Japanese News Simplification: Task Design, Data Set Construction, and Analysis of Simplified Text

Isao Goto goto.i-es@nhk.or.jp
Hideki Tanaka tanaka.h-ja@nhk.or.jp
Tadashi Kumano kumano.t-eq@nhk.or.jp
NHK Science & Technology Research Laboratories,
1-10-11 Kinuta, Setagaya-ku, Tokyo 157-8510, Japan

Abstract

In this paper we explore a Japanese news simplification task. We designed a Japanese news simplification task, constructed the data set for the task, and analyzed the manual simplification process. We designed the task focusing on sentence-level simplification, which is part of the process of manual simplification of Japanese news for non-native speakers. We constructed the data set consisting of Japanese news sentences and their corresponding simplified Japanese news sentences, and verified the effectiveness of the data set for automatic simplification by conducting preliminary experiments using phrase-based statistical machine translation. To reveal the processes behind manual simplification, such as simplification associated with word order (syntactic structure), we analyzed manually simplified Japanese news sentences.

1 Introduction

Simplified texts increase readability and understandability for non-native speakers and children. Simplified news texts are especially useful for daily living because news delivers information needed for life. There are a number of simplified texts available in English, such as Learning English provided by Voice of America and the BBC, which are multimedia sources of news and information geared toward learners, Simple English Wikipedia, and simplified technical English based on the ASD-STE100 standard[1]. There is much research on automatic English simplification (Chandrasekar et al., 1996; Carroll et al., 1998; Petersen and Ostendorf, 2007; Zhu et al., 2010; Coster and Kauchak, 2011; Woodsend and Lapata, 2011; Wubben et al., 2012; Kauchak, 2013; Narayan and Gardent, 2014). To realize automatic simplification using statistical machine translation (SMT), a parallel corpus consisting of sentences and their corresponding simplified sentences is needed. As such, an English parallel corpus was constructed using English Wikipedia and Simple English Wikipedia (Coster and Kauchak, 2011). In contrast, there is less research on the simplification of languages other than English: Portuguese (Aluísio et al., 2008), Spanish (Bott and Saggion, 2011; Bott et al., 2012), Italian (Dell'Orletta et al., 2011), French (Seretan, 2012), German (Klaper et al., 2013), and Japanese (Inui et al., 2003; Moku and Yamamoto, 2012; Tanaka et al., 2013). In these studies, work focusing on Japanese simplification can be summarized as follows. Inui et al. (2003) collected paraphrase readability rankings by consulting with teachers at schools for the deaf, and trained a paraphrase ranking model from the collected rankings. Moku and Yamamoto (2012) performed the automatic simplification of official documents in Japanese. They concluded through their pre-experiment, however, that

[1] http://www.asd-ste100.org/

their method was ineffective. Tanaka et al. (2013) reported on the Internet news service NEWS WEB EASY, which provides manually simplified Japanese news produced through the rewriting of Japanese news. Because the simplification is conducted manually by humans, automation becomes the key to solving the issue of efficiency. To our knowledge, there is currently no work on the automatic simplification of Japanese news texts using SMT.

In this paper, we explore a Japanese news simplification task. We design the task, construct the data set for it, then reveal the manual simplification process by analyzing the simplified text. In the process of manually simplifying Japanese news for the NEWS WEB EASY service (Tanaka et al., 2013), designed for foreigners living in Japan, there are two types of operations: article-level shortening for conciseness (especially when articles are long), and sentence-level simplification of expressions. News reporters mainly carry out article-level shortening for conciseness, while Japanese instructors mainly carry out sentence-level simplification of expressions (Section 2). We focus on the process of sentence-level simplification of expressions, which is thought as closer to being serviceable for practical use than automatic article-level shorting, and define this sentence-level process here as the simplification task (Section 3). We construct a data set of parallel sentences consisting of sentences from Japanese news prior to being initially simplified by the Japanese instructors, and the resulting simplified sentences (Section 4). Then, we conduct preliminary experiments on automatic simplification using the data set and phrase-based SMT to verify the effectiveness of the data set (Section 5). Additionally, we produce manually annotated word alignments between parallel sentences and reveal the processes of manual simplification, such as rewriting with a different word order (i.e., syntactic structure) by analyzing the annotated word alignments (Section 6). Our contributions are summarized as follows:

- We proposed a Japanese news simplification task, constructed a data set for the task, and verified the effectiveness of the data set through experiments on automatic simplification using phrase-based SMT.
- We revealed the processes of manual simplification, such as rewrites associated with word order (syntactic structure), which are needed to design effective automatic simplification.

