Japanese Lexical Simplification for Non-Native Speakers

Muhaimin Hading, **Yuji Matsumoto,**
Graduate School of Information Science
Nara Institute of Science and Technology
{muhaimin.hading.mc, matsu}@is.naist.jp

Maki Sakamoto
Graduate School of Informatics and Engineering
The University of Electro-Communications
maki.sakamoto@uec.ac.jp

Abstract

This paper introduces Japanese lexical simplification. Japanese lexical simplification is the task of replacing complex words in a given sentence with simple words to produce a new sentence without changing the original meaning of the sentence. We propose a method of supervised regression learning to estimate complexity ordering of words with statistical features obtained from two types of Japanese corpora. For the similarity of words, we use a Japanese thesaurus and dependency-based word embeddings. Evaluation of the proposed method is performed by comparing the complexity ordering of the words.

1 Introduction

According to the statistical data collected by Ministry of Justice, there are about two million foreigners living in Japan today. Around half of them do not have Japanese proficiency. It has been a problem for foreigners since most of information is provided in Japanese. Lexical simplification is a process to make a sentence more readable for non-native speakers by replacing complex words to simpler ones that retain the same meaning. A number of studies on lexical simplification have been conducted in recent years. Since the task of lexical simplifications is related with development of other tasks in natural language processing such as words similarity or paraphrasing words. Most of studies in simplification are conducted in English. Since there are much larger data and many tools available for English compared with other languages. In this paper, we tackle the sentence simplification problem in Japanese. We propose a method for estimating complexity of Japanese words and for obtaining semantically similar words for the replacement of complex words.

2 Related Work

Dominant approaches of previous work in simplification are hybrid approaches, which combine deep semantic and monolingual machine translation (Narayan and Gardent, 2014), word alignment approach (Coster and Kauchak, 2011; Paetzold and Specia 2013; Horn, et al., 2014) or language modeling approach (Kauchak, 2013). The main limitation of these methods is that they depend on parallel corpus between simple and complex sentences such as English Wikipedia, Simple English Wikipedia or Newsela corpus (Xu et al., 2015).

Another approach is to implement the simplification task of complex words by replacing them with the extract synonym words obtained from databases such as thesaurus (Devlin and Tait, 1998; De Belder and Moens, 2010), from dictionary definition or from WordNet (Kajiwara et al.,2013). Thesauri can provide good synonyms while their coverage is limited. Recent work (Glavaš and Stajner, 2015; Paetzold and Specia, 2016) approached the word embedding model (Mikolov et al., 2013) to estimate word similarity and aims to mitigate the limitation of thesauri and parallel corpora.

The next approach is to identify complex words and choose simpler words for replacing them, while keeping the same meaning of the original complex words. Identifying complex words is

Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications, pages 92–96, Osaka, Japan, December 12 2016.

common work before doing the simplification task (Carroll et al., 1998; Baustista et al., 2009; Paetzold and Specia, 2016). Estimation of word complexity is mostly based on their frequencies (Devlin and Tait, 1998; De Belder and Moens, 2010; Kauchak et al., 2014, Kajiwara et al., 2015), their length (Bautista et al.,2009), judgment by user study (Paetzold and Specia, 2016), technical words in specific domains (Kauchak et al., 2014), basic vocabularies for children (Kajiwara et. al, 2013), Japanese Language Proficiency Test levels, or Easy Japanese corpus (Moku et al.,2012). The recent work of (Kodaira et al., 2016) used crowdsourcing to for collecting simplification candidates of words.

Our approach uses more than one resource. We first use a thesaurus since thesauri produce the best candidates of synonyms. For addressing the limitation of a thesaurus, we utilize dependency-based word embedding (Levy and Goldberg, 2014) since it is shown that dependency-based embedding highlights less topical and more functional similarity than the skip-gram models. For identification of complexity of words, we use existing approaches like frequency, words used by children, technical words, and Japanese Language Proficiency Test levels.

