
• We explore a referenceless alternative for AMR-to-English generation: parsing-based evaluation
• Also suggested by Opitz & Frank (EACL 2021); we evaluate the approach in new ways:
  • Comparison to human judgments
  • Manual editing experiment
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.
NEWER IDEA: PARSING-BASED EVALUATION

Original Sentence (Reference)
- The US government has consulted the Brazilian Government about the provisions of US law.

Gold AMR

Generated Sentence
- the US government has consulted with the Brazil government over the provisions of US law.

Parsed AMR

Requires: parser

Compare
Requires: similarity metric
METHODS: DATA

• Human judgment data produced in previous study (Manning et al. 2020)
• Judgments on 100 sentences each from 5 generation systems (+ references)
• Scalar judgments of fluency and adequacy
  • For this we’re most interested in adequacy!
REFERENCE-BASED BASELINES

- Correlations of popular RBMs with our human judgments

<table>
<thead>
<tr>
<th>Metric</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>
<tr>
<td>BLEURT↑</td>
<td>0.60</td>
<td>0.69</td>
</tr>
</tbody>
</table>
EXPERIMENT 1: AUTOMATIC PARSING

What should we use for the parser and similarity metric? How much does it matter?

• Tried 3 automatic parsers:
  • Baseline: JAMR (Flanigan et al., 2014, 2016)
  • Medium: Lyu & Titov (2018)
  • Best: Cai & Lam (2020)

• And 3 similarity metrics:
  • Standard: Smatch (Cai & Knight, 2013)
  • Minor Variant: Smatch$_{100} +$seed – Smatch, but more reliable & reproducible
  • Bigger Variant: $S^2$match (Opitz et al., 2020)
EXPERIMENT 1: RESULTS

• Better parser -> better results!
• Similarity metric doesn’t matter much
• Correlation with adequacy lower than for most automatic metrics 😞

<table>
<thead>
<tr>
<th></th>
<th>S\text{match}_4</th>
<th>S\text{match}_{100+ \text{seed}}</th>
<th>S^2\text{match}</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAMR</td>
<td>0.358</td>
<td>0.356</td>
<td>0.362</td>
</tr>
<tr>
<td>Lyu &amp; Titov</td>
<td>0.462</td>
<td>0.460</td>
<td>0.465</td>
</tr>
<tr>
<td>Cai &amp; Lam</td>
<td>0.495</td>
<td>0.492</td>
<td>0.494</td>
</tr>
</tbody>
</table>
The US government has consulted with the Brazilian Government about the provisions of US law.
EXPERIMENT 2: MANUAL EDITING

- State-of-the-art AMR parser introduces a lot of errors for this data!
- How well could this work with an even better parser?
- Idea: Correct the parses myself to approximate an upper bound!
the US government has consulted with the Brazil government for the provisions of the South Korean law.
MANUAL EDITING RESULTS

- After editing, correlation with adequacy increases to 0.66 😎
- Indicates that this will work better automatically as parsers continue to improve
CONCLUSION: MAIN TAKEAWAYS

• Of existing automatic reference-based metrics, BLEURT and BERTScore look pretty good for AMR generation
  • Please stop relying on BLEU!
  • But: concerns about transparency, bias, etc.
• Parsing-based referenceless evaluation has potential, but is currently limited by parser accuracy