Sentence simplification with core vocabulary

Takumi Maruyama and Kazuhide Yamamoto
Nagaoka University of Technology
1603-1, Kamitomioka Nagaoka, Niigata 940-2188, JAPAN
{maruyama, yamamoto}@jnlp.org

***Abstract*—We attempt automatic text simplification with vocabulary restriction on the output side using a machine translation approach based on a simplified corpus that we built. This is the first machine translation approach in Japanese because no Japanese simplification corpus has been created to date. This corpus focuses only on paraphrases of sentence units and phrase units. It is the first time that this type of simplification has been used with such a corpus. This approach makes it possible to simplify better than existing systems do. We also compared models that changed the quantity and quality of the training data and development data. The result shows that data having a medium S-BLEU score between the original sentence and a simple sentence is most effective for automatic text simplification by a machine translation approach.**

***Index Terms*—text simplification; paraphrasing;**

I. Introduction

Automatic text simplification is a task that reduces the complexity of a vocabulary and expressions while preserving the main meaning of the text. This technique can be used to make many text resources available for a wide range of readers including children, nonnative speakers, and the disabled. As a preprocessing step, simplification can improve the performance of natural language processing (NLP) tasks including parsing [1], summarization [2], [3], semantic role labelling [4], information extraction [5], and machine translation [6]. Automatic text simplification is generally divided into three approaches: lexical simplification, rule based, and machine translation.

In Japanese, there are few studies on text simplification; only lexical simplification has been researched. Existing Japanese lexical simplification (LS) systems often produce results that do not suit the context [7]. In addition, systems like [8] and [7] are unable to simplify complex syntactic structures. Thus, it is necessary to address not only lexical simplification but also sentence simplification. However, there is no parallel corpus for Japanese text simplification.

Therefore, we experiment with a machine translation approach based on a simplification corpus (Easy Japanese Parallel Corpus) that we built. Because there is a restriction that a simple sentence must consist of core vocabulary (2,000 words), this corpus is mainly composed of simplifications in units of phrases and sentences. We expect that simplification that considers context becomes possible by using this corpus. In addition, we think that such a corpus is easier to expand than a simplification corpus that includes many kinds of simplification operations and rules. In order to expand the corpus size in the future, it is important to know how the quantity and quality of training data affect the performance of the machine translation system. Therefore, we investigate two important issues in text simplification:

1) Can an automatic text simplification system of sentence units output more simplified sentences when compared to existing LS systems?
2) Do the quality and quantity of training data and development data affect sentence-level simplification systems?

To explore these problems, we conducted two experiments. First, we compared machine translation systems with existing Japanese LS systems (Section IV-B). Second, we built 32 models by changing the quantity and quality of the training data and development data, and compared them (Section IV-C).

II. Related Works

LS is an approach to replace complex words in sentences with simple synonyms. In Japanese LS, generating synonyms from dictionaries and WordNet [8], as well as methods for obtaining synonyms using word embedding [7], have been attempted. In order to compare our system with a LS system, we use Kajiwara's Japanese LS system, which is the only publicly available system [8]. This system simplifies based on the following four components: (1) identification of complex words, (2) substitution generation, (3) word sense disambiguation, and (4) synonym ranking. In these systems, there are problems with paraphrases that do not suit the context.

III. Text Simplification Corpus

We used the Easy Japanese Parallel Corpus, which was constructed manually from a Japanese-to-English Machine Translation corpus[1]. The original Japanese side of the corpus contains a sentence length of 4 to 16 words. The simple sentences of this corpus consist of only core vocabulary and named entities.

Table I shows example sentences in our corpus and the S-BLEU [2] [9] score between the original sentence and simplified sentence. Underlined words are not core vocabulary (complex words) in original sentences. One original sentence contains one complex word or two complex words. In each sentence, the number of complex words is approximately the same, but the scale of change when simplifying is different. Therefore, S-BLEU scores show variations. When a pair with an original

[1]small parallel enja: 50k En/Ja Parallel Corpus for Testing SMT Methods
[2]S-BLEU is Sentence-wise BLEU score

978-1-5386-1981-0/17/$31.00 ©2017 IEEE

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on November 14,2020 at 19:09:17 UTC from IEEE Xplore. Restrictions apply.

