Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6744–6752
Marseille, 20-25 June 2022

A Benchmark Dataset for Multi-Level Complexity-Controllable Machine Translation

Kazuki Tani[1], Ryoya Yuasa[1], Kazuki Takikawa[2], Akihiro Tamura[1], Tomoyuki Kajiwara[2], Takashi Ninomiya[2], Tsuneo Kato[1]
[1]Doshisha University, [2]Ehime University
{ctwh0176@mail4, ctwh0190@mail4, aktamura@mail, tsukato@mail}.doshisha.ac.jp
{takikawa@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp

Abstract

This paper introduces a new benchmark test dataset for multi-level complexity-controllable machine translation (MLCC-MT), which is an MT that controls the output complexity at more than two levels. In previous studies, MLCC-MT models have been evaluated on a test dataset automatically generated from the Newsela corpus, which is a document-level comparable corpus with document-level complexity. There are three issues with the existing test dataset: first, a source language sentence and its target language sentence are not necessarily an exact translation pair because they are automatically detected. Second, a target language sentence and its simplified target language sentence are not always perfectly parallel since they are automatically aligned. Third, a sentence-level complexity is not always appropriate because it is derived from an article-level complexity associated with the Newsela corpus. Therefore, we created a benchmark test dataset for Japanese-to-English MLCC-MT from the Newsela corpus by introducing an automatic filtering of data with inappropriate sentence-level complexity, manual check for parallel target language sentences with different complexity levels, and manual translation. Furthermore, we implement two MLCC-NMT frameworks with a Transformer architecture and report their performance on our test dataset as baselines for future research. Our test dataset and codes are released.

Keywords: Machine Translation, Natural Language Generation, Corpus (Creation, Annotation, etc.)

1. Introduction

In recent years, neural machine translation (NMT), which translates an input sentence (a source language sentence) into a sentence in the target language (a target language sentence) using neural networks, has become increasingly developed and widespread to a wide range of users. While conventional NMT uniformly generates a target language sentence for any user and any situation, style-controllable NMT, which controls the style of output rather than its content, has recently received a lot of attention. Sennrich et al. (2016), for example, proposed an NMT model that controls the politeness of a target language sentence through the incorporation of special tokens representing politeness, and Kuczmarski et al. (2018) proposed a method to control the gender of pronouns in target language sentences. Schioppa et al. (2021) proposed a method to control multiple attributes of MT outputs. Furthermore, complexity-controllable NMT (CC-NMT), which controls the complexity of a target language sentence to allow translation tailored to the user's reading level, has been attracting much attention. Most of the existing CC-NMT models control the complexity of a target language sentence at two levels (i.e., complex and simple), which we refer to as two-level complexity-controllable NMT (2LCC-NMT). In the previous research, English-European language pairs have been primarily focused on. For example, Marchisio et al. (2019) have proposed a 2LCC-NMT model between Spanish and English. Other than English-European language pairs, Maruyama and Yamamoto (2018) and Katsuta and Yamamoto (2018) have created and released simplified sentences of Japanese sentences in the English-Japanese parallel corpus, which can be used as a dataset for English-Japanese 2LCC-MT.

Recently, Agrawal and Carpuat (2019) have proposed a multi-level complexity-controllable NMT (MLCC-NMT) model , which controls the complexity of a target language sentence at three or more levels. Since it is more flexible than 2LCC-NMT, MLCC-NMT is expected to be further developed and may be more suitable for applications, such as foreign language learning and cross-lingual document processing. However, there are few resources available for training and evaluating MLCC-MT models. Agrawal and Carpuat (2019) have used the dataset automatically constructed from the Newsela corpus,[1] which contains news articles in English and Spanish written at multiple complexity levels, for training and evaluating their models by using Spanish-to-English machine translation and automatic sentence alignment in English.[2]

The automatically constructed dataset has the following three problems: first, a source language sentence and its target language sentence are not necessarily an exact translation pair because sentence alignment

[1]https://newsela.com/data/

[2]Although the paper (Agrawal and Carpuat, 2019) does not explicitly mention the test data, the automatically constructed data is considered to be used both for training and testing.

