Revisiting Iterative Back-Translation from the Perspective Of Compositional Generalization

Yinuo Guo1, Huailei Zhu2, Zeqi Lin3, Bei Chen1, Jian-Guang Lou3, Dongmei Zhang3
Peking University1 & Beihang University2 & Microsoft Research3

Compositional Generalization
Compositional generalization (CG) is an ability to understand and produce unseen combinations of seen components.

Benchmark
• SCAN (Lake et al. 2018): input natural language commands paired with output action sequences.
• CFQ (Keysers et al. 2020): input natural language questions paired with SPARQL queries (Freebase knowledge graph).

Three Research Questions
• RQ1: How does IBT affect neural seq2seq models’ ability to generalize to more combinations beyond parallel data?
• RQ2: If IBT is useful from the perspective of compositional generalization, what is the key that contributes to its success?
• RQ3: Is there a way to further improve the performance?

Semi-Supervised Learning & Iterative Back-Translation
• Parallel data are limited and expensive.
• Monolingual data are cheap and abundant, containing lots of unseen combinations.

• We focus on Iterative Back-Translation (IBT), a simple yet effective semi-supervised method that has been successfully applied in machine translation.

Evaluate on CFQ & SCAN

<table>
<thead>
<tr>
<th>Model</th>
<th>MCD1</th>
<th>MCD2</th>
<th>MCD3</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM+Attn</td>
<td>28.9 ± 1.8</td>
<td>5.0 ± 0.8</td>
<td>10.4 ± 0.6</td>
</tr>
<tr>
<td>Transformer</td>
<td>34.9 ± 1.1</td>
<td>8.2 ± 0.3</td>
<td>10.4 ± 1.1</td>
</tr>
<tr>
<td>Uni-Transformer</td>
<td>37.1 ± 2.2</td>
<td>8.1 ± 1.6</td>
<td>11.3 ± 0.9</td>
</tr>
<tr>
<td>TT-ITB</td>
<td>41.4 ± 4.9</td>
<td>39.1 ± 2.2</td>
<td>31.4 ± 5.7</td>
</tr>
<tr>
<td>GRU+Attn (Our) - encoder/decoder</td>
<td>52.6 ± 0.22</td>
<td>6.9 ± 0.25</td>
<td>9.3 ± 0.25</td>
</tr>
<tr>
<td>encoder/decoder</td>
<td>58.4 ± 4.9</td>
<td>57.8 ± 4.9</td>
<td>64.6 ± 4.9</td>
</tr>
<tr>
<td>encoder-only</td>
<td>55.2 ± 3.1</td>
<td>11.5 ± 6.9</td>
<td>81.3 ± 1.6</td>
</tr>
<tr>
<td>transformer</td>
<td>55.4 ± 0.7</td>
<td>61.4 ± 6.5</td>
<td>68.3 ± 2.2</td>
</tr>
</tbody>
</table>

• IBT Substantially improves the performance on CG benchmarks.
• Better monolingual data, better results.

Impact of Error-Prone Data & Perturbations

• BT: Even noise pseudo-parallel data can bring gains.
  • As they bring implicit knowledge of unseen combinations.
• BT+OTF: Perturbations brought by OTF (on-the-fly) is very important.
  • Pseudo-parallel data are generated dynamically, which prevent learning specific incorrect bias.

Curriculum Iterative Back-Translation

• Help reduce errors more efficiently.
• CIBT: during the training process:
  • start out with easy monolingual data
  • then gradually increase the difficulty.
• Curriculum learning benefits iterative back-translation.
• Curriculum learning is more beneficial to difficult data than simple data.

References

Contact
• gyn0806@pku.edu.cn
• Zeqi.Lin@microsoft.com

Paper