TABLE I
EXAMPLE SENTENCE PAIRS OF EASY JAPANESE CORPUS AND S-BLEU SCORES BETWEEN THE ORIGINAL AND THE SIMPLIFIED

	S-BLEU	Version	Sentence	Literal English translation
(1)	0.000	Original	疑い の 余地 はない。	There is no room for doubt.
		Simplified	明らかだ。	It is clear.
(2)	0.090	Original	日本では、月給 です。	In Japan, salary is on monthly basis.
		Simplified	日本では、月に１度、働いた分のお金がもらえます。	In Japan, you get money once for working a month.
(3)	0.517	Original	交通 渋滞 のため、私は遅れました。	Because of the traffic jam, I was late.
		Simplified	道路が混んでいたため、私は遅れました。	I was late because the road was crowded.
(4)	0.598	Original	いつも 手近 に辞書を持っていなさい。	Always have your dictionary near at hand.
		Simplified	いつでも使えるように辞書を持っていなさい。	Have your dictionary so that you can use it anytime.
(5)	0.702	Original	彼は 一生懸命 英語を勉強したに違いない。	He must have studied English with utmost effort.
		Simplified	彼は頑張って英語を勉強したに違いない。	He must have studied English hard.
(6)	0.791	Original	十分に 休養 をとることは、非常に大切です。	It is very important to take a rest.
		Simplified	十分に休みをとることは、非常に大切です。	It is very important to take a rest.

TABLE II
CORPUS STATISTICS

	Original	Simplified
Total #sentences	50,000	50,000
Total #tokens	490,021	516,881
Total #words (unique tokens)	8,786	2,238
Avg. #characters per sentence	14.79	15.35
Avg. #words per sentence	9.80	10.34

sentence and a simple sentence do not match and are classified by an S-BLEU score, they are 17.7% for [0.0, 0.1), 17.5% for [0.5, 0.6], and 16.3% for [0.7, 0.8). In [0.0, 0.1), original sentences are greatly transformed, such as in (1) and (2). In addition, in [0.5, 0.6), there is a tendency to transform phrase units such as (3) and (4), and in [0.7, 0.8), transformation is done for only one word, such as in (5) and (6). Thus, in this corpus, operations ranging from word units to sentence units are approximately equally included.

In addition, we classified the simplification operation for about 100 sentences randomly selected from this corpus. The percentage of only paraphrasing operations accounted for 97%. When we examined the three sentences for which insertion is performed, we found that the insertion and the paraphrase are combined to make the simple sentence. Insertion is an operation that inserts only a postpositional particle to make a fluent sentence. Therefore, this simplification corpus is a corpus focusing only on paraphrases.

IV. EXPERIMENT

A. Experimental Setup

For text simplification by the machine translation approach, we use phrase-based statistical machine translation (PB-SMT) and neural machine translation (NMT). We use the Easy Japanese Parallel Corpus for training and testing. The settings for each machine translation system are shown below.

1) PB-SMT: We use Moses [10] for PB-SMT. We tune the systems using minimum error rate training (MERT) [11]. For the language model (LM), we use the training data and build a 3-gram language model with Modified Kneser-Ney smoothing trained with KenLM [12].

2) NMT: We use OpenNMT for NMT. The number of LSTM layers in the Encoder/Decoder is 2, and the number of LSTM units in the hidden layer is 500. In addition, the number of dimensions of word embedding is 500. We use stochastic gradient descent for optimization.

The following two experiments were carried out using these machine translation systems.

1) Comparison of automatic simplification using machine translation system and existing LS system (Section IV-B);
2) Comparison of models with different quantity and quality of training data and development data (Section IV-C);

B. Comparison of Machine Translation Approach and Existing LS System

We compare four systems: the baseline, LS, SMT, and NMT. For the baseline, we adopt a system that does not change input sentences. That is, it is the input sentence itself. For the LS system, we use the only available Japanese LS system by Kajiwara et al. [8]. In the machine translation system, the training data contains 40,000 sentences, while the development and test data each contain 5,000 sentences.

We evaluate each system by three metrics: S-BLEU, SARI, and Simplicity.

1) SARI: A recently proposed simplification metric compares the **S**ystem output **A**gainst **R**eferences and against the **I**nput sentence. This is an arithmetic average of n-gram precision and the recall of three rewrite operations: addition, retention, and deletion. It rewards addition operations where system output was not in the input but occurred in the references. In addition, it rewards words retained/deleted in both the system output and the references. In an experimental evaluation, Xu et al. indicated that SARI correlates with judgments of human simplicity [13].

2) Simplicity: In this paper, we define simplicity as the measure of the content of sentences from the 2,000 core words. We define a sentence by word w_i as follows:

$$Sentence = \{w_0, w_1 .... w_N\} \quad (1)$$

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on November 14,2020 at 19:09:17 UTC from IEEE Xplore. Restrictions apply.