2 Simplification in Easy Japanese News Service

In this section we describe the Japanese news simplification processes in the simplified Japanese (easy Japanese) news service called NEWS WEB EASY (Tanaka et al., 2013); our research target for this study. The service is offered by NHK. NEWS WEB EASY is an online service that provides easy Japanese news for foreigners living in Japan who have reached pre-intermediate-level Japanese. Easy Japanese news is mainly produced through two operations: (1) shortening original news articles for conciseness, especially when the articles are long, and (2) simplifying expressions. Long Japanese news articles can be a burden for non-native speakers to read. Japanese news articles can be long or short, but there are few long articles that feature easy to understand Japanese. When longer news articles are shortened, the resulting articles are sometimes one-half or one-third the original length. The degree of shortening depends on the original length. News reporters mainly shorten with the goal of conciseness, while designated Japanese instructors[2] well-versed in easy Japanese news mainly simplify expressions found in the text. For each original article to be simplified, a news reporter and a Japanese instructor take turns rewriting the article. This rewriting process by the reporter and the Japanese instructor is repeated two or more times, as needed.[3]

[2]These instructors have experience teaching Japanese to non-native speakers.

[3] The percentage of original articles first rewritten by Japanese instructors before being rewritten by news reporters changed over time, as follows. In April 2012, when the NEWS WEB EASY test service started, the rate was 58%.

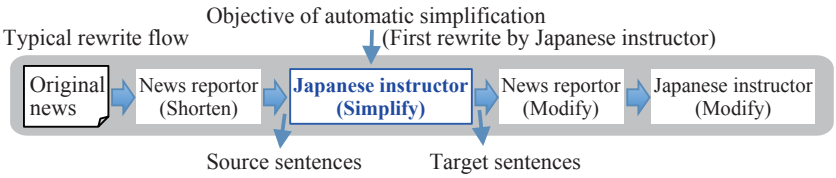

Figure 1: Typical rewrite flow and objective of automatic simplification

3 Sentence-Level Japanese News Simplification Task

Shortening news articles can be technically simulated to some extent using an automatic summarization technique. However, there are problematic issues related to the use of automatic summarization for actual proper news services. Namely, to shorten long articles into lengths of one-half or one-third of their original lengths, not only are repetitive sections removed, but some of the actual news content is also eliminated. Therefore, there is a risk that some of the news contents that should be delivered to the reader may be deleted. This is a significant risk in terms of accurately conveying news. Additionally, when news reporters shorten an article, the article is reorganized as needed, and this process requires sophisticated editing. For these reasons, it is considered difficult to obtain practical-level quality for news services through automatic summarization. In contrast, the operation Japanese instructors perform does not require shortening and centers mainly on sentence-level simplification of expressions.

Therefore, we divided these two operations, that is, article-level shortening of length and sentence-level simplification of expressions into two separate tasks. In this paper, we define the sentence-level simplification of expressions, which is performed by Japanese instructors, as the simplification task. Automating this operation is thought to be more practical for news services, because sentence-level rewrites do not require removing certain content and a statistical machine translation technique can simulate this operation to some extent. In this task, the input sentences for automatic simplification systems are the sentences prior to being simplified by the Japanese instructors, while the simplified reference sentences are the sentences that are the results of the initial simplification process performed by the Japanese instructors.

4 Data Set for Japanese News Simplification

We produced a data set for the Japanese news simplification task by constructing parallel sentences. These comprised sentences taken from Japanese news articles and their corresponding simplified articles. In this section, we describe the construction method and the constructed data set.

4.1 Construction Method

Here we explain how parallel sentences taken from Japanese news articles and their corresponding simplified Japanese sentences were constructed. In this paper, the sentences prior to being initially simplified by the Japanese instructors are called *source sentences*, and their corresponding output simplified sentences are called *target sentences*. Additionally, a source sentence and its corresponding target sentence(s) are called a *parallel sentence pair*; similarly, an article prior to being initially simplified by a Japanese instructor and its corresponding output simplified article are called a *parallel article pair*.

However, by May 2013, when the official NEWS WEB EASY service started, the rate had dropped to 6%. In September 2014, the rate was down to 1%. Because both Japanese instructors and news reporters began rewriting at the same time when the service was just beginning, approximately half of the news articles were first rewritten by Japanese instructors. As time went on, however, most of news articles came to be first rewritten by news reporters because it was discovered that it was more efficient to simplify expressions after articles had already been shortened.

