Japanese grammatical simplification with simplified corpus

Yumeto Inaoka and Kazuhide Yamamoto
Nagaoka University of Technology
Nagaoka, Japan
{*inaoka, yamamoto*}*@jnlp.org*

***Abstract*—We construct a Japanese grammatical simplification corpus and established automatic simplification methods. We compare the conventional machine translation approach, our proposed method, and a hybrid method by automatic and manual evaluation. The results of the automatic evaluation show that the proposed method exhibits a lower score than the machine translation approach; however, the hybrid method garners the highest score. According to those results, the machine translation approach and proposed method present different sentences that can be simplified, while the hybrid version is effective in grammatical simplification.**

***Keywords*-text simplification, paraphrasing, controlled languages**

I. Introduction

The number of foreign residents in Japan has reached approximately 2.64 million, increasing over the past six years[1]. Recently, in Japan, guide plates and official documents have often been written in both Japanese and English. However, 56% of foreigners in Japan do not understand English [1], which is more than the foreigners who do not understand Japanese (37%). Therefore, Japanese that is easy for foreigners to understand (Easy Japanese) is presented in the field of humanities[2].

Meanwhile, automatic simplification is studied in the field of natural language processing. This is a task that automatically converts text with complex vocabulary and grammar into simple text. [2] This technology is useful for foreign language learners, children, the elderly, and disabled people. Text simplification can be regarded as a subtask into various language processing applications because the input and output are the same language. It is known to improve performance by using text simplification as a subtask of syntactic parsers [3] and machine translation [4].

First, we constructed a Japanese grammatical simplification corpus by using crowdsourcing. Crowdsourcing is a method of widely recruiting workers using the Internet and outsourcing the work. With this method, a large amount of work can be completed in a short time and inexpensively. For instance, this method is used in the construction of language resources and manual evaluation of machine translation [5], [6].

Second, we compared simplification methods by using the constructed corpus. In conventional text simplification, the machine translation approach is used. Because the input and output languages are the same, it is not necessary to generate an entire sentence from scratch. Therefore, we compared the following three methods.

- Machine translation approach that is often used in conventional research
- Our proposed method by extracting differences in sentence pairs
- Hybrid method combining the above two methods

II. Related work

Text simplification research is mainly written in English. There is also some research in Japanese that we target. Simple PPDB [7] is a dictionary for simplification, constructed by collecting simplified paraphrases from a large-scale paraphrase dictionary (PPDB) [8] in English, and a Japanese version(Simple PPDB: Japanese [9]) is available as well. Paraphrasing to simplify vocabulary, as in those dictionaries, is called lexical simplification. In sentence-level simplification, a machine learning method using a parallel corpus is applied in which pairs of difficult sentences and plain sentences are constructed. Statistical machine translation (SMT) [10] and neural machine translation (NMT) [11] are used to train parallel corpora. In English sentence simplification, parallel corpora are often constructed by sentence pairs extracted using English Wikipedia[3] and Simple English Wikipedia[4] [12]. In Japanese research, construction of a parallel corpus by hand, construction using crowdsourcing, and simplification by machine translation method using a parallel corpus have been studied [13]–[15].

English used in Simple English Wikipedia is constrained by simple grammar and vocabulary. Meanwhile, Japanese simplified corpora do not consider simple grammar. Therefore, they cannot be used as a parallel corpus for grammatical simplification.

We extracted paraphrases for simplification from parallel corpus. The extraction of paraphrases from parallel corpus by using the alignment method [16] is proposed [17]. We used the method for grammatical simplification by using a manually constructed parallel corpus.