TABLE III
EXAMPLE OF OUTPUT

	System	Output sentence	Literal translation of the output
(1)	Original (Baseline)	そこに 署名 してください。	Please sign there.
	LS	そこに 署名 してください。	Please sign there.
	SMT	そこに 署名 してください。	Please sign there.
	NMT	そこに名前を書いてください。	Please write your name there.
(2)	Original (Baseline)	彼は大家族を 養う ために 懸命 に働いた。	He worked hard to feed a large family.
	LS	彼は大家族を 養う ために 懸命 に働いた。	He worked hard to feed a large family.
	SMT	彼は大家族を 養う ために頑張って働いた。	He worked hard to feed a large family.
	NMT	彼は大家族を支えるために頑張って働いた。	He worked hard to support the large family.
(3)	Original (Baseline)	彼は人生に 確固 とした目的を持っている。	He has a firm purpose in life.
	LS	彼は人生に強いとした目的を持っている。	He has a purpose which it is strong in life.
	SMT	彼は人生に 確固 とした目的を持っている。	He has a firm purpose in life.
	NMT	彼は人生にしっかりとした目的を持っている。	He has a firm purpose in life.
(4)	Original (Baseline)	和合 して生活している。	They live together in unity.
	LS	和合 して生活している。	They live together in unity.
	SMT	和合 して生活している。	They live together in unity.
	NMT	系を集めている。	They are collecting threads.

TABLE IV
AUTOMATIC EVALUATION

System	S-BLEU	SARI	Simplicity
Baseline	0.706	0.278	0.924
LS	0.661	0.335	0.932
SMT	0.757	0.499	0.969
NMT	**0.794**	**0.585**	**0.998**

TABLE V
HUMAN EVALUATION

	Grammar			Meaning preservation		
System	Mean	Mode	Median	Mean	Mode	Median
LS	3.45	4	4	3.37	4	4
SMT	3.76	4	4	3.79	4	4
NMT	3.11	4	4	3.04	4	4
Manual	**3.81**	4	4	**3.72**	4	4

We define *Simplicity* as follows:

$$Simplicity = \frac{\sum_{i=0}^{N} f(w_i)}{N} \tag{2}$$

$$f(w_i) = \begin{cases} 1 & (w_i : \text{core word}) \\ 0 & (w_i : \text{other}) \end{cases} \tag{3}$$

We also conduct human evaluations on grammar and meaning preservation. We set the evaluation as four stages of 1 to 4 (higher marks indicate better output). We randomly extracted 100 sentences from the corpus. We asked five native speakers using crowdsourcing. Each subject evaluated 400 sentences in two aspects (grammar and meaning preservation). The 400 sentences consisted of four kinds of outputs (i.e., Baseline, LS, SMT, NMT), each of which had 100 sentences.

C. Comparison of Models with Different Quantity and Quality of Training Data and Development Data

We constructed models that vary by the quantity and quality of training data and development data. We classified them into four ranges [0.0, 0.4), [0.4, 0.6), [0.6, 0.8), and [0.8, 1.0] according to the S-BLEU score between the original sentence and the simple sentence. For each section, we set the training data size as 2,000, 4,000, 6,000, and 8,000 sentences. Furthermore, we set the development data size to 100, 200, 300, and 400 sentences. From these settings, we constructed models for each combination of quality and quantity for the machine translation system (SMT and NMT). The test data was 400 sentences which we randomly extracted 100 sentences from each section of the S-BLEU score. We evaluated the output of each model with S-BLEU and SARI.

V. RESULT AND DISCUSSION

In the LS system, since 200 sentences could not be analysed, the results shown exclude these 200 sentences. Table IV shows the results of an automatic evaluation.

Although SMT and NMT are trained with a corpus composed of 2,000 words on the simple sentence side, the score of simplicity is not 1.0. This is caused by words with different parts of speech and UNK tokens [3]. The machine translation systems (SMT and NMT) outperform the LS system in all automatic evaluation metrics. LS rewrites "確固 (firm)" as "強い (strong)" in Table III (3). Although the meanings of "確固 (firm)" and "強い (strong)" are similar, it is not grammatical to have "とした" after "強い (strong)". On the other hand, NMT rewrites "確固 (firm)" as "しっかり (firm)". These are equivalent meanings and are transformations suitable for grammar. Even in cases where the LS system cannot take context into consideration, the machine translation system can translate appropriately.

Table V shows the results of a human evaluation. The score of "Manual" for grammar and meaning preservation is not 4.00. Owing to vocabulary restrictions, it is not always possible to represent the meaning of the original sentence perfectly. Thus, we think that the meaning preservation score

[3]The UNK token is an out-of-vocabulary word. Words not included in training data are replaced by this token.

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on November 14,2020 at 19:09:17 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
RESULTS OF NMT

S-BLEU	Size of the training set			
	2,000	4,000	6,000	8,000
[0.0, 0.4)	0.000	0.0428	0.241	0.408
[0.4, 0.6)	0.005	0.131	0.472	0.588
[0.6, 0.8)	0.008	0.305	0.553	**0.606**
[0.8, 1.0]	0.002	0.076	0.544	0.580

TABLE VII
RESULTS OF SMT

S-BLEU	Size of the training set			
	2,000	4,000	6,000	8,000
[0.0, 0.4)	0.569	0.581	0.592	0.592
[0.4, 0.6)	0.579	0.599	0.613	**0.621**
[0.6, 0.8)	0.595	0.608	0.603	0.606
[0.8, 1.0]	0.566	0.563	0.569	0.570

became 3.72. In addition, for grammar, there was a tendency that low scores were given to sentences that became longer and ambiguous owing to vocabulary restrictions.

