Towards Minimal Supervision BERT-based Grammar Error Correction

Yiyuan Li, Antonios Anastasopoulos and Alan W Black

Carnegie Mellon University, Pittsburgh, PA
yiyuanli@andrew.cmu.edu {aanastas, awb}@cs.cmu.edu

Motivation

Utilizing grammatical information captured by unsupervised contextual model pre-trained on large corpora, like BERT [1] and extending to GEC in many languages with minimal supervision.

Dataset

- Corpus: The First Certificate in English (FCE), preprocessed to single error sentence-edit pairs (each edit and last edit).

Proposed Method

Two stages:

- Error Identification:
  - BERT to detect.
  - Mask placement.
- Mask Prediction:
  - Predict token at masked position.
  - Rerank candidates.

Result

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
</tr>
</thead>
</table>
| Grammatical Error Correction (GEC): Detect errors (misspelling, subject-verb agreement, determiner, etc.) in the sentences and correct them. | 0 1 0 1 0 0 0 0

Challenges

- Current grammatical error correction methods require large amount of annotated data, which may not be accessible in many languages.

Error Analysis

Common BERT prediction errors, with the original error and the prediction highlighted.

Example 23: Synonym

- Source: Of course there’s also a number 8 bus in front of the hotel, which is also suitable, but it leaves only every half an hour.
- Mask: Of course there’s also a number 8 bus [MASK] in front of the hotel, which is also suitable, but it leaves only every half an hour.
- Target: Of course there’s also a number 8 bus in front of the hotel, which is also suitable, but it leaves only every half an hour.
- Ours: Of course there’s also a number 8 bus [MASK] in front of the hotel, which is also suitable, but it leaves only every half an hour.

Future Work

- Error fertility: Accurate mask placement.
- Better span detection: To leverage redundant edits.

Table 1: Sentence-level evaluation.

<table>
<thead>
<tr>
<th>Masking Strategy</th>
<th>each edit</th>
<th>last edit</th>
</tr>
</thead>
<tbody>
<tr>
<td>P@1 R@1 F@1</td>
<td>P@1 R@1 F@1</td>
<td>P@1 R@1 F@1</td>
</tr>
<tr>
<td># origin</td>
<td>0.632</td>
<td>0.725</td>
</tr>
<tr>
<td># target</td>
<td>0.668</td>
<td>0.807</td>
</tr>
<tr>
<td>single</td>
<td>0.763</td>
<td>0.931</td>
</tr>
</tbody>
</table>

Table 2: Token-level evaluation.

<table>
<thead>
<tr>
<th>Masking Strategy</th>
<th>each edit</th>
<th>last edit</th>
</tr>
</thead>
<tbody>
<tr>
<td>P@1 R@1 F@1</td>
<td>P@1 R@1 F@1</td>
<td>P@1 R@1 F@1</td>
</tr>
<tr>
<td># origin</td>
<td>0.292</td>
<td>0.295</td>
</tr>
<tr>
<td># target</td>
<td>0.313</td>
<td>0.348</td>
</tr>
<tr>
<td>single</td>
<td>0.365</td>
<td>0.554</td>
</tr>
</tbody>
</table>

- Sentence level evaluated by ERRANT (Table 1).
- Token level evaluated by performance@5 and performance@1 (Table 2).
- Multilingual BERT achieves precision over 0.7 without fine-tuning.
- Further potential improvement by reranking.

Conclusion

- Pre-trained BERT can achieve more than 0.7 precision in single-error grammatical error correction without fine-tuning, and could be potentially improved by re-ranking.
- Advanced masking and fluency measure are needed to leverage information lost by masking and setting up ending criterion in iterative editing.

Table 3: Performance of fine-tuning BERT on FCE for grammatical error detection (GED).

<table>
<thead>
<tr>
<th>Pre-trained Model</th>
<th>F</th>
<th>R</th>
<th>F@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base-uncased</td>
<td>0.464</td>
<td>0.319</td>
<td>0.426</td>
</tr>
<tr>
<td>BERT-multilingual-uncased</td>
<td>0.664</td>
<td>0.574</td>
<td>0.656</td>
</tr>
<tr>
<td>Kaneko and Komachi 2019 [3]</td>
<td>0.298</td>
<td>0.312</td>
<td>0.320</td>
</tr>
</tbody>
</table>

Common BERT prediction errors, with the original error and the prediction highlighted.

Example 23: Synonym

- Source: Of course there’s also a number 8 bus in front of the hotel, which is also suitable, but it leaves only every half an hour.
- Mask: Of course there’s also a number 8 bus [MASK] in front of the hotel, which is also suitable, but it leaves only every half an hour.
- Target: Of course there’s also a number 8 bus in front of the hotel, which is also suitable, but it leaves only every half an hour.
- Ours: Of course there’s also a number 8 bus [MASK] in front of the hotel, which is also suitable, but it leaves only every half an hour.

Future Work

- Error fertility: Accurate mask placement.
- Better span detection: To leverage redundant edits.

Table 3: Performance of fine-tuning BERT on FCE for grammatical error detection (GED).

- Pre-trained BERT can achieve more than 0.7 precision in single-error grammatical error correction without fine-tuning, and could be potentially improved by re-ranking.
- Advanced masking and fluency measure are needed to leverage information lost by masking and setting up ending criterion in iterative editing.