Empirical Analysis of Noising Scheme based Synthetic Data Generation for Automatic Post-editing

Hyeonseok Moon†, Chanjun Park, Seolwha Lee, Jaehyung Seo, Jungseob Lee, Sugyeong Eo, Heuiseok Lim*
Korea University, Republic of Korea

Overview

• Why Automatic Post Editing? (APE)
  - Inherent limitation of the machine translation system
  - Alleviating human level editing effort

Wrong Translation

순창은 고추장으로 유명하다.
Sunchang is famous for red pepper paste.

Human Revision

고추장 ➔ Kochujang

• Limitation of APE Research - Data Acquisition
  - High human resource is required in generating APE data (Especially, Post-Edited sentence)
  - No publicly released dataset for most of the language pairs
  - Existing dataset have small data size

Our Approach

- Noising scheme based data augmentation
  ➔ Does not require NMT model in data generation
- Mimicking human-like errors
  ➔ Considers practical human editing process

Convention

Noising Scheme ➔ Combining all types noise

Experimental Results

• Inspection on the Edit-Distance based noise

- Combining all lead to degradation
  - Substitution noise leads to better performance

• Noising Scheme Utilizing POS Tagging

- Best: SubPOS / Worst: ShiftPOS
  - Noise injection conserving the linguistic structure

• Semantic Noising Utilizing Wordnet

- Performance gap between Sub_syn and SubAnto
  - Semantics coherence in noise injection

• Combining Noising Schemes

- Shifting noise can act positively when semantical and structural coherence is maintained

<table>
<thead>
<tr>
<th>Noising Scheme</th>
<th>BLEU (↑)</th>
<th>TER (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>33.115</td>
<td>51.929</td>
</tr>
<tr>
<td>Dynamic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edit Distance Based</td>
<td>42.862</td>
<td>44.348</td>
</tr>
<tr>
<td>SubPOS</td>
<td>42.941</td>
<td>44.040</td>
</tr>
<tr>
<td>SubPOS + ShiftPOS</td>
<td>43.002</td>
<td>44.043</td>
</tr>
<tr>
<td>Static</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edit Distance Based</td>
<td>42.853</td>
<td>44.153</td>
</tr>
<tr>
<td>SubPOS</td>
<td>42.874</td>
<td>44.216</td>
</tr>
<tr>
<td>SubPOS + ShiftPOS</td>
<td>42.703</td>
<td>44.430</td>
</tr>
</tbody>
</table>

Generating new APE data for each iteration ➔ Training various aspect of the noising schemes.