Construction of a Quality Estimation Dataset for Automatic Evaluation of Japanese Grammatical Error Correction

Daisuke Suzuki ♡, Yujin Takahashi ♡, Ikumi Yamashita ♡, Taichi Aida ♡, Tosho Hirasawa ♡, Michitaka Nakatsuji ♡, Masato Mita ♡, ♡, Mamoru Komachi ♡

♡ Tokyo Metropolitan University; ♡ RIKEN

Introduction

• Quality estimation (QE) models can evaluate grammatical error correction (GEC) systems without relying on reference sentences
• Yoshimura+ (2020) showed English QE model has high correlation with human, but Japanese QE models are not available owing to the absence of datasets

Contribution

• Created a Japanese QE dataset for GEC via manual evaluation based on the NAIST Lang-8 Learner Corpora (Mizumoto+, 2011)
• Constructed a QE model using the dataset and demonstrated the usefulness of a QE model for Japanese GEC

Annotation

1. Get learner-sentences from Lang-8 derived corpora
   • TEC-JL (Koyama+, 2020): Evaluation dataset for Japanese GEC (2,042 sents)
   • FLUTEC (Kiyama+, 2022): Fluency-edited evaluation dataset (2,100 sents)
2. Use a variety of GEC systems to generate system outputs
   • SMT (moses), RNN (fairseq), CNN (fairseq), Transformer (fairseq)
3. Retain unique pairs of learner-sentences and system-outputs
4. Assign QE scores (0-4) to the pairs by three Japanese-speaking university students

Flow for the annotation

- Learner-sentence
- GEC systems
- Corrected sentences
- Human annotation

Histogram of our dataset

<table>
<thead>
<tr>
<th>Evaluation score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>[1.0, 2.0)</td>
<td>816</td>
</tr>
<tr>
<td>[2.0, 3.0)</td>
<td>554</td>
</tr>
<tr>
<td>[3.0, 4.0)</td>
<td>1426</td>
</tr>
<tr>
<td>4</td>
<td>1689</td>
</tr>
</tbody>
</table>

Average of evaluation scores

- TEC-JL
- FLUTEC

Experiments

• Built a BERT-based QE model using our dataset (4,391 sents)
• Baseline: GLEU (Napoles+, 2016)
• Compare Pearson’s and Spearman’s coefficients against human scores

Results of the QE model

<table>
<thead>
<tr>
<th>Method</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLEU</td>
<td>0.320</td>
<td>0.362</td>
</tr>
<tr>
<td>Fine-tuned BERT</td>
<td>0.580</td>
<td>0.413</td>
</tr>
</tbody>
</table>

Successful examples (S = src, C = correction, R = reference, E = evaluation)

Example 1

S ソンさんが好きがっているのは推理小説です。
(Son A wants to like reading mysteries.)
C ソンさんが好きになっているのは推理小説です。
(Son A is becoming fond of mystery novels.)
R ソンさんが好きでっているのは推理小説です。
(Sonna likes to read mysteries.)
E GLEU, BERT, Human / 0.59, 3.30, 3.00

https://github.com/tmu-nlp/jqe4gec/