Developing a Dataset of Overridden Information in Wikipedia
Masatoshi Tsuchiya
Yasutaka Yokoi
Toyohashi University of Technology

Example of overridden information

Example of overridden information

Two possible task settings

Existence of neutral sentence

Decompose the sentence into a tuple

Contradiction occurs when replacing the time information of the target sentence

Contradiction DOES NOT occur when replacing the time information of the target sentence

Definition of information override

Construction procedure of the dataset

Statistics of the dataset

Conclusion

Sentence level changes between two versions of Japanese Wikipedia were collected as annotation targets.
2. Remove changes that do not contain either reference to updated articles or time and date expressions.
3. Remove changes caused by minor editing
4. Remove changes from low-quality articles
5. Human annotation!

<table>
<thead>
<tr>
<th>Article Title</th>
<th>Target Sentence</th>
<th>Reference Sentence</th>
<th>Textual entailment</th>
<th>Information override</th>
</tr>
</thead>
<tbody>
<tr>
<td>A binary classification problem taking only a target sentence as an input.</td>
<td>A binary classification problem taking both a target sentence and a reference sentence as an input.</td>
<td>The latter setting is employed, although the former setting looks better, because the subtask to discover a suitable reference sentence is expected to make the whole task too difficult.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Two possible task settings

We propose a new task of detecting overridden information and formalize it as a binary classification problem. The operation status of Oshida Station described in sentence \( s_2 \) has been overridden by sentence \( s_1 \).

A binary classification problem taking both a target sentence and a reference sentence as an input.

We can suppose that the operation status of Oshida Station described in sentence \( s_2 \) has been overridden by sentence \( s_1 \).

The definition of information override is necessary.

We cannot suppose that the operation status of Oshida Station described in sentence \( s_2 \) has been overridden by sentence \( s_1 \).

We can suppose that the operation status of Oshida Station described in sentence \( s_2 \) has been overridden by sentence \( s_1 \).

The definition of information override is necessary.

Conclude

We propose a new task of detecting overridden information and formalize it as a binary classification problem to determine whether a reference sentence has overridden a target sentence.

We offer a formal definition of information override between two sentences while relating it to textual entailment.

We propose a procedure to construct a dataset of overridden information by collecting sentence pairs from the difference between two versions of Wikipedia.