Automatic Classification of Russian Learner Errors
Alla Rozovskaya
Queens College, CUNY

Standard Reference-Based Evaluation for Grammatical Error Correction (GEC)

<table>
<thead>
<tr>
<th>Source</th>
<th>The settings are very realistic and the actors had a great performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Gold (RG)</td>
<td>The settings were very realistic and the actors gave a great performance</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>The settings are very realistic and the actors had great performance</td>
</tr>
</tbody>
</table>

Gold edits: (1) realistic -> realistic, (2) had -> gave, (3) are -> were

System edits: (1) realistic -> realistic, (2) had -> great, (3) are -> were

Corrected edits: (1) realistic -> realistic

Precision: 1/2 = 0.5
Recall: 1/3 = 0.33

No information on how the system performs on specific error types.

Classifying the edits
- For the gold edits, the annotators can be asked to provide a linguistic category during annotation.

Gold edits:
- (1) realistic -> realistic
- (2) a -> 0
- (3) are -> were
- Spelling
- Determiner
- Verb tense

Classifying system edits
- State-of-the-art GEC systems are based on neural machine translation (NMT) architecture.
- Classifying system edits is not trivial since the systems are not restricted in the types of edits that can be made.

The problem
- System performance can only be evaluated overall on all edits.
- Automatic edit classification is necessary to perform type-based evaluation of system performance.
- Type-based evaluation
  - Can provide insight into further system development.
  - Is necessary in order to provide useful feedback to language learners, when a mistake is identified.
  - Allows for a standardization of multiple GEC datasets that may have been annotated with different error taxonomies.

Error classification tool for Russian
- Our approach is inspired by ERRANT.
  - We use POS and morphological information to classify edits.
- Adapted to the specific challenges of Russian.
- The tool is applied to classify edits in two Russian learner corpora.
  - Manual evaluation with human annotators reveals that the accuracy of the edit classification is 93%.
- The tool is applied to 2 GEC systems.
  - Type-based performance evaluation shows "easy" and "challenging" errors in RussianGEC.

Overview of the rules
- Morphological analyzer is applied to source and target token of an edit.

Sample edit:

CTOP$\_$(table, sg, nom).

The analyzer produces:
- base form
- list of morphological properties (e.g., number, gender, case, aspect, voice, person, tense)

Predicting error type
- When base form is the same on source and target token, error type is predicted based on the morphological property that has different values in the source and target token.

- When base forms are different
  - Predict spelling error if the source word is not in the dictionary.
  - Predict lexical error otherwise.

- See paper for more details on the rules.

Challenges specific to Russian
- Some Russian surface forms have multiple analyses.
  - E.g., sg., gen, can be confused with pl., nom.
- Such cases are problematic since depending on the chosen analysis a different mismatch in the grammatical category (case or number) will be identified.
- We handle these cases by predicting 2 mismatch categories.

Manual evaluation of automatic error categories

<table>
<thead>
<tr>
<th>Feature</th>
<th>RUL-REC</th>
<th>RUL-Text</th>
<th>RUL-Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Accept.</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>Bad Accept.</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
</tr>
</tbody>
</table>

Manual evaluation of the automatically assigned error categories by each reader and each database on a set of 100 edits, randomly selected.

Type-based evaluation

<table>
<thead>
<tr>
<th>Error type</th>
<th>CWS P</th>
<th>R F1</th>
<th>Transformer P</th>
<th>R F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling</td>
<td>66.2</td>
<td>53.3</td>
<td>63.3</td>
<td>35.3</td>
</tr>
<tr>
<td>Len. choice</td>
<td>86.3</td>
<td>91.3</td>
<td>92.7</td>
<td>100</td>
</tr>
<tr>
<td>Punct.</td>
<td>5.4</td>
<td>23.1</td>
<td>13.5</td>
<td>5.4</td>
</tr>
<tr>
<td>Prep.</td>
<td>5.0</td>
<td>0.0</td>
<td>0.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Nouns</td>
<td>49.2</td>
<td>64.6</td>
<td>57.0</td>
<td>35.3</td>
</tr>
<tr>
<td>Verbs</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Adj.</td>
<td>49.0</td>
<td>25.6</td>
<td>39.0</td>
<td>64.1</td>
</tr>
</tbody>
</table>

Type-based evaluation on the RULREC dataset.