Evaluation Dataset and System for Japanese Lexical Simplification

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

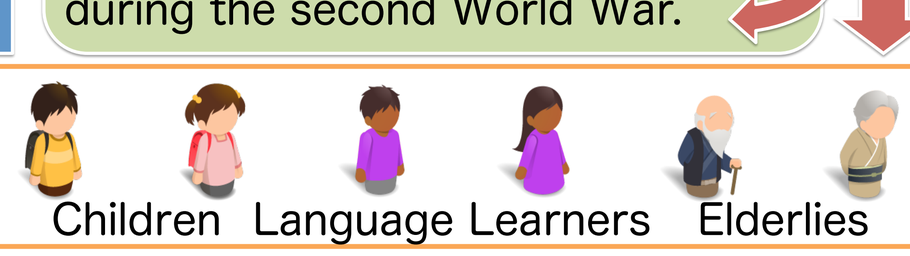
Tomoyuki Kajiwara and Kazuhide Yamamoto
Nagaoka University of Technology, Japan

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➢ Unpublished system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➢ Unpublished dataset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

- This dataset consists of 2,330 sentences, 233 target words each with 10 sentences as contexts
- These contexts were randomly selected from Japanese newspaper articles

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

Dataset	Sentence	Noun	Verb	Adjective	Adverb
SemEval 2012 Task1	2,010	580 (28.9%)	520 (25.9%)	560 (27.9%)	350 (17.4%)
Ours	2,330	630 (27.0%)	720 (30.9%)	500 (21.5%)	480 (20.6%)

System	Precision	Recall	F-measure
Our Original	0.89	0.08	0.15
w/o WSD	0.84	0.71	0.77

Properties of the dataset	
The average number of substitutions	4.50
The average number of levels of difficulty	4.94
There are synonyms that are more simpler	69.4 %
There are synonyms that are more complex	83.5 %

Context dependency ratio		
①: context pairs	10,485	100 %
②: ① with same list	1,593	15 %
③: ② with different rankings	948	60 %
④: ③ with different top word	463	49 %

ACL Student Research Workshop 2015, Beijing, China, July 28th, 2015.