Table 1: Noise in parallel sentences consisting of one source sentence and one or more target sentence(s)

	Number of sentence pairs	Rate
Without noise	2,012	0.697
With noise	873	0.303

Table 2: Number of sentences in each manually annotated sentence alignment

Number of sentences (source–target)	Frequency	Rate
1–1	2,201	0.698
1–2	575	0.182
1–3	90	0.029
1–0	71	0.023
0–1	67	0.021
2–2	52	0.016
2–1	46	0.015
2–3	17	0.005
1–4	16	0.005
Others	20	0.006

Manual Extraction of Parallel Sentence Pairs for Test Data

We constructed test sentences and their simplified reference sentences by manual extraction to ensure reliability. We manually aligned source sentences with their corresponding target sentences in parallel article pairs. Parallel sentence pairs can be extracted using these sentence alignments. The objective of the simplified reference sentences is to evaluate the quality of automatically simplified sentences, as each test sentence is automatically simplified independently. Thus, we selected test sentences and their simplified reference sentences from parallel sentence pairs under the following conditions:

- One source sentence corresponds to one or more target sentences.
- We eliminate parallel sentence pairs for which major content in the source sentence is not included in its corresponding target sentence(s), or for which content that is not included in a source sentence is included in its corresponding target sentence(s). There are the following exceptions. When major news content in a source sentence is included in its corresponding target sentence(s), the omission of related detailed information is allowed. In the target sentences, additional information not dependent on context and which is always true is allowed.

The first condition is called the alignment condition, and the second condition is called the noise condition. Here we show an example of additional information that does not depend on context and is always true: “Aomori Prefecture” in a source sentence and “Aomori Prefecture in the Tohoku region” in its target sentence.

We manually aligned sentence pairs for 490 parallel article pairs. Among these article pairs, test sentences and their simplified reference sentences could be extracted from 485 article pairs under the abovementioned conditions. The rates for without noise (satisfying the noise condition) and with noise (not satisfying the noise condition) for the parallel sentences that satisfied the alignment condition are shown in Table 1. The rate for without noise was approximately 0.7.

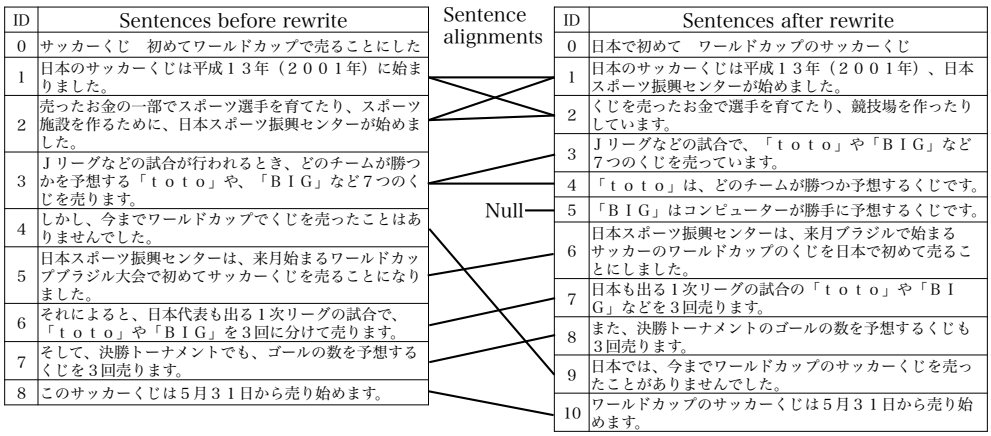

Figure 2: Example of sentence alignments and a parallel article pair before and after being simplified by a Japanese instructor (sentence ID of 0 indicates title)

Requirements of Automatic Sentence Alignment

We use an automatic sentence alignment method for constructing training data because it reduces cost and allows us to automatically use news data produced daily. We therefore examined the requirements of automatic sentence alignment.

As explained in Section 2, news reporters mainly shorten news articles for conciseness, while Japanese instructors mainly simplify the expressions used within the text. However, these roles are not clearly divided. News reporters sometimes simplify parts of expressions, while Japanese instructors not only replace difficult words with simple words and divide sentences into simpler syntactic structures, but they also sometimes omit detailed content, add supplemental explanations, or change the sentence order in an article. Figure 2 shows an example of sentence alignments and a parallel article pair, which consist of articles before and after being simplified by a Japanese instructor.[4] In the example, the source sentence ID (which specifies a sentence) of 3 is split into the two target sentence IDs of 3 and 4. The target sentence ID of 5 is an added sentence. The order of the source sentence ID of 4 was changed in the respective target sentence.