[1] https://www.e-stat.go.jp/

[2] http://human.cc.hirosaki-u.ac.jp/kokugo/tagengoenglish.html

[3] https://en.wikipedia.org/

[4] https://simple.wikipedia.org

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

Table I
SENTENCES THAT FOLLOW MINIMUM GRAMMAR AND THOSE THAT DO NOT

	Sentence that does NOT follow	Sentence that follows	English translation
1	彼 は 元気である 。	彼 は 元気です 。	He is fine.
2	その 動物 は 滅ぼさ れた 。	その 動物 は 滅び ました 。	The animal has been destroyed.
3	お 好きな だけ 取って ください 。	好きな 分 を 取って ください 。	Take as much as you like.
4	知らない です 。	知り ませ ん 。	I don't know.
5	彼女 は 必死 に なって 走り ました 。	彼女 は 必死 に 走り ました 。	She ran desperately.
6	私 は 次 に どう す べき か 分かりません 。	私 は 次 に どう する 方 が いい か 分かりません 。	I do not know what to do next.

Table II
DETAILS OF THE MACHINE TRANSLATION MODEL

Parameter	Value
Architecture	Encoder-Decoder with Attention
Encoder	Bi-directional LSTM
Decoder	LSTM
Number of RNN layers	2
Hidden size	256
Word embedding size	256
Dropout	0.4
Optimizer	Stochastic gradient descent
Learning rate	1.0
Others	Shared embedding layers Copy-mechanism Replace UNK[5]

III. GRAMMATICALLY SIMPLIFIED CORPUS

A. Easy Japanese grammar checker

One of the examples of simple grammar in Japanese is called "minimum grammar" [18]. It defines the minimum grammar necessary to express one's own thoughts in Japanese. We constructed a parallel corpus by paraphrasing sentences to satisfy the grammatical constraints. We created a checker that automatically checks whether a sentence follows the grammatical constraints. Input sentences are analyzed using MeCab[6], the JUMAN dictionary, and CaboCha[7]. MeCab is an engine used for Japanese morphological analysis. The JUMAN dictionary is a dictionary used in morphological analysis by MeCab. CaboCha is a Japanese dependency parser; however, we used it to chunk sentences. The checker checks in chunk units without considering dependencies. Consequently, it has not been possible to implement all of the grammatical constraints. In addition, although the nominal verb is not included in "minimum grammar," we included it because, otherwise, it would be difficult to paraphrase through crowdsourcing. The paraphrasing of a nominal verb will be an issue for the future. Table I shows examples of sentences that follow minimum grammar and those that do not.

B. Construction of a grammatically simplified corpus

We constructed a grammatically simplified corpus for research into Japanese grammatical simplification. The corpus is constructed by paraphrasing complex sentences to be simple. We used "small_parallel_enja: 50k En/Ja Parallel Corpus for Testing SMT Methods[8]" as the original text to paraphrase. It is a Japanese-English bilingual corpus constructed by filtering the Tanaka corpus[9]. The workers to paraphrase were recruited using CrowdWorks[10] (Japanese Crowdsourcing service). We assigned 5,000 sentences per person to 13 workers and collected 64,738 paraphrasing pairs. However, 100 of the sentences were given to every worker. Therefore, 100 sentences were paraphrased by all 13 workers. The workers used an easy Japanese grammar checker to paraphrase. Specifically, the sentence is paraphrased by repeating paraphrasing, such that a chunk classified by the checker as not easy is determined to be easy.

IV. METHODS

A. Machine translation approach

In the case of simplification, the machine translation approach is often used, considering before and after simplification as different languages. We trained a neural machine translation model using the parallel corpus explained above and constructed the simplification model. The details of the machine translation model used are shown in Table II.

B. Proposed method

Because text simplification outputs sentences with the same language and meaning as the input, there are many common parts between the input and output. Therefore, there is no need to generate an output sentence from scratch. NMT inputs and outputs entire sentences, but those do not require simplification. Thus, we propose a method to extract and apply the difference in sentence pairs for simplification.

1) Extraction of paraphrases: First, a sentence pair is considered before and after simplification. Next, each sentence is separated into chunks with CaboCha. A sentence is expressed by a sequence of chunks, which represent sequences of words. The edit distance from each sentence-pair sequence is calculated. Then, according to the edit distance, the substitution is extracted among the editing

[6] https://taku910.github.io/mecab/
[7] https://taku910.github.io/cabocha/

[8] https://github.com/odashi/small_parallel_enja
[9] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus
[10] https://crowdworks.jp/
[10] Replace the generated UNK tokens with the source token that had highest attention weight.

