Crowdsourced Corpus of Sentence Simplification with Core Vocabulary

Akihiro Katsuta, Kazuhide Yamamoto
Nagaoka University of Technology
1603-1, Kamitomioka Nagaoka, Niigata 940-2188, JAPAN
{katsuta,yamamoto}@jnlp.org

Abstract

We present a new Japanese crowdsourced data set of simplified sentences created from more complex ones. Our simplicity standard involves all rewritable words in the simplified sentences being drawn from a core vocabulary of 2,000 words. Our simplified corpus is a collection of complex sentences from Japanese textbooks and reference books together with simplified sentences generated by humans, paired with data on how the complex sentences were paraphrased. The corpus contains a total of 15,000 sentences, in both complex and simple versions. In addition, we investigate the differences in the simplification operations used by each annotator. The aim is to understand whether a crowdsourced complex-simple parallel corpus is an appropriate data source for automated simplification by machine learning. The results, that there was a high level of agreement between the annotators building the data set. So, we believe that this corpus is a good quality data set for machine learning for simplification. We therefore plan to expand the scale of the simplified corpus in the future.

Keywords: Corpus, Crowdsourcing, Simplification

1. Introduction

Simplification task is the process of rewriting a complex text into a simpler form while preserving its meaning. Simplified texts play an important role in providing accessible and easy-to-understand information for a wide range of users who find it difficult to understand texts that have not been simplified due to their linguistic complexity. Attempts have been made to automate the simplification process for various languages, including English, Spanish, Brazilian Portuguese, and Portuguese. We have been conducted research on simplification since a few years ago (Moku et al., 2012). Recent studies have treated text simplification as a monolingual machine translation problem in which simple and synonymous sentences are generated using statistical machine translation (Wubben et al., 2012; Xu et al., 2016).

As with statistical machine translation using bilingual parallel corpora, text simplification therefore requires a monolingual parallel corpus for training. In the case of English, there are PWKP (Zhu et al., 2010) and Wikipedia Datasets (Coster and Kauchak, 2011; Kauchak, 2013) made from wiki and Swiki. In German, a simplification corpus with a scale of 7,000 sentences (Caseli et al., 2009), In Italian the PaCCSS-IT (Brunato et al., 2016) with a scale of 63,000 sentences, In Spanish a simplification corpus do exist made by hand by the few rules (Mitkov and Štajner, 2014; Štajner et al., 2015), and so on.

However, only English corpora are publicly available, such as a large-scale simplified English corpus obtained from pairs of Wikipedia and Simple English Wikipedia entries. For Japanese, there is no other than our simplification corpus (Maruyama and Yamamoto, 2018). We have already conducted the experiment of simplification using corpus (Maruyama and Yamamoto, 2017). Currently, most of the simplification research that has been done for Japanese has emphasized paraphrasing word units (Kajiwara et al., 2013; Kajiwara and Yamamoto, 2015; Hading et al., 2016). For that reason, we consider building a simplified corpus to be the most important task for studying automated simplification of Japanese.

In this paper, we therefore present a human-generated simplified corpus where the simplification operations are based on simple vocabulary restriction rules. Recent research on corpus building has shown that simplification processes based on short lists of simple rules are more time efficient and consistent (Mitkov and Štajner, 2014).

In order to express all the usual things of daily conversation level, we set a frame of total number in advance, and artificially selected what we call core vocabulary. We unify these simplification criteria by defining a core vocabulary of 2,000 words and asking annotators to simplify complex sentences into plainer ones using only these 2,000 words. Our simplification is aimed at vocabulary compression of sentences as we usually use everyday. Therefore, we do not consider other simplification parameters such as difficult grammar and sentence length (E.g. simplification including summarization tasks such as Simple Wikipedia is not covered). Also, since we are interested in the approaches different annotators take to the simplification process, we investigate the differences in the simplification operations used for the simplified corpus.

2. Experimental Design

2.1. Core Vocabulary

We chose 2,000 words, based on the UniDic[1] word units that preserved the most meaning for sentences in the Tanaka corpus[2], and defined these words as the core, or simple Japanese, vocabulary. The Tanaka corpus of Japanese-English parallel corpus that was translated as part of the classwork by the Japanese college students and contains

[1] https://ja.osdn.net/projects/unidic
[2] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus

	Version		Example
(a)	Original		彼女は、あなたが考えているような 女の子 ではない。
	Simplified	(1)	彼女は、あなたが考えているような 女 ではない。
		(2)	彼女は、あなたが考えているような 女性 ではない。
		(3)	彼女は、あなたが考えているような 少女 ではない。
(b)	Original		このスープは ほんのわずか 塩が たりない。
	Simplified	(1)	このスープは 本当に少しだけ 塩が 不足している。
		(2)	このスープは ちょっとだけ 塩が 不十分だった。
		(3)	このスープは 少し 塩が 少ない。
(c)	Original		今日私は道で見つけた キー を 拾い上げた。
	Simplified	(1) - (3)	今日私は道で見つけた 鍵 を 拾った。

Table 1: Example of simplification sentences in the corpus

150 thousand sentences. It is many of the sentence pairs have been derived from textbooks. Therefore, we consider that many the sentences of the daily conversation level are included. And it is also possible to combine simple sentences with English sentences.
This core vocabulary had the following features.