We checked the number of sentences in each manually annotated sentence alignment. The results are shown in Table 2. The rate of one source sentence corresponding to one and more target sentence(s) is high (over 0.91). However, there are some cases of sentence omission, sentence addition, or the inclusion of two source sentences. The average of Kendall's τ, which here represents the correlation between the order of sentences in a parallel article pair, was 0.987.[5] Because the value is close to 1, we can confirm that the sentence orders were the same in most cases.

From these results, we could conclude that the requirements of sentence alignment between the source and target sentences are as follows. In addition to one sentence to one sentence (one-to-one) alignments, alignments including two or more sentences can also be created. Sentence omission and sentence addition can be accounted for, and sentence order can also be considered.

[4] In the example in Figure 2, many expressions in the sentences on the left were already simplified by a news reporter. Reporters sometimes simplify expressions like these.

[5] This value was calculated as follows. We removed target sentences that were not aligned with source sentences. We then projected source sentence IDs to their aligned target sentences. Here, we define the projected order of the target sentences as the order of sentences based on the projected IDs. When two target sentences had the same projected ID, the projected order follows the original target sentence order. We calculated Kendall's τ between the projected order and the order of the target sentence order.

Table 3: Number of sentences in each automatic sentence alignment

Number of sentences (source–target)	Frequency	Rate
1–1	7,893	0.727
1–2	2,336	0.215
1–3	383	0.035
2–1	111	0.010
2–2	44	0.004
1–4	39	0.004
0–1	32	0.003
1–0	19	0.002
3–1	5	0.000

Table 4: Precision and recall of automatic sentence alignment

Number of sentences (source–target)	Precision	Recall
Whole	0.872	0.885
One–One or more	0.881	0.942

Automatic Sentence Alignment for Training Data

Source sentences and target sentences include many identical words, and the sentence order is almost identical. Therefore, we decided to use an alignment method using identical words and dynamic programming, which can efficiently estimate sentence alignments with consideration of sentence order so that sentence alignments do not include crossing alignments. We used Champollion (Ma, 2006), which is an implementation of such a method, for automatic sentence alignment. This method can treat alignments including two or more source or target sentences, sentence omission, and sentence addition.[6] The number of sentences in each automatic sentence alignment for 1,559 parallel article pairs for training data are shown in Table 3.

Quality Evaluation of Automatic Sentence Alignments

We evaluated the quality of automatic sentence alignments using the 490 parallel article pairs that were manually annotated with sentence alignments. Precision and recall for all sentence alignments and for sentence alignments consisting of one source sentence and one or more target sentences are shown in Table 4. Here, the unit of sentence alignment is defined as one parallel sentence pair (e.g., one-to-one, one-to-two, two-to-one, or one-to-null).

Automatic simplification based on monolingual translation has the following characteristics. When output sentences include errors, the quality of the output sentences decreases compared with the input sentences. However, when input sentences are output without modifications, the quality of the output sentences does not decrease compared with the input sentences. Therefore, it is important for simplification to remove as much noise as possible from the training data.

4.2 Data Set for Evaluating Automatic Simplification

Here we describe the specifications of the constructed data set for evaluating automatic simplification. We constructed the data set using news archives from April 2012 to September 2014. We used news articles with edit histories showing simplification by Japanese instructors. In

[6]Target sentences whose sentence order was changed from the source sentence order cannot be aligned perfectly. However, if such target sentences are not aligned with any source sentences, then the extraction of parallel sentences of low quality can be avoided.

Table 5: Specifications of the data set

	Term (Article number)	Number of article pairs	Number of sentence pairs	Extraction Method
Training	2012/04/02 (1) – 2014/02/26 (3)	1,559	10,651	Automatic
Development	2014/02/26 (4) – 2014/04/24 (2)	170	723	Manual
Test	2014/04/24 (3) – 2014/09/30 (5)	485	2,012	Manual

keeping with practical use, we divided the data as follows: Data in the latest term were used as the test data, data in the term immediately prior to the test data term were used as the development data, and data in the term immediately prior to the development data term were used as the training data. The specifics of the data are shown in Table 5. The test data consist of 2,012 manually extracted source sentences with their corresponding simplified reference sentences. The development data consist of 723 parallel sentence pairs manually extracted in the same way as the test data. The training data consist of 10,651 automatically extracted parallel sentence pairs. The training data also include 1,559 news articles with full versions of edit histories in the training term.