Table III
EXTRACTED PARAPHRASES FROM TABLE I

	Source	Target	English Translation
1	元気である 。	元気です 。	fine.
2	滅ぼされた 。	滅びました 。	has been destroyed.
3	お 好きな だけ	好きな 分 を	as much as you like
4	知らない です 。	知り ません 。	do not know.
5	必死 に なって	必死 に	desperately
6	どう す べき か	どう する 方 が いい か	What shoud I do?

Table IV
EXAMPLE THAT CANNOT BE PARAPHRASED WITHOUT GENERALIZATION

	Surface	English Translation
Chunk1	食べた	ate (plain form)
Chunk2	食べました	ate (polite form)
Chunk1’	飲んだ	drank (plain form)
Chunk2’	飲みました	drank (polite form)

operations (substitution, insertion, and deletion) between the two sequences (implemented using Python difflib[11]). In this manner, paraphrases between sentences can be extracted. The extracted paraphrases are not necessarily a substitution of one chunk for another chunk. Multiple chunks may be substituted by one chunk or one chunk may be substituted by multiple chunks. The paraphrases extracted from the corpus are shown in Table III.

2) Generalization of paraphrase by conjugation: The applicable range of paraphrases extracted using only surface of words is narrow. The example that cannot be paraphrased without generalization is shown in Table IV. Even if it is possible to extract a paraphrase from Chunk1 to Chunk2, it is not possible to paraphrase from Chunk1’ to Chunk2’. This means that the applicable range of the extracted paraphrase is narrow. Such examples also appear elsewhere. Therefore, we generalize the verb and adjective words to the form of “part-of-speech + conjugation” without using surface. Thus, even if only one of these paraphrases can be extracted, the other paraphrase is also applicable. The generalization is applied only to the first verb or adjective in a chunk, because the second and subsequent verbs are often non-independent verbs. It is often the same with adjectives. The example of generalization is shown in Table V.

3) Application to the paraphrases: The paraphrases are applied using the pair before and after paraphrasing. The applied paraphrase is performed in the following procedure.

(i) Divide the input sentence into chunk sequences with CaboCha.
(ii) Search for the part that matches the chunk sequence before paraphrasing from the input. The search is performed in order from the beginning of the input sentence.
(iii) Substitute the chunk sequence with the paraphrased chunk sequence.
(iv) Change to the correct conjugation using the surface of the input word.

If there are multiple matches in step 2, we repeat step 2 and step 3. In addition, the search is sequentially performed from the top of the input chunk sequence and the paraphrase whose source sequence of the chunk is the longest. If there is still more than one paraphrase, the most frequent paraphrase is selected.

4) Filtering paraphrase by frequency: Even if there are any applicable paraphrases, they are not always applied. In the proposed method, the frequency of occurrence of the source chunk sequence of the paraphrase in the data set($= n_{all}$) and the frequency of the applied paraphrase($= n_{applied}$) are counted. The ratio($\alpha = n_{applied}/n_{all}$) is determined using those values. Paraphrases are filtered using the ratio α and threshold($= \alpha_{th}$). By applying the paraphrases only when α exceeds the threshold($= \alpha_{th}$), it is possible to adjust how much the paraphrases are applied. In this study, we set α_{th} to 0.4 using validation data.

C. Hybrid method

Our proposed method described above can be hybridized with conventional machine learning methods. In the hybrid method, the proposed method is used as preprocessing for the input of the machine translation model. In other words, simplification is performed in the two steps of the proposed method and the machine translation approach. The hybrid method can simplify sentences that can be simplified only by either of the two methods. In this study, we set α_{th} to 0.6 using validation data.

V. EXPERIMENTS

We performed experiments using the dataset described below to compare the approaches described above. The evaluation was performed by automatic evaluation and manual evaluation described below. In automatic evaluation, a system that outputs without changing the input sentences was set as a baseline.