3 Data

In our experiments, we use Japanese raw corpora. We combine Balanced Corpus of Contemporary Written Japanese (BCCWJ)[1] and Mainichi Newspaper Corpus to estimate word similarity, Japanese Language Proficiency Test vocabularies list, a corpus of compositions written by Japanese Elementary and Junior High School students (Miyagi, 2015), a corpus of compositions written by Japanese Elementary School Children (Sakamoto, 2010), and Bunrui Goi Hyo Database Japanese thesaurus[2].

4 Proposed Method

4.1 Grouping Similar Words

The purpose of this task is to find groups of words that have similar meaning. In our experiment we use Bunrui-Goi-Hyo (BGH for short), a Japanese thesaurus. This thesaurus is manually constructed. It comprises about 100K words. In BGH, all the words are arranged by their meaning. We extract all the groups of words at the bottom level as similar words. Since the number of words are limited, some words do not appear in BGH. As we mentioned in Section 2, we propose to use dependency-based embedding approach (Levy and Goldberg, 2014) to improve the grouping of words especially for those words that do not appear in BGH.

4.2 Level of Word Complexity

The purpose of this task is to predict the complexity level of words based on Japanese Language Proficiency Test (JLPT) level (Hmeljak. 2009). JLPT is the standard test of Japanese for foreign learners and classifies words into five levels: N1 is the most advanced level, N2 is the high level, N3 is the intermediate level, N4 is the lower level, N5 is the beginner level. In the JLPT list, there are approximately 800 N5 level words, 1,500 N2 level words, 3,750 N3 level words, 6,000 N2 level words, and 10,000 N1 level words.

Vocabulary lists always have limited coverage of lexical entries. To cope with this limitation, we use a machine-learning approach to predict the complexity levels of words that are not included in the JLPT vocabulary list. We choose a linear regression model in this task. We examined several features for the linear regression model to predict the JLPT level of a given word. The features we use are:

1. Unigram frequency: Most of research in simplification uses frequency of words to determine the complexity of words. Sentences that are simple to understand mostly use high frequency words or well-known words (Kauchack et al., 2014; Glavasˇand Stajner, 2015).
2. Words in children's corpora: Children basically use simple words, kanjis, and expressions. Words used by children are considered as easy words.
3. Technical Words: Technical words are in many cases complex words and commonly used in specific domains. To measure specificity, we use Jensen–Shannon divergence of words over domains. The frequency distribution of a target word over domains is compared with the

[1] http://pj.ninjal.ac.jp/corpus_center/bccwj/en/
[2] https://www.ninjal.ac.jp/archives/goihyo/

average distribution of all words over domains. Those words that have low scores have similar distributions over domains, meaning they appear various domains and are considered as general words. In contrast, those that have high scores are considered as technical and complex words because they tend to appear in specific domains.

$$SD(P||Q) = \frac{\sum P \log \frac{P}{Q} + \sum Q log \frac{Q}{P}}{2}$$

In this formula of Jensen-Shannon divergence, P is the normalized frequency distribution of target word over all domains and Q is averaged normalized distribution of all words in the corpus. We used the genres defined in the BCCWJ corpus as different domains.

All these features can be the measurement of word complexity levels and we use them in our linear regression model.

5 Experiment

We divide our experiment into four steps. We start with pre-experiment, words similarity, complex words identification, and word replacement.

5.1 Pre-Experiment

This section describes preparation of the data. We use the Japanese morphological analyser MeCab to word segmentation and POS tagging of Mainichi Newspaper, BCCWJ, and the children's corpora. Since some words do not need to be the target of simplification, we make rules based on POS tags of words. In the following rules, we changed all words sharing the same POS as one word:

1. All words with POS '記号' (symbol) such as "、,。,「,」" are changed to 'Symbol'.
2. All words with POS '数' (number) such as "十(ten), 四(four), ５" are changed to 'Number'.
3. All words with POS '人名' (people name) such as "山田(Yamada) are changed 'People'.
4. All words with POS '組織'(organization) such as 東芝(Toshiba) are change to 'Organization'.
5. All words with POS '地域'(Region) such as "奈良(Nara) are change to 'Place'.
6. All names of day, month and year such as "月曜日(Monday), ９月(September), 九月(September), ２０１６年(2016), ２８平成(2016)" are change to 'Date'.