1. It consisted mainly of simple, frequently used words.

2. It also included some words that did not meet the first condition, but were necessary to explain certain concepts. (e.g. red, green, and blue were necessary to explain the concept of color.)

Symbols (such as punctuation marks), unique words (such as proper nouns), and words that were not present in UniDic (such as English words) were excluded from the core vocabulary.

2.2. Simplification Task

We used "CrowdWorks[3]," the crowdsourcing platform in Japan to gather Japanese workers. The purpose of this research is to build a human-generated simplified corpus. The text that we targeted for simplification consisted of the Japanese sentences in the Tanaka corpus. We used crowdsourcing to take complex source sentences containing words outside the core vocabulary, and translate them into simple sentences. We extracted sentences of between 7 and 65 words from the Tanaka corpus and simplified them using crowdsourcing. A total of 34,300 sentences were divided into seven parts, and these parts were then assigned to different annotators for simplification (4,900 sentences each). For evaluation, we additionally asked them to simplify the same 100 sentences. The rules we requested the annotators to use for simplification were as follows.

1. Translate each complex sentence using only the core vocabulary, preserving as much of its meaning as possible.

2. Do not paraphrase words that have been excluded from the core Vocabulary (see Section 2.1.), or the target word when explaining the meaning of a word in a sentence.

[3]https://crowdworks.jp/

	Original	Simplified
#Sentences	(4,900)34,300	(4,900)34,300
Vocabulary size	(5,846)14,429	(2,084) 3,649
Avg. sentence length	19.18	21.16

Table 2: Statistics for the simplified corpus (In parentheses are the average for each annotator)

	Original	Simplified
#Sentences	100	(100) 700
Vocabulary size	397	(398) 681
Avg. sentence length	16.56	19.41

Table 3: Statistics for the evaluation simplified corpus (In parentheses are the average for each annotator)

Ideally, machine translation systems should be trained on a corpus built by a small number of annotators, to prevent annotation variability. We therefore asked the annotators to translate most of the sentences by themselves.

3. Simplification Corpus

Major phenomena of simplification in the evaluation corpus is shown in Table 1. The same optimal word may not always be substituted, because the annotator may not know all 2,000 words and there are multiple ways of expressing many concepts. Despite having asked the annotators to make substitutions using the core words, there were instances where, for example, "女の子 (girl)" was replaced by "女 (female)" or "女性 (woman)" rather than "少女 (girl)", as shown in (a) of the table, it is observed that the optimal word is not always chosen for substitution, because the annotator may not know it is in the core words. In example (b), all the annotators paraphrased the sentences to preserve its meaning, but they used different words. Example (c) shows an instance where all three annotators simplified a sentence in the same way.
Overall, 34,300 complex sentences using a vocabulary of 14,429 words, were selected for simplification. After simplification, the new corpus used a vocabulary of only 3,649 words (in Table 2). So, there are 1,649 words excluded from substitution in the 34,300 sentences. And, when 7

BLEU	Frequency
[0.0 , 0.2)	0
[0.2 , 0,3)	4
[0.3 , 0.4)	2
[0.4 , 0.5)	8
[0.5 , 0.6)	7
[0.6 , 1.0]	0

Table 4: Frequency of inter-agreement annotator

at annotators each simplified 100 sentences, the number of vocabulary increased from 397 to 681 (in Table 3). We consider that the reason for the increase is due to the difference of simplified such as (a) or (b) in Table 1.
The length of the simplified sentence is longer in both Table 2 and 3, so we can see that sentences tend to be longer if they paraphrase without losing their meaning. For example in the following cases:

Original:
“そこに署名してください。”
Simplified:
“そこに名前を書いてください。”
“そこにあなたの名前を書いてください。”

Original:
“彼はこのビジネスで名声を築いた。”
Simplified:
“彼はこの仕事で有名になった”
“彼はこの仕事で社会的に評価され尊敬を集める立場を作った。”

Although the latter is closer meaning, we can see that a sentence is longer.
In the next section, we investigate the 100 sentences of evaluation corpus that were given to all annotators.