A. Datasets

The dataset used for the experiments is the grammatically simplified corpus described above. The sentence pairs with the common source sentence paraphrased by the 13 workers are used as test data, and 58,438 sentence pairs are used for

[11] https://docs.python.org/3/library/difflib.html

Table V
GENERALIZED PARAPHRASES OF TABLE III

	Before	After
1	元気である 。	Adjective(dearu-lemma form)
2	滅ぼされた 。	Verb(irrealis form) れた
3	お 好きな だけ	お Adjective(adverbial form) だけ
4	知ら ないで す 。	Verb(irrealis form) ないで す
5	必死に なって	Adjective(adverbial form) なって
6	どう す べき か	どう Verb(literally lemma form) べき か

Table VI
CRITERIA FOR HUMAN EVALUATION

	Fluency
4	It is a grammatically correct sentence.
3	It has some grammatical mistakes, but the meaning of the sentence can be understood.
2	The grammar is incorrect, but you can guess the meaning.
1	It has many grammatical mistakes, and the meaning cannot be understood.

	Meaning preservation
4	The meanings of the two sentences are the same.
3	The specific meanings of the two sentences are different, but the overall meaning is the same.
2	The specific meanings of the two sentences are different, but the meanings of parts are the same.
1	The meanings of the two sentences are quite different.

the extraction of paraphrases while 5,000 sentence pairs are used to determine hyperparameters.

B. Automatic evaluation

We evaluate each method with BLEU, SARI [19], and the simplification rate. Although BLEU is a metric originally used to evaluate machine translation, it is also used to evaluate text simplification [10], [11], [13], [15]. SARI is a metric for simplicity that evaluates n-gram F-measure (addition, keeping) and relevance in the operation of simplification. We used multi-bleu.perl (Moses) [12] to calculate BLEU and SARI.py[13] to calculate SARI. We define the simplification rate as follows. First, the number of not-simple chunks (= N_input) included in the input sentence is counted. The same is done to output sentences (= N_output). Then, $(N_input - N_output)/N_input$ is calculated and defined as the simplification rate. The simplification rate has a value between 0 and 1. If the number of not-simple chunks does not change between the input and output, the value is 0. If the output sentence does not contain not-simple chunks, the value is 1.

C. Human evaluation

We manually evaluated the “fluency” and “meaning preservation.” The criteria for manual evaluation are shown in Table VI, which is the same as that used in [15]. Manual evaluation was performed by the first author of this paper. The scores are published to confirm whether the evaluation is reasonable[14].

Table VII
AUTOMATIC EVALUATION

Methods	BLEU	SARI	Simplicity [%]
Baseline (same as input)	50.78	21.98	0
MT approach	72.59	67.65	77.8
Proposed method	67.12	55.98	54.5
Hybrid method	**73.63**	**68.84**	**82.8**

Table VIII
HUMAN EVALUATION

Methods	Fluency	Meaning preservation
MT approach	3.69	3.62
Proposed method	**3.87**	**3.90**
Hybrid method	3.75	3.70

VI. RESULTS

Some of the input and output results are shown in the appendix. This section shows the results of automatic evaluation and manual evaluation.

A. Automatic evaluation

The automatic evaluation results of each method are shown in Table VII. As can be seen from the baseline’s high BLEU, input sentences and references are similar. Therefore, BLEU of each method is higher than a general machine translation task. The results show that the proposed method is lower in both BLEU, SARI, and simplification rates than the machine translation approach. However, the hybrid approach attains the highest score.

B. Human evaluation

The human evaluation results of each method are shown in Table VIII. When the input sentence and output one are the same, it is excluded from the result. Because, in

[12] http://www.statmt.org/moses/
[13] https://github.com/cocoxu/simplification
[14] Hidden for anonymization now

that case, the fluency and the meaning preservation always have a score of 4. The machine translation approach has low fluency and meaning preservation, while those of the proposed method are relatively high. The hybrid method has a score between the two methods'. The hybrid method has higher fluency and meaning preservation than that of the machine translation approach, which indicates that the sentences that can be simplified are different in the machine translation approach and the proposed method. Therefore, the hybrid method works effectively.