5 Experiments on Sentence-Level Automatic Simplification

To confirm the effectiveness of the constructed data set, we conducted preliminary experiments on sentence-level automatic simplification using the data set. We conducted automatic Japanese news simplification as a monolingual translation task using phrase-based SMT, as Coster and Kauchak (2011) did.

5.1 Setup

We used MeCab[7] for Japanese segmentation. Continuous Arabic numerals were merged to one word. We used the Moses implementation (Koehn et al., 2007) as the phrase-based SMT system. The translation model was trained using sentences that are 80 words or less. GIZA++ and grow-diag-final-and heuristics were used to obtain word alignments. To assist the word alignments for low frequency words, we added pairs of the same word to the training data when word alignments were estimated. This is because in the monolingual parallel sentences there are many words that should be aligned to the same words. We used a 5-gram language model that was trained using the target side of the training data. The SMT weighting parameters were tuned via MERT (Och, 2003) using the development data. We used distortion limits of 0 or 6 (default value), which limit the number of words for word reordering to a maximum number. We used the MSD bidirectional lexicalized reordering models of Moses (Koehn et al., 2005).

We compare the trained SMT system (MOSES) to the baseline that does not simplify input sentences and outputs the input sentences without any change (BASELINE). We evaluate using the test data of the data set.

5.2 Results and Discussion

We evaluated the simplification quality based on the automatic evaluation scores from the BLEU-4 (Papineni et al., 2002) and RIBES v1.02.4 (Isozaki et al., 2010), which are commonly used for evaluating translation quality. Percentage scores were used for these scores. RIBES is an automatic evaluation measure based on word-order correlation coefficients between reference sentences and output sentences. Evaluation results are shown in Table 6. Bold numbers indicate that values are significantly higher than the result of BASELINE in each evaluation measure. To assess this, we used the bootstrap resampling test at a significance level of $\alpha = 0.01$

[7] http://taku910.github.io/mecab/

Table 6: Results of sentence-level automatic simplification

System	BLEU	RIBES
BASELINE	41.62	84.04
MOSES (distortion limit 0)	**46.06**	**85.41**
MOSES (distortion limit 6)	**46.02**	**85.39**

(Koehn, 2004).

MOSES (distortion limit 0), which learned simplification from the training data, obtained a BLEU score that is 4.4 points higher than that of BASELINE. This confirms the effectiveness of the data set for automatic simplification. MOSES (distortion limit 0) also obtained a RIBES score that is 1.3 points higher than that of BASELINE. In an experiment on English simplification using Wikipedia (Coster and Kauchak, 2011), the improvement of the BLEU score was 0.5 points. The improvement of the BLEU score using our data set is higher than theirs, indicating that our data set had a larger effect than their English Wikipedia data set. One of the reasons for this larger effect is thought to be that more words in simplified Japanese news were rewritten than those in the Simple English Wikipedia data set because the BLEU score (41.62) of Japanese news BASELINE is lower than the BLEU score (59.37) of English Wikipedia BASELINE. The training data size is smaller than for parallel corpora often used in SMT experiments on translation between languages. When the training data are small, certain methods can improve the translation quality (Xiang et al., 2010; Irvine, 2013; Irvine and Callison-Burch, 2014). Such methods will be useful for our Japanese news simplification task.

When MOSES (distortion limit 6), which allows phrase reordering, is compared with MOSES (distortion limit 0), which does not allow phrase reordering, the automatic scores did not improve. However, this result does not ensure that phrase reordering is unnecessary because there is room for improvement of the RIBES scores. We cannot know what types of word reordering are needed based solely on these results. Therefore, in the next section we investigate what types of rewrites are needed for simplification.

6 Analysis of Manually Simplified News Sentences

To reveal what types of rewrites are needed for automatic Japanese news simplification, we analyzed what types of rewrites, such as those associated with word order (i.e., syntactic structure), were conducted by Japanese instructors in the manual simplification processes. For this analysis, we produced manually annotated word alignments for 50 parallel article pairs. The 50 article pairs include 309 source sentences and 530 target sentences. As explained in Section 4.1, news reporters sometimes simplify expressions, despite it not being their main role. Thus, when analyzing the simplification of expressions by comparing sentences before and after simplification by Japanese instructors, if we use sentences that were rewritten by news reporters as pre-simplification sentences, then parts of the simplification process may go undetected. This is because it is possible that a certain degree of expression simplification has already been conducted before being simplified by the Japanese instructors. Therefore, to exhaustively detect the simplification of expressions, we only used simplified news articles that Japanese instructors had rewritten prior to being rewritten by news reporters.[8] Here, articles prior to being rewritten by Japanese instructors are called ORG, and articles that have been rewritten by such instructors are called EASY.

