Referenceless Quality Estimation for Natural Language Generation

Ondřej Dušek, Jekaterina Novikova & Verena Rieser
Interaction Lab, Heriot-Watt University, Edinburgh

Quality Estimation for NLG
- estimate NLG system output quality by comparing with source MR only – no reference texts needed
- useful for system development: word-overlap metrics such as BLEU unreliable + need costly references
- useful at runtime: reranking, triggering fallback strategies

<table>
<thead>
<tr>
<th>Instance</th>
<th>Humans</th>
<th>BLEU</th>
<th>METEOR</th>
<th>ROUGE</th>
<th>CIDEr</th>
<th>Our system (S4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR System output Reference</td>
<td>inform(name=&quot;la ciccia\area=bernal heights\price_range=moderate&quot;) la ciccia, is in the bernal heights area with a moderate price range.</td>
<td>5.51</td>
<td>0.000</td>
<td>3.711</td>
<td>3.521</td>
<td>2.117</td>
</tr>
<tr>
<td>MR System output Reference</td>
<td>inform(name=&quot;intercontinental san francisco\price_range=pricey&quot;) the intercontinental san francisco is in the pricey price range.</td>
<td>2.45</td>
<td>0.707</td>
<td>0.433</td>
<td>5.52</td>
<td>2.318</td>
</tr>
</tbody>
</table>

Our NLG Quality Estimation Model
1) RNN GRU encoders for source MR + NLG system output to be rated
2) fully connected tanh layers
3) final layer – linear, predicting rating as a floating point number
- trained by minimising mean square error against human-assigned ratings
- delexicalization to fight data sparsity

Experiments & Results
- 5-fold cross-validation
- always better correlations than metrics
- lower than MT (less data & harder)
- 21% improvement with synthetic data
- with synthetic data:
  better MAE/RMSE than constant baseline
- cross-domain & cross-system performance poor, but small amounts of in-set data help greatly

Our Dataset
- outputs of 3 NLG systems on 3 datasets
- TGen & LOLS & RNNLG
- BAGEL & SFHot & SFRest
- CrowdFlower used to obtain human ratings
  - overall quality rating on a 1–6 Likert scale
  - 3+ ratings per system output
  - using medians for consistency
  - 2,460 instances total
- synthesising additional data:
  a) introducing artificial errors & lowering ratings
  - artificial errors in training system outputs
  - human references for training MRs & errors
  - human refs from source NLG training sets & errors
  - human references from test MRs & errors
  - human refs from whole source NLG sets & errors
  - code available at: https://github.com/tuetschek/ratpred
  - a) additional human references from source NLG datasets (with “perfect” ratings)
  - up to 78k synthesised instances

Conclusions
- 1st quality estimation system for NLG
- no need for references, better segment-level correlations than word-overlap metrics
- improvements over constant baseline suggest occasional large errors
- code available at: https://github.com/tuetschek/ratpred
- future work: better networks, better error synthesis, more data (E2E NLG challenge), post-edits prediction