VII. Conclusion

We constructed a Japanese grammatical simplification corpus and compared conventional machine translation approach, proposed method by extracting differences between sentence pairs, and hybrid method of them as simplification methods. We obtained the result that the hybrid method is effective in both automatic evaluation and manual evaluation. This means that sentences that can be simplified are different in the machine translation approach and proposed method. In grammatical simplification, it is important to take advantage of the fact that input and output sentences change only partially. In this study, focus was placed only on grammatical simplification; however, general simplification may have the same tendency.

Meanwhile, the proposed method still requires improvement. For example, when multiple paraphrases can be selected, selecting the most frequent paraphrase may not be the most suitable for the context. Therefore, it is necessary to use more information than frequency to find the most suitable paraphrase. In this experiment, paraphrasing is applied sequentially from the top of the input sentence; however, the optimal order is not clear. If there are more than one paraphrasing orders, more contextual ones need to be selected.

Future study will involve finding a method to select the optimal paraphrase by considering the context when multiple paraphrases can be applied. If it is possible, both the fluency and meaning preservation will be improved.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grants-in-Aid for Challenging Research (Exploratory) Grant ID 17K18481.

References

[1] K. Iwata, "The preference for English in linguistic services: 'Japanese for living: Countrywide survey' and Hiroshima (<special issue> changing Japanese society and language issues)," in *The Japanese Journal of Language in Society*, vol. 13, no. 1, 2010, pp. 81–94.

[2] H. Saggion, "Automatic text simplification. synthesis lectures on human language technologies." Morgan & Claypool Publishers, 2017.

[3] R. Chandrasekar, C. Doran, and B. Srinivas, "Motivations and methods for text simplification," in *COLING*, 1996, pp. 1041–1044.

[4] S. Štajner and M. Popovic, "Can text simplification help machine translation?" in *EAMT*, 2016, pp. 230–242.

[5] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks," in *EMNLP*, 2008, pp. 254–263.

[6] C. Callison-Burch, "Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk," in *EMNLP*, 2009, pp. 286–295.

[7] E. Pavlick and C. Callison-Burch, "Simple PPDB: A paraphrase database for simplification," in *ACL*, 2016, pp. 143–148.

[8] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, "PPDB: The paraphrase database," in *NAACL-HLT*, 2013, pp. 758–764.

[9] T. Kajiwara and M. Komachi, "Simple PPDB: Japanese," in *Proc. of the 23rd Natural Language Processing of Japan*, 2017, pp. 529–532.

[10] S. Štajner, H. Bechara, and H. Saggion, "A deeper exploration of the standard PB-SMT approach to text simplification and its evaluation," in *ACL-IJCNLP*, 2015, pp. 823–828.

[11] S. Nisioi, S. Štajner, S. P. Ponzetto, and L. P. Dinu, "Exploring neural text simplification models," in *ACL*, 2017, pp. 85–91.

[12] W. Hwang, H. Hajishirzi, M. Ostendorf, and W. Wu, "Aligning sentences from standard Wikipedia to simple Wikipedia," in *NAACL-HLT*, 2015, pp. 211–217.

[13] T. Maruyama and K. Yamamoto, "Simplified corpus with core vocabulary," in *LREC*, 2018, pp. 1153–1160.

[14] A. Katsuta and K. Yamamoto, "Crowdsourced corpus of sentence simplification with core vocabulary," in *LREC*, 2018, pp. 461–466.

[15] T. Maruyama and K. Yamamoto, "Sentence simplification with core vocabulary," in *IALP*, 2017, pp. 363–366.

[16] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," in *ACL*, 1991, pp. 177–184.

[17] R. Barzilay and K. R. McKeown, "Extracting paraphrases from a parallel corpus," in *ACL*, 2001, pp. 50–57.

[18] I. Iori, "Issues on the study of "Yasashii-nihongo": Today and tomorrow," in *The Hitotsubashi journal for Japanese language education*, 2014, pp. 1–12.