All of those words are not simplified. The result of this step is used in other steps.

5.2 Words Similarity

Using the BCCWJ and Mainishi corpora, we calculate the similarity of words by using the available tools of dependency-based word embeddings (Levy and Goldberg, 2014). We prepare the training data using Japanese dependency parser CaboCha for finding the dependency relation of words in sentences. Then we calculate the similarity of words as we showed in Section 4.1. This task is to augment the groups of similar words that do not appear in BGH. The following is an example of grouped words by the dependency-based word embedding: {処分, 認定, 申請, 給付, 承認, 決議, 届出, 登記, 規制, 譲渡}

5.3 Word Complexity Order

We counted all unigram frequency of words in Children's corpus, and combination of BCCWJ-Mainichi Newspaper. Then we divided the BCCWJ and Mainichi corpora into 19 categories. Then we calculate the frequency of each word in each category to know the distribution of the word over the categories. Based on the distribution of words over the categories, we calculated the Jensen-Shannon divergence values of words. This task is to know the technicality of words.

About 10,000 words in the JLPT vocabulary list are already divided into 5 levels. For each word in JLPT level, we calculated its log frequency in the BCCWJ-Mainichi corpus, log frequency in the children's corpus, and J-S divergence value, and use them for the features of linear regression.

After training, we apply the regression function to the other words that are not in the JLPT list to predict their complexity levels. Since the level of the easiest words is 5 and that of the most complex words is 1, higher values on test data indicate easier and lower value indicate more complex words.

5.4 Word Replacement

We already have groups of similar words (Section 5.2) and complexity ordering of words (Section 5.3). In this section, we combine these results. In order to replace a complexity word with a simpler synonym, we first start an experiment with words of POS 名詞 (Noun) tag.

When we input a sentence, first thing to do is morphological analysis of the sentence using MeCab, then select all words with POS 名詞 (Noun). For each selected word, we check it in the same group of similar words and compare the complexity of the selected word with other candidates, then choose the one with the highest complexity value. Table 1 shows the result of replacement of complex words with simpler synonyms so as to make it more readable for foreigners. From Table 1, we see that '人情本'(Novel) is a complex word that has similar meaning with 小説 (Novel). While 人情本 has meaning of old novels that were written in the Edo era, learners do not need to know that kind of words, the important information is that that word means a kind of novels.

Original Sentence	Simplified sentences
私は**人情本**より詩の方が好きです	私は**小説**より詩の方が好きです
芸の**秘奥**をきわめる	芸の**秘伝**をきわめる
代数は僕の得意な学科だ	**数学**は僕の得意な学科だ
ご**尊名**はよく承っております	ご**名前**はよく承っております

Table 1 : Result of word replacement

6 Evaluation

We evaluate word complexity orders provided by the trained liner regression model. We use another JLPT data collected from JLPT books, summing up to 20,000 words. We divide this data into training and test data. Then we test how the complexity levels of pairwise test words are correctly predicted. We experimented the comparison as shown in Table 2. From the table, we can see that in one different level, the average of accuracy is about 61.89%, that in two different levels is 72.13%, that in three different levels is 79.8%, and that in four different levels is 87%. We did not do the evaluation of group of similarity in BGH since it is a thesaurus constructed by human.