[8] As explained in footnote 3, recent articles were first rewritten mostly by news reporters and Japanese instructors first rewrote around half of the articles produced in the early phase. We wanted to conduct this analysis independently of the construction of the data set because we can then start analyzing when small data become available. For these reasons, we selected the 50 article pairs from the data produced in the early phase; that is, the 50 pairs are not a subset of the manually sentence-aligned article pairs in the data set described in Section 4.

Table 7: Rewrite categories associated with word order

Category	Frequency	Reordering distance type
Changing adnominal clauses to sentences	39	global
Reordering case elements	21	global
Reversed relation of modification	13	local
Complement for sentence splitting	13	global
Changing case of case elements	10	global
Indicating relation of continuous nouns	8	local
Changing compound nouns	7	local
Changing part of speech	7	local
Adnominal clause modifying formal nouns	5	global
Verbalizing nouns	4	global
Changing quantity expressions	3	global
Extraction of difficult expressions	3	global
Moving from EOS to BOS	3	global
Changing clause to noun modifier	2	local
Others	66	global/local

We then analyzed the simplifications associated with word order along with those not associated with word order.

6.1 Simplification Associated with Word Order

We categorized the simplification with respect to its association with word order for the sentences in the 50 article pairs. We disregarded expressions that were too largely summarized or too largely reorganized. Rewrite categories associated with word order are shown in Table 7. For the item of reordering distance type, "local" represents word reordering in a phrase pair or the reordering of contiguous phrases, and "global" represents longer word reordering compared with local, or the duplication of words. Explanations and examples of each category are shown in Appendix A. These results indicate that the most frequent rewrite type with word reordering is the extraction of adnominal clauses to become independent sentences. When adnominal clauses are extracted to become independent sentences, the syntactic structures of the resulting sentences become simpler. Thus, the effect of the extraction on readability is thought to be large. The second most frequent type, reordering case elements, does not result in a reduction in the complexity of syntactic structure. Thus, the effect of extraction of adnominal clauses on readability is thought to be larger than that of reordering case elements. Rewrites with local reordering also do not reduce the complexity of syntactic structures in many cases. From these analyses, the extraction of adnominal clauses to become independent sentences was found to be the most frequent and effective type of simplification with word reordering.

Although it is possible for phrase-based SMT to perform rewrites with local word reordering, converting adnominal clauses into independent sentences is difficult for phrase-based SMT because it requires long-distance word reordering and the duplication of the words modified by the adnominal clauses. Therefore, we believe that the following two-step conversion is suitable for use in automatic simplification. First, conversion with long-distance word reordering or word duplication, such as extracting adnominal clauses to become independent sentences, is conducted using syntactic structures and conversion rules. Second, sentences are simplified using phrase-based SMT.[9]

[9]The idea of splitting the process into two steps is the same as that of pre-ordering in SMT (Xia and McCord, 2004).

Table 8: Rate of tokens (words) in EASY that were unchanged, changed, or added

	Rate
Unchanged (identical) tokens	0.56
Changed (different) tokens	0.37
Added tokens	0.07

Table 9: Changing rates for passive voice expressions and causative expressions

	Change rate (Frequency)
Passive voice to active voice	0.92 (139/151)
Causative to non-causative	1.0 (11/11)

Table 10: Number of sentences produced from one sentence in simplification

Number of sentences (source–target)	Rate
1–1	0.40
1–2	0.46
1–3	0.11
1–4 or more	0.02

6.2 Simplification Not Associated with Word Order

We also analyzed the word rewrite rates, the rewrite rates of passive voice and causative expressions, and sentence splitting.

We checked the rates of tokens (words) in EASY that were unchanged, changed, or added by Japanese instructors.[10] The results are shown in Table 8. The results indicate that 44% of the tokens in EASY were either changed or added. We also checked the rate of omitted tokens in ORG. It was 0.08.[11,12]

We checked the rate of passive voice expressions that were changed into active voice expressions, as well as the rate of causative expressions that were changed into non-causative expressions.[13,14] The results are shown in Table 9. It was found that most of the passive voice expressions were changed into active voice expressions, and that most of the causative expressions were changed into non-causative expressions.