[19] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," *TACL*, vol. 4, pp. 401–415, 2016.

Appendix: Some simplified outputs and the reference

	Japanese	English
Input	私 は 午後 ずっと その 本 を 読み 続けて いる 。	I have been reading the book all afternoon.
Our proposed method	私 は 午後 ずっと その 本 を 読み 続けて いる 。	I have been reading the book all afternoon.
MT approach	私 は 午後 ずっと その 本 を 読む 続けて い ます 。	I have been read the book all afternoon.
Hybrid method	私 は 午後 ずっと その 本 を 読んで い ます 。	I have been reading the book all afternoon.
Reference	私 は 午後 ずっと その 本 の 読書 を 続けて い ます 。	I have been reading the book all afternoon.

	Japanese	English
Input	すぐ に 出て 行け ！	Get out soon!
Our proposed method	すぐ に 出て 行け ！	Get out soon!
MT approach	すぐ に 出て ください ！	Please get out soon!
Hybrid method	すぐ に 出て ください ！	Please get out soon!
Reference	すぐ に 出て いき なさい ！	Get out soon!

	Japanese	English
Input	君 に 会い たかった よ 。	I wanted to see you.
Our proposed method	君 に 会い たかった よ 。	I wanted to see you.
MT approach	君 に 会い たい と 思い ました よ 。	I wanted to see you.
Hybrid method	君 に 会い たい です よ 。	I wanted to see you.
Reference	君 に 会い たい と 思って い ました 。	I wanted to see you.

	Japanese	English
Input	すべて は 結局 同じ よ 。	Everything is the same after all.
Our proposed method	すべて は 結局 同じ よ 。	Everything is the same after all.
MT approach	すべて は 結局 です よ 。	Everything is after all.
Hybrid method	すべて は 結局 続き ました よ 。	Everything went on eventually.
Reference	すべて は 結局 同じ です よ 。	Everything is the same after all.

	Japanese	English
Input	私 が 外出 して いる あいだ 、 犬 の 面倒 を みて くれ ない 。	Will you take care of my dog while I'm out?
Our proposed method	私 が 外出 中 は 、 犬 の 面倒 を みて ください 。	Please take care of my dog while I'm out.
MT approach	私 が 外出 した とき 、 犬 の 面倒 を みて ください 。	Please take care of my dog when I'm out.
Hybrid method	私 が 外出 中 は 、 犬 の 面倒 を みて ください 。	Please take care of my dog while I'm out.
Reference	私 が 外出 する 間 に 、 犬 の 面倒 を みて ください 。	Please take care of my dog while I'm out.

	Japanese	English
Input	私 は 先生 に 叱ら れた 。	I was scolded by my teacher.
Our proposed method	私 は 先生 に 叱り ました 。	I scolded my teacher.
MT approach	先生 を 私 は 叱り ました 。	I scolded my teacher.
Hybrid method	私 は 先生 を 叱り ました 。	I scolded my teacher.
Reference	私 は 先生 に 叱ら れ ました 。	I was scolded by my teacher.

	Japanese	English
Input	この 価格 に は 、 運賃 は 含ま れて い ませ ん 。	This price does not include fares.
Our proposed method	この 価格 は 、 運賃 を 含んで い ませ ん 。	This price does not include fares.
MT approach	この 価格 に は 、 運賃 は 含ま い ませ ん 。	This price does not <broken> fares.
Hybrid method	この 価格 は 、 運賃 を 含んで い ませ ん 。	This price does not include fares.
Reference	この 価格 は 、 運賃 を 含み ませ ん 。	This price does not include fares.

	Japanese	English
Input	小銭 を お 持ち です か 。	Do you have change?
Our proposed method	小銭 を 持って い ます か 。	Do you have change?
MT approach	小銭 を 教えて ください 。	Please tell me your change.
Hybrid method	小銭 を 持って い ます か 。	Do you have change?
Reference	小銭 を 持って い ます か 。	Do you have change?