Revisiting Iterative Back-Translation from the Perspective of Compositional Generalization

郭一诺 (gyn0806@pku.edu.cn)
Our Team - Main Contributors

Peking University
Yinuo Guo

Beihang University
Hualei Zhu

Microsoft Asia
Zeqi Lin

Microsoft Asia
Bei Chen

Microsoft Asia
Jian-Guang

Microsoft Asia
Dongmei Zhang
Compositional Generalization

• The algebraic ability to understand and produce unseen combinations of seen atoms.

Infinite use of finite means. —— Chomsky

<table>
<thead>
<tr>
<th>Natural Language</th>
<th>Programming Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>run twice</td>
<td>RUN RUN</td>
</tr>
<tr>
<td>jump and walk</td>
<td>JUMP WALK</td>
</tr>
</tbody>
</table>

Train

Test

jump twice and run      \(\Rightarrow\) JUMP JUMP RUN
Background: Seq2seq Tasks in NLP

- Machine Translation
- Semantic Parsing
- Summarization
- ...

Diagram:
- Encoder
  - RNN
  - h₁ → h₂ → h₃
  - x₁ → x₂ → x₃
- Decoder
  - y₁ → y₂
  - RNN
  - Encoder Vector
Semi-Supervised Learning

- Parallel data are limited and expensive
- Monolingual data are cheap and abundant, containing lots of unseen combinations
- Hypothesis: semi-supervised learning can enable models to understand and produce much more combinations beyond labelled data, thus tackling the bottleneck of lacking compositional generalization.
Iterative Back-Translation

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

• RQ1: How does IBT affect compositional generalization of seq2seq models?
  • Yes

• RQ2: What is the key that contributes to the success of IBT?
  • Quality of pseudo parallel data & Perturbations

• RQ3: How to further improve the performance of IBT?
  • Curriculum Iterative Back-translation
Evaluate on CFQ & SCAN

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

<table>
<thead>
<tr>
<th>Models</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>10.8 ± 0.6</td>
</tr>
<tr>
<td>Transformer</td>
<td>34.9 ± 1.1</td>
<td>8.2 ± 0.3</td>
<td>10.6 ± 1.1</td>
</tr>
<tr>
<td>Uni-Transformer</td>
<td>37.4 ± 2.2</td>
<td>8.1 ± 1.6</td>
<td>11.3 ± 0.3</td>
</tr>
<tr>
<td>CGPS</td>
<td>13.2 ± 3.9</td>
<td>1.6 ± 0.8</td>
<td>6.6 ± 0.6</td>
</tr>
<tr>
<td>T5-11B</td>
<td>61.4 ± 4.8</td>
<td>30.1 ± 2.2</td>
<td>31.2 ± 5.7</td>
</tr>
<tr>
<td>GRU+Attn (Ours)</td>
<td>32.6 ± 0.22</td>
<td>6.0 ± 0.25</td>
<td>9.5 ± 0.25</td>
</tr>
<tr>
<td>+mono30</td>
<td>64.8 ± 4.4</td>
<td>57.8 ± 4.9</td>
<td>64.6 ± 4.9</td>
</tr>
<tr>
<td>+mono100</td>
<td>83.2 ± 3.1</td>
<td>71.5 ± 6.9</td>
<td>81.3 ± 1.6</td>
</tr>
<tr>
<td>+transductive</td>
<td>88.4 ± 0.7</td>
<td>81.6 ± 6.5</td>
<td>88.2 ± 2.2</td>
</tr>
</tbody>
</table>
Quality of Pseudo Parallel Data

- Iterative back-translation can increasingly correct errors in pseudo-parallel data
Impact of Error-Prone Data & Perturbations

- Even noise pseudo-parallel data can bring gains!
  - As they bring implicit knowledge of unseen combinations

- Perturbations brought by OTF (on-the-fly) is very important!
  - Pseudo-parallel data are generated dynamically, which prevent learning specific incorrect bias

![Graphs showing accuracy and BLEU scores for Source-to-Target (S2T) and Target-to-Source (T2S) models.](image-url)

(a) Accuracy of S2T models  
(b) BLEU of T2S models
Curriculum Iterative Back-Translation

• We want to help reduce errors more efficiently

• **CIBT**: during the training process:
  • start out with easy monolingual data,
  • then gradually increase the difficulty.
Curriculum Iterative Back-Translation

- Curriculum learning benefits iterative back-translation.
- Curriculum learning is more beneficial to difficult data than simple data.

Table 3: Performance (accuracy) of curriculum iterative back-translation.

<table>
<thead>
<tr>
<th></th>
<th>IBT</th>
<th>CIBT with hyperparameter c (steps in each stage)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2000</td>
<td>2500</td>
</tr>
<tr>
<td>MCD1</td>
<td>64.8 ± 4.4</td>
<td>66.1 ± 5.0</td>
</tr>
<tr>
<td>MCD2</td>
<td>57.8 ± 4.9</td>
<td>68.6 ± 2.6</td>
</tr>
<tr>
<td>MCD3</td>
<td>64.6 ± 4.9</td>
<td>70.2 ± 4.9</td>
</tr>
<tr>
<td>Mean</td>
<td>62.4 ± 6.1</td>
<td>68.3 ± 4.1</td>
</tr>
</tbody>
</table>

Figure 6: Performance on different subsets. This figure indicates that curriculum learning is more beneficial to difficult data (larger $k$) than simple data (smaller $k$).
Takeaways

• Iterative back-translation can significantly improve CG.

• Why IBT works well:
  • Unseen combinations
  • Increasingly improving the quality of pseudo-parallel data
  • Perturbations

• We propose curriculum iterative back-translation to further improving the performance.
THANKS

2020.12.19

Related papers from our team (MSRA DKI):

Hierarchical Poset Decoding for Compositional Generalization in Language (NeurIPS 2020)

Compositional Generalization by Learning Analytical Expressions (NeurIPS 2020 Spotlight)

Iterating Utterance Segmentation for Neural Semantic Parsing (AAAI 2021)