The E2E Dataset: New Challenges for End-to-End NLG

Jekaterina Novikova, Ondrej Dusek, Verena Rieser

The E2E Dataset

- Restaurant information domain
- Collected via CrowdFlower
- using pictorial instructions to get more diverse texts

Dataset Size

<table>
<thead>
<tr>
<th></th>
<th>No. of instances</th>
<th>No. of unique MRs</th>
<th>Refs/MR</th>
<th>Slots/MR</th>
<th>W/Ref</th>
<th>W/Sent</th>
<th>Sents/Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2E</td>
<td>50,602</td>
<td>5,751</td>
<td>8.1</td>
<td>(2–16)</td>
<td>5.43</td>
<td>20.1</td>
<td>14.3</td>
</tr>
<tr>
<td>SFRest</td>
<td>5,192</td>
<td>1,950</td>
<td>1.82</td>
<td>(1–101)</td>
<td>2.86</td>
<td>8.53</td>
<td>8.53</td>
</tr>
<tr>
<td>Bagel</td>
<td>404</td>
<td>202</td>
<td>2</td>
<td>(2–2)</td>
<td>5.41</td>
<td>11.54</td>
<td>11.54</td>
</tr>
</tbody>
</table>

Syntactic Variation

- Slightly fewer syntactically simple sentences
- Slightly more sentences with complex syntactic structure

Baseline System

- LSTM sequence-to-sequence model with attention
- Beam search
- Reranking by MR classification
- penalize differences w.r.t. the input MR
- Requires delexicalization

Lexical Richness

- Higher diversity (Mean Segmental Type-Token Ratio)
- Higher proportion of rare words (Lexical Sophistication)
- A challenge to word-based generators that don't handle out-of-vocabulary words

Semantic Adequacy

More challenging – requires content selection:
- Crowd workers asked to realize relevant information
  - Some (redundant) information is omitted
  - Some information is added

Development set performance:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.6925</td>
</tr>
<tr>
<td>NIST</td>
<td>8.4781</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.4703</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.7257</td>
</tr>
<tr>
<td>CIDER</td>
<td>2.3987</td>
</tr>
</tbody>
</table>

E2E NLG Challenge

- take part!
- download the data at: http://bit.ly/e2e-nlg
- submit results by October 15