3.1. Agreement between Annotators

3.1.1. Automated Evaluation

We assessed the level of agreement between annotators to analyze annotation variability. We computed this in terms of the BLEU score in the same way as (Mitkov and Štajner, 2014). This looks at the level of agreement between the simplifications, with higher values indicating more similar results. Because annotator is 7 people, BLEU between 21 patterns annotator was calculated. Table 4 shows the level of inter-annotator agreement in terms of the BLEU scores. Most of BLEU is distributed between 0.4 and 0.6, and there are a few around 0.3.
Mitkov and Štajner (2014) built a simplified corpus based on simple rules, finding that the BLEU scores of their three annotators were between 0.44 and 0.53. The BLEU of our simplified corpus (Maruyama and Yamamoto, 2018) range between 0.58 to 0.63. Since the corpus we created is more controllable, it is highly agreement. In crowdsourcing, since difference how much preserve meaning for each annotator, the case that BLEU become low occurs. However, it is considered that the quality of simplification corpus is high because about 70% of inter-agreement exceeds 0.4.

3.1.2. Manual Evaluation

We also analyzed the quality of the simplified sentences and the simplification operations applied by each annotator. We used the frequencies with which the simplification operations were selected and the qualities of the resulting sentences as indicators of agreement between annotators.
First, we analyzed the simplification operations applied by the annotators to each of the 100 common sentences. We did not count the number of operations applied in a sentence, but just whether or not it was applied at least once. The simplification operations were mainly classified into the following four types:

1. Word substitution (WS)
   A word is paraphrased by a synonym.
   Original:
   “彼から ずいぶん 長い間 便り がない。”
   Simplified:
   “彼から かなり 長い間 手紙 がない。”

2. Phrase substitution (PS)
   Two or more consecutive words are paraphrased.
   Original:
   “私たちのところに、不意の来客 が**あった**。”
   Simplified:
   “私たちのところに、突然訪ねてきた客 が**いた**。”

3. Deletion (D)
   A word is removed from the original sentence.
   Original:
   “彼は私の肩を いっぱつ 打った。”
   Simplified:
   “彼は私の肩を打った。”

4. Insertion (I)
   A word is added to the original sentence.
   Original:
   “冬休みはどのように過ごしましたか。”
   Simplified:
   “冬 の 休みはどのように過ごしましたか。”

Second, we opted to manually evaluate the quality of each sentences, in addition to the above analysis. Following the criteria in Table 5, we asked human evaluators to assess, on a scale of 1-4 (where higher marks denote better sentences), two aspects of the presented sentences: grammaticality (G) and meaning preservation (M).
The resulting numbers of simplification operations, together with the average evaluations from five annotators, are shown for each annotator in Table 6.
The total number of simplification operations for each annotator is not 100, because in some cases the annotators performed several different operations on a single sentence. The most common operation we observed was replacing a word/phrase in the original sentences with a word/phrase from the core vocabulary. In addition, the grammar quality and level of meaning preservation were both high overall, and exhibited little variation between annotators, so the simplification process can be regarded as both consistent and reliable.

Grammar (G)	
Evaluation	Criterion
1	The sentence does not make any sense at all.
2	The sentence is hard to understand due to grammatical mistake.
3	The sentence is fairly good except for minor mistakes.
4	The sentence is free from grammatical mistakes.
Meaning (M)	
Evaluation	Criterion
1	The meaning is unrelated to that of the original sentence.
2	The meaning is related, but the original sentence cannot be guessed.
3	The meaning of the sentence is roughly the same, but it is a little ambiguous.
4	The sentences have the same meaning.

Table 5: Criteria for evaluating the simplified sentences

	WS	PS	D	I	G	M
(1)	60	52	3	5	3.75	3.57
(2)	74	41	2	5	3.87	3.65
(3)	68	39	11	5	3.92	3.57

Table 6: Simplification operation frequencies and human evaluation results (the "G" and "M" columns show the mean grammaticality, and meaning preservation scores, respectively)

4. Conclusions and Future Work

In this paper, we have created a simplified corpus for automated simplification research by crowdsourcing. This is a large simplification corpus, suitable for machine learning, that was produced by a small number of annotators and which shows low annotation variability. Examining the level of agreement between annotators, we found almost high agreement in either the manual or automated evaluations, and therefore regard, this corpus as consistent and reliable. In addition, we found that limiting the core vocabulary to 2,000 words was advantageous for controlling annotation variability. This work also demonstrates that a high-quality simplification corpus can easily be built by crowdsourcing.

In the future, we aim to expand this simplification corpus and use it to create an automated simplification system by machine learning. This corpus will be released.

