A Benchmark Dataset for Multi-Level Complexity-Controllable Machine Translation

Kazuki Tani, Ryoya Yuasa, Kazuki Takikawa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, Tsuneo Kato (Doshisha University, Ehime University)

Introduction

- Style-controllable NMT has recently received much attention.
- Multi-Level Complexity-Controllable MT (MLCC-MT)
  - Controls the complexity of a target language sentence at three or more levels.
  - To allow translation tailored to the user’s reading level.

Problem: Existing test dataset cannot precisely measure model performance.

Objectives:
- Construct a test dataset to properly evaluate MLCC-MT models.
- Provide benchmark performance by evaluating two MLCC-MT models (i.e., pipeline and multi-task models).

Problems of Existing Test Dataset (Agrawal et al., EMNLP 2019)

- Es → En dataset automatically generated from Newsela corpus.
- Newsela corpus: a document-level comparable corpus with document-level complexity (i.e., grade level), composed of English and Spanish news articles.
- Creation procedure:
  - Step 1: Generates the sets of English sentences with the same content written at multiple complexity levels.
  - Step 2: Automatically removes exactly the same sentences or grade levels diff ≤ 1 (→ solves Issue 3).
  - Step 3: Manually removes the sets where new content appears (→ solves Issue 2).

Benchmark Test Dataset

- A new benchmark test dataset for Ja-To-En MLCC-MT.
- Proposed creation procedure:
  - Step 1: Generates the sets of English sentences with the same content written at multiple complexity levels.
  - Step 2: Automatically removes exactly the same sentences or grade levels diff ≤ 1 (→ solves Issue 3).
  - Step 3: Manually removes the sets where new content appears (→ solves Issue 2).

Issue 1: Incorrect translation pairs
- Difference of granularity of information among target language sentences with different complexity levels.

Examples of Issue 2 (proper noun insertion):

<table>
<thead>
<tr>
<th>Level</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>However, she says that there are times when “you just need to get away.”</td>
</tr>
<tr>
<td>2</td>
<td>Yet she says that there are times when “you just need to get away.”</td>
</tr>
<tr>
<td>3</td>
<td>Still, Bopp says that there are times when “you just need to get away.”</td>
</tr>
</tbody>
</table>

Issue 3: Incorrect sentence-level complexity
- Examples of Issue 3:

<table>
<thead>
<tr>
<th>Level</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>So few Indians drink brewed coffee that virtually all its best crop is exported to countries such as Italy, where the beans are used in name-brand espresso blends and sold at a huge markup.</td>
</tr>
<tr>
<td>9</td>
<td>There the beans are used in name-brand espresso blends and sold at a huge price increase.</td>
</tr>
<tr>
<td>10</td>
<td>There the beans are used in name-brand espresso blends and sold for a huge price increase.</td>
</tr>
</tbody>
</table>

Benchmark Experiments

- Implement two Transformer-based MLCC-MT models and evaluate them on our test dataset to serve as benchmark performance for future research.

Benchmark models:
- Pipeline model: Ja-to-En NMT → En multi-level simplification.
- Ja-to-En NMT: Transformer NMT (Kiyono et al., wmt2020).
- En multi-level simplification model: Incorporation of special tokens representing target complexity (Scarton et al., ACL 2018).
- Multi-task model: Based on the following three losses:
  - \( \text{loss} = L_{\text{MT}} + L_{\text{Simplification}} + L_{\text{FKGL}} \)
  - \( L_{\text{MT}} \): the loss for conventional MT
  - \( L_{\text{Simplification}} \): the loss for text simplification
  - \( L_{\text{FKGL}} \): the loss for complexity-controllable MT

Evaluation Metrics:
- SARI (Xu et al., TACL 2016): Text simplification performance.
- \( \text{MAE}_{\text{FKGL}} \) (Mean absolute error of FKGL) (Nishihara et al., ACL RW 2019): complexity controlling performance.

Results:

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU (%)</th>
<th>SARI (%)</th>
<th>( \text{MAE}_{\text{FKGL}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipeline</td>
<td>15.12</td>
<td>23.89</td>
<td>5.10</td>
</tr>
<tr>
<td>Multi-task</td>
<td>20.17</td>
<td>26.78</td>
<td>4.83</td>
</tr>
</tbody>
</table>

Conclusion

- Create a new benchmark test dataset for Ja-En MLCC-MT.
- The proposed creation procedure includes automatic filtering, manual check, and manual translation to make our test dataset more appropriate than existing test datasets.
- Implement two Transformer-based MLCC-MT (pipeline and multi-task) and evaluate them as benchmark performance.
- Future work:
  - Increase the size of our dataset and create a multi-lingual dataset for MLCC-MT.