The number of sentences produced from a single sentence during simplification are shown in Table 10. Approximately half of the sentences were split into two separate sentences. The most frequent cause of sentence splitting was the changing of continuous clauses into independent sentences without word reordering, and the second most frequent was the changing of the above-mentioned adnominal clauses into independent sentences. The rate of sentence splitting is larger than the rate shown in Table 2. The main reason for this result is thought to be the fact that the pre-simplification sentences were different. The current analysis used original sentences as the pre-simplification sentences, whereas the analysis of Table 2 mainly used sentences that

[10] We used MeCab with an IPA dictionary as the morphological analyzer.

[11] Expressions that do not contain main content may be dropped. [E.g. 20] in Appendix A is such an example.

[12] The lead sentence, which is the first sentence in a news article that describes a summary of the news, may be dropped because its content overlaps with the text of the main body. To check the rate of dropped words under the condition in which the reason for deletion was not overlapping, we checked the rate of omitted tokens using all sentences except the lead sentences.

[13] We only checked causative expressions in which the base forms were *seru* or *saseru* and the part of speech was verb-postfix.

[14] [E.g. 6] in Appendix A is an example of passive voice being changed into active voice. [E.g. 1] in Appendix A is an example of causative being changed into non-causative.

had already been rewritten to some degree by news reporters as the pre-simplification sentences.

7 Conclusion

We designed a Japanese news sentence-level simplification task and constructed a Japanese news simplification data set, which consisted of Japanese news sentences and their corresponding simplified sentences. This is, to the best of our knowledge, the first time a parallel corpus consisting of sentences sourced from Japanese news and their simplified Japanese counterparts has been constructed. We verified the effectiveness of the constructed data set through preliminary experiments on automatic simplification using phrase-based SMT, and we confirmed that our result was 4.4 BLEU points higher than that of the baseline, which does not change the input sentences. Additionally, we produced manually annotated word alignments between parallel sentences to analyze the human operation of simplifying expressions, and provided what types of rewrites, such as rewrites associated with word order, were conducted in the manual simplification processes. The constructed data set and the knowledge obtained by our analysis of manual simplification will be useful for future research on Japanese news simplification and will serve as assistance in the production of simplified Japanese news.

References

Aluísio, S. M., Specia, L., Pardo, T. A., Maziero, E. G., and Fortes, R. P. (2008). Towards Brazilian Portuguese automatic text simplification systems. In *Proceedings of the Eighth ACM Symposium on Document Engineering*, DocEng '08, pages 240–248.

Bott, S., Rello, L., Drndarevic, B., and Saggion, H. (2012). Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proceedings of COLING 2012*, pages 357–374.

Bott, S. and Saggion, H. (2011). An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 20–26.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 1041–1044.

Coster, W. and Kauchak, D. (2011). Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). Text simplification for reading assistance: A project note. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 9–16.

Irvine, A. (2013). Statistical machine translation in low resource settings. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 54–61.

Irvine, A. and Callison-Burch, C. (2014). Hallucinating phrase translations for low resource MT. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 160–170.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1537–1546.

Klaper, D., Ebling, S., and Volk, M. (2013). Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 388–395.

Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 489–492.

Moku, M. and Yamamoto, H. (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.

Narayan, S. and Gardent, C. (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.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Petersen, S. E. and Ostendorf, M. (2007). Text simplification for language learners: a corpus analysis. In *Proceedings of Workshop on Speech and Language Technology for Education*.

Seretan, V. (2012). Acquisition of syntactic simplification rules for French. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 4019–4026.

Tanaka, H., Mino, H., Kumano, T., Ochi, S., and Shibata, M. (2013). News service in simplified Japanese and its production support systems. In *Proceedings of the IBC2013 Conference*.

Woodsend, K. and Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 1015–1024.

Xia, F. and McCord, M. (2004). Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of Coling 2004*, pages 508–514.

Xiang, B., Deng, Y., and Zhou, B. (2010). Diversify and combine: Improving word alignment for machine translation on low-resource languages. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 22–26.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1353–1361.

Appendix A. Explanations and Examples of Each Category in Table 7

Here we explain the categories associated with word order in Table 7 and show their examples. Red and blue are used for the parts reordered based on the factor of each category.

Changing adnominal clauses to sentences This refers to cases in which adnominal clauses are extracted as independent sentences.

