Evaluating the Factual Correctness for Abstractive Summarization

Yuhui Zhang*  
yuhuiz@stanford.edu

The Problem
• Text summarization: extractive, abstractive.
• Applications: news, laws, clinical, biomedical.
• However, 30%[3] of summaries generated by abstractive models contain factual inconsistencies.

This is a critical issue for neural abstractive summarization.

How can we evaluate the factual correctness?

Abstractive Summarization
Most recent works about abstractive summarization are based on sequence-to-sequence (seq2seq) architecture:

**Seq2Seq**: Basic seq2seq architecture.
**Pointer-Generator**[4]: Allow to copy from source text.
**ML**[8]: Attend over source and target text separately.
**ML+RL**[3]: Training with reinforcement learning.

Summaries are generated and sampled from CNN/DM dataset using these models.[4]

Factual Score
**Fact Extractor**: We use AllenNLP open information extraction (OpenIE) toolkit to extract facts from text. Each fact is a triple (argument, predicate, argument).
**Fact Encoder**: We concatenate the fact triple and use Google universal sentence encoder to generate fact embedding.
**Factual Scorer**: We use cosine-similarity to estimate the relevance of each fact pair, and then compute precision, recall and F1 by averaging across facts from generated summary and facts from reference summary.

Falsity Attack
We manually generate false examples with 5 simple text transformations:

Truth: Andrew is a professor at Stanford, and he teaches CS 230 for many years.
Falsity: Andrew is not a professor at Berkeley, and she teaches CS 231 for many years.

- Negation
- Entity Swap
- Pronoun Swap
- Number Swap
- Noise Injection

Results
**Evaluations of abstractive summarization with…**
- **ROUGE-L Score** (n-gram hard-match evaluation)
- **BERT Score**[5] (token soft-match evaluation)
- **Factual Score** (factual correctness evaluation)

<table>
<thead>
<tr>
<th>System</th>
<th>ROUGE-L</th>
<th>BERT Score</th>
<th>FACT Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2seq</td>
<td>19.94</td>
<td>55.01</td>
<td>39.61</td>
</tr>
<tr>
<td>Pointer-Generator</td>
<td>27.62</td>
<td>60.20</td>
<td>43.49</td>
</tr>
<tr>
<td>ML</td>
<td>26.57</td>
<td>60.35</td>
<td>42.83</td>
</tr>
<tr>
<td>ML+RL</td>
<td>28.63</td>
<td>61.72</td>
<td>45.13</td>
</tr>
</tbody>
</table>

Factual score is consistent with human evaluation:  
ML+RL > Pointer-Generator >≈ ML > Seq2seq

Relation of factual score with…
- **ROUGE-L Score**
- **BERT Score** (more strongly correlated)

Discussion and Future Work
- Encoder is much more sensitive to noun phrases than number, pronoun and negation → Design better **fact encoder** architecture.
- OpenIE outputs contain duplicated facts and noisy facts → Try different ways to **denoise** OpenIE outputs.
- **Reinforcement learning** on factual score.

* Research project with Yuhao Zhang and Christopher D Manning.