In automatic evaluation, NMT has the best result, but in human evaluation, NMT is inferior to the LS system and SMT. The cause of this difference is UNK tokens. SMT and LS systems often output complex words without changing them. For example, in Table III (1), SMT and LS do not rewrite "署名 (signature)". In addition, in (2), "養う (to feed)" and "懸命 (hard)" are not core words, but the LS system outputs the same sentence as the original sentence. SMT rewrites "懸命 (hard)" as "頑張って (hard)", but does not convert "養う (to feed)". On the other hand, NMT converts each complex word. Therefore, NMT often outputs sentences having completely different meanings with the original sentence as Table III (4). As a result, SMT outperforms NMT in human evaluations. In addition, we think that the SMT output has better results than the "Manual" in human evaluations because there is no restriction on the SMT output vocabulary. According to the human evaluation criteria, Grammar score 3 or more is an evaluation that "you can understand the meaning of sentences". In addition, meaning preservation score is 3 or more, is an evaluation that "most of the meaning is the same between original sentence and simple sentence". Although output vocabulary of NMT is restricted, the output sentence of NMT has sufficient results in grammar and meaning preservation.

Table VI and Table VII show the results evaluated by S-BLEU for the model in which the quantity and quality of the data are changed. Focusing on the quality of the data, it is found that the model trained with a medium S-BLEU score data is good in both NMT and SMT. The score is low for [0.0, 0.4) and [0.8, 1.0]. In addition, it is shown that the scale of the corpus influences the output result in the sentence simplification. If the corpus size expands, collecting data with a medium S-BLEU score is most effective for improving the simplification performance.

VI. CONCLUSION

We performed automatic text simplification by using a machine translation approach with a simplified corpus in Japanese for the first time. This approach also makes it possible to make correct simplifications for cases where existing LS systems cannot simplify properly. This proposed system greatly outperforms those systems that adopt conventional methods. In addition, we constructed 32 models according to the quantity and quality of training data, development data, and the machine translation system. A comparison of these models showed that data with a medium S-BLEU score are most effective for automatic text simplification by a machine translation approach.

ACKNOWLEDGMENT

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

REFERENCES

[1] R. Chandrasekar, C. Doran, and B. Srinivas, "Motivations and Methods for Text Simplication," *Proc. COLING '96*, vol. 2, pp. 1041–1044, 1996.
[2] A. Siddharthan, A. Nenkova, and K. McKeown, "Syntactic simplification for improving content selection in multi-document summarization," *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, pp. 896—-es, 2004.
[3] W. Xu and R. Grishman, "A Parse-and-Trim Approach with Information Significance for Chinese Sentence Compression," *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pp. 48–55, 2009.
[4] D. Vickrey and D. Koller, "Sentence Simplification for Semantic Role Labeling," *Proceedings of ACL-08: HLT*, pp. 344–352, 2008. [Online]. Available: http://www.aclweb.org/anthology/P/P08/P08-1040
[5] M. Miwa, R. Sætre, and Y. Miyao, "Entity-Focused Sentence Simplification for Relation Extraction," *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 788–796, 2010.
[6] H.-b. Chen, H.-H. Huang, H.-H. Chen, and C.-T. Tan, "A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications," *Coling-2012*, vol. 2, pp. 545–560, 2012.
[7] M. Hading and Y. Matsumoto, "Japanese Lexical Simplification for Non-Native Speakers," *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications*, pp. 92–96, 2016.
[8] T. Kajiwara and K. Yamamoto, "Evaluation Dataset and System for Japanese Lexical Simplification," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 35–40, 2015.
[9] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, no. July, pp. 311–318, 2002.
[10] P. Koehn, W. Shen, M. Federico, N. Bertoldi, C. Callison-Burch, B. Cowan, C. Dyer, H. Hoang, O. Bojar, R. Zens, A. Constantin, E. Herbst, and C. Moran, "Open Source Toolkit for Statistical Machine Translation," *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, no. June, pp. 177–180, 2006.
[11] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, vol. 1001, pp. 160–167, 2003.
[12] K. Heafield, "KenLM : Faster and Smaller Language Model Queries," *Proceedings of the 6th Workshop on Statistical Machine Translation*, pp. 187–197, 2011.
[13] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing Statistical Machine Translation for Text Simplification," *Transactions of the ACL*, vol. 4, pp. 401–415, 2016.

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on November 14,2020 at 19:09:17 UTC from IEEE Xplore. Restrictions apply.