Cleaning E2E training data fixes up to 97% NLG semantic errors

Semantic Noise Matters
for Neural Natural Language Generation

Ondřej Dušek*  David M. Howcroft†  Verena Rieser†

Research Questions
► Does noisy data matter for Neural Natural Language Generation (NNLG)?
► Can NNLG systems learn to ignore errors in training data by generalising away from them?

The Problem: Noisy Training Data
Crowdworkers introduce more noise than expected
► Insertions
► Deletions
► Alterations
Measured by calculating the Semantic Error Rate
\[
\text{SER} = \frac{\text{#added} + \text{#missing} + \text{#wrong value}}{\text{#slots}}
\]

End-to-End Generation Challenge Corpus (E2E)
► collected via crowdsourcing
► used by 17 teams in the E2E challenge
► used in 13 published papers since
11–17% SER in the E2E dataset
► approx. 40% of references include ≥ 1 error

Fixing the E2E NLG Challenge Dataset
We cleaned the data! (a little goes a long way)
► our heuristic script for SER also provides corrections
► good accuracy but not perfect
► SER 4.2%; 19.5% of references with errors
► some cleaned MRs from TRAIN&DEV overlapped TEST
► these instances were removed
► systems trained on cleaned data can be evaluated on original TEST
► cleaned data: fewer instances, more distinct MRs
► more challenging for training

Dataset statistics

<table>
<thead>
<tr>
<th>Dataset Part</th>
<th>MRs</th>
<th>Refs</th>
<th>SER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRAIN</td>
<td>4,862</td>
<td>42,061</td>
<td>17.69</td>
</tr>
<tr>
<td>Original</td>
<td>547</td>
<td>4,672</td>
<td>11.42</td>
</tr>
<tr>
<td>TEST</td>
<td>630</td>
<td>4,693</td>
<td>11.49</td>
</tr>
<tr>
<td>TRAIN</td>
<td>8,362</td>
<td>33,525</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Cleaned</td>
<td>1,132</td>
<td>4,299</td>
<td>(0.00)</td>
</tr>
<tr>
<td>TEST</td>
<td>1,358</td>
<td>4,693</td>
<td>(0.00)</td>
</tr>
</tbody>
</table>

Table: # of distinct MRs, # of reference texts, and SER as measured by our script.

Example MR Fixes by Our Script

Original MR
At the riverside near The Portland Arms, Cotto is a coffee shop that serves English food at less than £20 and has low customer rating.

Correction: removed price range; changed area

Example corrections

Reference: Located near The Portland Arms in riverside, the Cotto coffee shop serves English food with a price range of $20 and a low customer rating.
Correction: incorrectly(!) removed price range
– our script’s slot patterns are not perfect

Impact on Neural NLG Systems

Cleaned data can reduce errors by up to 97%

Results

<table>
<thead>
<tr>
<th>System</th>
<th>TRAIN</th>
<th>BLEU</th>
<th>NIST</th>
<th>A</th>
<th>M</th>
<th>V</th>
<th>SER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2Seq</td>
<td>Original</td>
<td>63.37</td>
<td>7.71</td>
<td>0.06</td>
<td>15.77</td>
<td>0.11</td>
<td>15.94</td>
</tr>
<tr>
<td></td>
<td>Cleaned added</td>
<td>64.40</td>
<td>7.96</td>
<td>0.01</td>
<td>13.08</td>
<td>0.00</td>
<td>13.09</td>
</tr>
<tr>
<td></td>
<td>Cleaned missing</td>
<td>66.28</td>
<td>8.52</td>
<td>0.14</td>
<td>2.26</td>
<td>0.22</td>
<td>2.61</td>
</tr>
<tr>
<td></td>
<td>Cleaned</td>
<td>65.87</td>
<td>8.64</td>
<td>0.20</td>
<td>0.56</td>
<td>0.21</td>
<td>1.27</td>
</tr>
<tr>
<td>TGen</td>
<td>Original</td>
<td>66.41</td>
<td>8.55</td>
<td>0.14</td>
<td>4.11</td>
<td>0.03</td>
<td>4.27</td>
</tr>
<tr>
<td></td>
<td>Cleaned added</td>
<td>66.23</td>
<td>8.55</td>
<td>0.04</td>
<td>3.04</td>
<td>0.00</td>
<td>3.09</td>
</tr>
<tr>
<td></td>
<td>Cleaned missing</td>
<td>67.00</td>
<td>8.68</td>
<td>0.06</td>
<td>0.44</td>
<td>0.03</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Cleaned</td>
<td>66.24</td>
<td>8.68</td>
<td>0.10</td>
<td>0.02</td>
<td>0.00</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table: A = % instances with added slots, M = missed slots, V = wrong values

Data clean: fewer instances, more distinct MRs
► more challenging for training

Conclusions and Future Work

Semantic noise matters
► Crowdsourced datasets are noisy, so clean your data!

Get our data & code here:
https://github.com/tuetschek/e2e-cleaning

What’s next?
► continuing to improve data & checking effects on diversity