Controlled and Balanced Dataset for Japanese Lexical Simplification

Tomonori Kodaira, Tomoyuki Kajiwara, and Mamoru Komachi
Tokyo Metropolitan University, Japan. https://github.com/KodairaTomonori/EvaluationDataset

Dataset for Japanese Lexical Simplification

We propose a new method of constructing a dataset for Japanese lexical simplification.
Each sentence in our dataset includes only one difficult word.
It is the first controlled and balanced dataset for Japanese lexical simplification with high correlation with human judgment.

Lexical Simplification

Substitutes a complex word or phrase in sentence with simpler synonym

He is a sincere man.

He is a honest man.

Example of dataset

sentence	もっとも 安上がり に サーファ を 装う 方法 The most simplest method that is **imitating** safer.

simp ranking	の ふり を する 1. professing	に 見せ かける 2.counterfeiting	の 真似 を する, の 振り を する 3.playing, professing	を 真似る 4.playing	を 装う 5.imitating

Example of annotation

1. Given Sentence

はるかに変化に富む (Far more varied)

2. Extracting Substitutes

が多い(numerous), が豊富(rich), が多い(wealthy)

3. Evaluating substitutes

が多い(numerous), が豊富(rich)

4. Ranking substitutes

substitutes	rank
に富む (varied)	2
が豊富 (wealthy)	3
が多い (numerous)	1

Flow of Constructing Dataset

1.Extracting Sentences

2,100 sentences from Balanced Corpus of Contemporary Written Japanese (Maekawa et al., 2010):
- Including only one complex word.
- 10 contexts of occurrence were collected for each complex words.

Difficult words:
- "High Level" words in the Lexicon for Japanese Language Education (Sunakawa et al., 2012).

Content words (7 parts of speech):
- nouns, verbs, adjectives, adverbs, adjectival nouns, *sahen* nouns, and *sahen* verbs.

Annotation Using Crowdsourcing

2. Extracting Substitutes

For each complex word, five annotators wrote substitutes that didn't change the sense of the sentence.
※ These substitutes could include particles in context. (Kajiwara and Yamamoto (2015) didn't include them.)

3. Evaluating Substitutes

Five annotators selected an appropriate word to include as a substitution that didn't change the sense of the sentence.
Substitutes that won a majority were defined as correct.

4. Ranking Substitutes

Five annotators arranged substitutes and complex word according to the simplification ranking.
These ranking could include ties.

Inter annotator agreement

	Spearman's score
K&Y	0.332
Our work	0.522

5. Ranking Integration

To remove extraordinary annotators,we used Maximum Likelihood Estimation method (Matsui et al., 2012).
Integrate rankings were made by average score after removing extraordinary annotators.

Correlation of ranking

	baseline	outlier removal
Spearman's score	0.541	0.580

Comparison of Datasets

	B&M (2012)	Specia et al. (2012)	K&Y (2015)	This work
# of sentences	430	2,010	2,330	**2,010**
lang	En	En	Ja	**Ja**
balanced dataset	Yes	Yes	No	**Yes**
complex word	multi	multi	multi	**one**
ties allowed	Yes	No	No	**Yes**
outlier excluded	Yes	No	No	**Yes**

B&M: Belder and Moens (2015), K&Y:Kajiwara and Yamamoto (2015)

Conclusion

(1) Our dataset is more consistent than the previous datasets.
(2) Lexical simplification methods using our dataset correlate with human annotation better than the previous datasets.

Future work includes increasing the number of sentences, so as to leverage the dataset for machine learning-based simplification methods.