[E.g. 1] NASA は去年１１月に**打ち上げた火星探査機「キュリオシティー」**を日本時間の来月６日、午後２時半過ぎに火星に着陸させます。

↓

NASA は去年１１月に**「キュリオシティー」という火星探査機を打ち上げました**。
「キュリオシティー」は、日本時間の来月６日、午後２時半すぎに、火星に着きます。

[E.g. 2] この遺伝子をマウスの脳の**記憶などをつかさどる海馬という部分に大量に組み込みました**。

↓

その遺伝子をマウスの脳の中の**海馬という部分にたくさん入れました。**
海馬は記憶などをコントロールする働きがあります。

Reordering case elements This refers to cases in which the order of the case elements is changed.

[E.g. 3] **富士山が大規模に噴火した場合**、**山梨県は** …
→ **山梨県は**、**富士山が大規模に噴火した場合** …

Reversed relation of modification This refers to cases in which the word reordering is caused by changing words into other words with reversed relations of modification.

[E.g. 4] **半年余り**で → **約半年**で

Complement for sentence splitting This refers to cases in which the case elements are replicated to complement split sentences when continuous clauses are changed into independent sentences.

[E.g. 5] **この有料サービスは**、**… の現在地を地図で把握できるというもので**、１０日から運用が始まりました。

↓

この有料サービスは、**… が今いる場所を地図で知ることができるというものです**。
このサービスは１０日から始まりました。

Changing case of case elements This refers to cases in which the word reordering is caused by changing the case of the case elements.

[E.g. 6] 笑った顔に見える埴輪が、パリで来月開かれる展覧会で展示されることになり、
↓
パリで来月開かれる展覧会に笑った顔に見える埴輪を出すことになりました。

[E.g. 7] 難しい役柄を表現力豊かに演じ → すばらしい表現力で難しい役を演じて

Indicating relation of continuous nouns This refers to cases in which the word reordering is caused by indicating the relations of continuous nouns.

[E.g. 8] 火星探査機「キュリオシティー」 →「キュリオシティー」という火星探査機

Changing compound nouns This refers to cases in which the word reordering is caused by changing compound nouns.

[E.g. 9] 家庭の電力消費 → 家庭で使う電力

[E.g. 10] 新たな被害想定を → 被害について、新しい予想を

Changing part of speech This refers to cases in which the word reordering is caused by changing the part of speech of words and their modifying points.

[E.g. 11] 具体的な場所を尋ね → 場所を細かく聞くと

Adnominal clause modifying formal nouns This refers to cases in which the verbs in adnominal clauses modifying formal nouns move backward.

[E.g. 12] 研究を行ったのは、 ••• のグループです。
→ この研究は ••• のグループが行いました。

Verbalizing nouns This refers to cases in which the word reordering is caused by verbalizing nouns (*sahen* nouns).

[E.g. 13] 着陸まであと２０日余り → 約２０日で着きます。

Changing quantity expressions This refers to cases in which the word reordering is caused by changing quantity expressions.

[E.g. 14] 避難が必要な人数の試算を進め → どのくらいの人が避難する必要があるか計算を進め

Extraction of difficult expressions This refers to cases in which difficult expressions are extracted and explained as independent sentences.

[E.g. 17] と述べ、異例の謝罪を行いましたが、
→ と言って謝りました。このようなことは今までにありませんでした。

Moving from EOS to BOS This refers to cases in which parts of expressions at the end of sentences move to the beginning of the sentence.

[E.g. 15] ••• に提供することを明らかにし、 → 発表によると、 ••• に渡したりしました。

[E.g. 16] ••• を考えているという調査結果がまとまりました。
→ その結果、 ••• を考えてみようとしていることが分かりました。

Changing clause to noun modifier This refers to cases in which the word reordering is caused by changing clauses to noun modifiers.

[E.g. 18] １１年８か月ぶりの円高水準を更新しました。
→ ２０００年１１月から今までで いちばんの円高です。

Others This refers to cases for which categorization is difficult because rewrites are complex or they require background information, or when the frequency of the category is one.

[E.g. 19] IT **企業の間では**、**この分野を強化する**動きが広がっています。
→ **地図に力を入れる** IT **企業が**増えています。

[E.g. 20] **電力会社が提供する**需給状況のデータに基づいて、**電力需要が少なく**価格が安い**時間**帯に
→ **電力が足りている時間**は、**電力会社の**データから分かります。