Acknowledgments

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

References

Brunato, D., Cimino, A., Dell'Orletta, F., and Venturi, G. (2016). Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A., Gasperin, C., and Aluísio, S. M. (2009). Building a brazilian portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science*, 41:59–70.

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: short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.

Hading, M., Matsumoto, Y., and Sakamoto, M. (2016). Japanese lexical simplification for non-native speakers. *NLPTEA 2016*, page 92.

Kajiwara, T. and Yamamoto, K. (2015). Evaluation dataset and system for japanese lexical simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40.

Kajiwara, T., Matsumoto, H., and Yamamoto, K. (2013). Selecting proper lexical paraphrase for children. In *Proceedings of The 25th Conference on Computational Linguistics and Speech Processing*, pages 59–73.

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 (volume 1: Long papers)*, volume 1, pages 1537–1546.

Maruyama, T. and Yamamoto, K. (2017). Sentence simplification with core vocabulary. *Proceedings of the International Conference on Asian Language Processing*, pages 363–366.

Maruyama, T. and Yamamoto, K. (2018). Simplified corpus with core vocabulary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'18)*.

Mitkov, R. and Štajner, S. (2014). The fewer, the better? a contrastive study about ways to simplify. In *Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 30–40.

Moku, M., Yamamoto, K., and Makabi, A. (2012). Automatic easy japanese translation for information accessibility of foreigners. In *Proceedings of Coling-2012 Workshop on Speech and Language Processing Tools in Education*, pages 85–90.

Štajner, S., Calixto, I., and Saggion, H. (2015). Automatic text simplification for spanish: comparative evaluation of various simplification strategies. In *Proceedings of the international conference recent advances in natural language processing*, pages 618–626.

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: Long Papers-Volume 1, pages 1015–1024. Association for Computational Linguistics.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

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. Association for Computational Linguistics.

Simplification corpus	
Original	ナンシーは最初に来た 女の子 だった。
	(Nancy was the first girl to come.)
Simplified	ナンシーは最初に来た少女だった
Original	追って 通知 があるまで 会合 は 延期 された。
	(The meeting was put off till further notice.)
Simplified	追って連絡があるまで会の予定は遅れることになった。
Original	ドアが パッ と 勢い 良く あい た。
	(The door burst open.)
Simplified	ドアが突然開いた。
Original	この 騒音 は我慢出来ない。
	(I cannot put up with this noise.)
Simplified	このようにうるさい所にはいられない。
Original	いやぁ、昨日は 入れ食い でねえ。
	(They were biting like crazy yesterday.)
Simplified	いや、昨日はたくさん魚が手に入ってねえ。
Original	例の スキャンダル はそういつまでも 臭い ものに フタ と いう わけにはいくまい。 いずれ 人は 嗅ぎつけ てしまうさ。
	(I don't think we can keep the lid on the scandal much longer; people are bound to find out.)
Simplified	例のあまり良くない話はそういつまでも無かったことにはできまい。 やがて人は気づいてしまうさ。
Evaluation simplifie corpus	
Original	明日の午後、いつでもお 出で ください。
	(Come to see me at any time tomorrow afternoon.)
Simplified	明日の午後、いつでも来てください。 (7)
Original	近頃 は いかが ですか。
	(How are you these days?)
Simplified	最近はどうですか。(4) 調子はどうですか。 最近は、調子はどうですか。 最近はどのようにお過ごしですか。
Original	彼はこの ビジネス で 名声 を 築い た。
	(He worked up a good reputation through this business.)
Simplified	彼はこの仕事で有名になった。(3) 彼はこの仕事で名を上げた。 彼はこの事業で有名になった。 彼はこの仕事で高い評価を手にした。 彼はこの仕事で社会的に評価され尊敬を集める立場を作った。
Original	彼から ずいぶん 長い間 便り がない。
	(I haven't heard from him for ages.)
Simplified	彼からかなり長い間手紙がない。 彼からかなり長い間手紙が来ない。 彼からかなり長い間連絡がない。 彼から長い間連絡がない。 彼からとても長い間連絡がない。 彼らからは長い間手紙が来ていない。 彼から非常に長い間、手紙が来ない。
Original	彼らは 雇い主 に 忠実 だ。
	(They are loyal to their master.)
Simplified	彼らは使用人に誠実によく働く人だ。 彼らは主人を信頼して従っている。 彼らは上司に従う。 彼らは働いているところの主人によく従っている。 彼らは使用者の言うことによく従う。 彼らは社長の命令をよく聞く。 彼らは経営者に大事にされている。

Table 7: Example of Simplification Corpus. The underline words in the original sentences are complex words. The numbers are the number of same simplified sentences.