Evaluation Dataset and System for Japanese Lexical Simplification

Motivation
- Extensive / various forms of texts
- Hitler committed terrible atrocities during the second World War.
- Hitler committed terrible cruelties during the second World War.

Children  Language Learners  Elderlies

Task: Lexical Simplification
- Substitutes a complex word or phrase in a sentence with a simpler synonym
- Supports the reading comprehension of a wide range of readers

Problems in Japanese

Unpublished system
- It is difficult for people who need reading assistance to obtain simple Japanese sentences

Unpublished dataset
- It is difficult for researchers and developers to evaluate the performance of different systems

Lexical Simplification System
- Input: 未来は若者が担う Young people bear the future
- Identification of Complex Words: 担う bear
- Substitution Generation: 担う: 支える, 引き継ぐ, 受け継ぐ, 伝承する
  - bear: hold, wear, carry, expect
- Word Sense Disambiguation: 担う: 支える, 受け継ぐ
  - bear: hold, carry
- Synonym Ranking: 1: 支える, 2: 受け継ぐ, 3: 担う
  - 1: hold, 2: carry, 3: bear
- Output: 未来は若者が支える Young people hold the future

- Analyzed by the Japanese morphological analyzer MeCab
- Out of 5,404 simple words (for elementary student)
- Selected from 4 Japanese lexical paraphrasing databases
- Analyzed by the Japanese predicate-argument structure analyzer SynCha
- The case frame dictionary has 4,413,089 proper pairs of (predicate, subject / object)
- Improper words are removed
- Lexical Properties of Japanese (77,328 words & its familiarity)
- The most simple synonym is used as a replacement for the original word

Evaluation Dataset

1. Constructing Japanese Lexical Substitution Dataset
  - Collecting Substitutions (crowdsourcing: 5 workers, 17.8%)
  - Evaluating Substitutions (crowdsourcing: 5 workers, 66.4%)

2. Transforming it into Lexical Simplification Dataset
  - Ranking Substitutions (crowdsourcing: 5 workers, 33.2%)
  - Merging All Rankings

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sentence</th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
<th>Adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemEval 2012 Task1</td>
<td>2,010</td>
<td>580 (28.9%)</td>
<td>520 (25.9%)</td>
<td>560 (27.9%)</td>
<td>350 (17.4%)</td>
</tr>
<tr>
<td>Ours</td>
<td>2,330</td>
<td>630 (27.0%)</td>
<td>720 (30.9%)</td>
<td>500 (21.5%)</td>
<td>480 (20.6%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Properties of the dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>The average number of substitutions</td>
</tr>
<tr>
<td>The average number of levels of difficulty</td>
</tr>
<tr>
<td>There are synonyms that are more simple</td>
</tr>
<tr>
<td>There are synonyms that are more complex</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context dependency ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>①: context pairs</td>
</tr>
<tr>
<td>②: ① with same list</td>
</tr>
<tr>
<td>③: ② with different rankings</td>
</tr>
<tr>
<td>④: ③ with different top word</td>
</tr>
</tbody>
</table>

Sample: Young people bear the future.
- Lexical Substitutions: carry, hold
- Rank of Simple Level: 1:hold, 2:carry, 3:bear

The gold-standard annotations were generated by averaging the annotations from all annotators.

Tomoyuki Kajiwara and Kazuhide Yamamoto
Nagaoka University of Technology, Japan

kajiwara@jnlp.org
http://www.jnlp.org/SNOW