Categories	Compared Levels	Accuracy	Average
One different level	N1 and N2	61.14%	61.89 %
	N2 and N3	61.61%	
	N3 and N4	58.92%	
	N4 and N5	65.90%	
Two different level	N1 and N3	70.90%	72.13 %
	N2 and N4	69.23%	
	N3 and N5	74.28%	
Three different level	N1 and N4	76.91%	79. 8 %
	N2 and N5	82.69	
Four different level	N1 and N5	87.81%	87.81%

Table 2: Results of word difficulty level comparison

7 Conclusion and Future Work

We proposed an approach for Japanese lexical simplification. Our main task is divided into two parts. The first is word similarity estimation and the second is word complexity ordering. We used combination of several Japanese corpora to implement word embedding and linear regression models. Our experiments showed the order of word complexity is usable to select simpler similar words. Because of limited space, we did not discuss problems caused by lexical replacement. The last row in Table 1 shows a problem caused by word replacement from "尊名" to "名前". The prefix existing in the original sentence "ご" fits with the former word but not with the latter word. This type of problem caused by word combination is one of the important problem to be tackled by our future work.

Reference

Colby Horn, Cathryn Manduca and David Kauchak. 2014. Learning a Lexical Simplifier Using Wikipedia. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 458–463.

David Kauchak. 2013. Improving Text Simplification Language Modeling Using Unsimplified Text Data. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1537–154.

David Kauchak, Obay Mouradi, Christopher Pentoney, Gondy Leroy, PhD. 2014. Text Simplification Tools: Using Machine Learning to Discover Features that Identify Difficult Text. *47th Hawaii International Conference on System Science*, pages 2616-2625.

Goran Glavaš, Sanja Stajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora?. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 63–68.

Gustavo H. Paetzold and Lucia Specia. 2016. Unsupervised Lexical Simplification for Non-Native Speakers. *In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3761-3767.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. *In Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. *In Proceedings of the SIGIR Workshop on Accessible Search Systems*, pages 19–26.

Kristina Hmeljak Sangawa. 2009. A corpus for Readability Measurement for Non-Native Learners of Japanese. *IEICE Technical Report*, pages 19-23

Manami Moku, Kazuhide Yamamoto, Ai Makabi. 2012. Automatic Easy Japanese Translation for Information accessibility of foreigners. *In proceedings of the workshop on speech and language processing tools in education*, pages 85-90.

Maki Sakamoto. 2010. Corpus of Texts Composed by Japanese Elementary School Children and its Application in Linguistics and Sociology. *Journal of Natural Language Processing Vol. 17 No. 5*, pages 75-98 (In Japanese).

Omer Levy and Yoav Golfberg. 2014. Dependecy-Based Word Embeddings. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 302–308

Shin Miyagi and Mizuho Imada. 2015. Design of a Written Composition Corpus of Japanese Elementary and Junior High School Students. *第7回コーパス日本語学ワークショップ予稿集*, pages 223-232 (In Japanese).

Shashi Narayan and Claire Gardent. 2014. Hybrid Simplification using Deep Semantics and Machine Translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 435–445.

Susana Bautista, Pablo Gervas, and R. Ignacio Madrid. 2009. Feasibility analysis for semi-automatic conversion of text to improve readability. *In Proceedings of the Second International Conference on Information and Communication Technology and Accessibility (ICTA)*, pages 33–40.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

Tomoyuki Kajiwara, Hiroshi Matsumoto, Kazuhide Yamamoto. 2013 Selecting Proper Lexical Paraphrase for Children. *In Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING)*, pages 59-73.

Tomoyuki Kajiwara and Kazuhide Yamamoto. 2015. Evaluation Dataset and System for Japanese Lexical Simplification. *In Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *In Advances in Neural Information Processing Systems 26 .Curran Associates, Inc*. pages 3111–3119.

Tomonori Kodaira, Tomoyuki Kajiwara, Mamoru Komachi. 2016. Controlled and Balanced Dataset for Japanese Lexical Simplification. In *Proceedings of the ACL 2016 Student Research Workshop*, pp.1-7.

Wei Xu, Chris Callison-Burch, Courtney Napoles. 2015. Problems in Current Text Simplification Research : New Data Can Help. *Transactions of the Association for Computational Linguistics, vol. 3*, pp. 283–297

William Coster and David Kauchak. 2011. Learning to simplify sentences using wikipedia. In Proceedings of the Workshop on Monolingual Text-To-Text Generation, pages 1–9.