Japanese Sentence	Complexity	English Sentence
今では、画期的な学術研究が彼女を裏付けているようです。	12	Now, landmark academic research appears to back her up.
	8	Now, important academic research appears to back her up.
	4	Now, an important study appears to back her up.
公衆衛生についての入門書が必要だ。	12	A primer about public health is in order.
	9	A short explanation about public health is in order.
	6	A short explanation about public health is needed.

Table 1: Samples in Our Test Dataset for Japanese-to-English MLCC-MT

across languages is automatic. Second, since the sentence alignment is automatic in the target language, a target language sentence and its simplified target language sentence are not always exactly parallel. Third, a sentence-level complexity is not always appropriate because it is transferred from an article-level complexity attached to the Newsela corpus. In general, because training data can contain some noise, such data could be used as a training dataset. However, because a test dataset should be able to precisely measure model performance, the data with the aforementioned issues are insufficient for a test dataset.

Therefore, in this study, we build a test dataset to properly assess the performance of an MLCC-MT model. In particular, our creation procedure manually translates English sentences in the Newsela corpus to achieve correct translation pairs. Furthermore, our procedure introduces an automatic filtering of data with inappropriate sentence-level complexity and manual check for parallel target language sentences with different complexity levels to obtain proper counterparts written at multiple complexity levels. In this study, we focus on the Japanese-English language pair as one case of language pairs not included in the Newsela corpus. As a result, we create a benchmark test dataset for Japanese-to-English MLCC-NMT, consisting of 1,014 sets of a Japanese sentence and its English sentences written at multiple complexity levels. Table 1 displays samples in our test dataset.

Furthermore, this study implements two MLCC-NMT frameworks, a pipeline framework and multi-task framework, proposed by Agrawal and Carpuat (2019), with a SOTA architecture (i.e., Transformer) and assesses their performance on our test dataset. We release our test dataset and the codes of the implemented models so that they can be used as a benchmark for future research on MLCC-MT.[3]

2. Related Work

Section 2.1 describes the existing research on monolingual text simplification. Section 2.2 overviews the existing studies on 2LCC-NMT. Finally, Section 2.3 describes the existing study on MLCC-NMT.

[3]Our test dataset and the codes are available at https://github.com/K-T4N1/A-BenchmarkDataset-for-ComplexityControllableNMT.git.

2.1. Text Simplification

Text simplification, which converts an input sentence into a simplified sentence in the same language, has been extensively studied in the field of NLP. Scarton and Specia (2018) proposed an English sentence simplification model in which a special token representing the level of simplicity is added to an input sentence, and Kato et al. (2020) proposed a BERT-based model in which predicates in Japanese sentences are simplified. Other models have been proposed for Spanish (Štajner et al., 2015), Italian (Brunato et al., 2016), and German (Klaper et al., 2013). Surya et al. (2019) proposed an unsupervised text simplification model trained on unlabeled English Wikipedia text.

2.2. Two-Level Complexity-Controllable NMT

A 2LCC-NMT model is an NMT model that takes a source language sentence and one of the two complexity levels (i.e., "simple" and "complex") as input and then generates a target language sentence according to the input complexity level. There have been numerous models proposed for English-European language pairs. Marchisio et al. (2019), for example, proposed a 2LCC-NMT model between Spanish and English that includes a decoder for each complexity level: a simple-decoder and a complex-decoder.

Previous studies have attempted to build resources for 2LCC-MT because only small parallel corpora are available for 2LCC-MT, unlike standard MT. In the study by Maruyama and Yamamoto (2018), Japanese sentences of the Tanaka corpus, which is an English-Japanese parallel corpus, have been manually simplified by students. Through crowdsourcing, Katsuta and Yamamoto (2018) have expanded the corpus of Maruyama and Yamamoto (2018). As resources for English-Japanese 2LCC-MT, these corpora provide sets of triplets that include an English sentence, its Japanese sentence, and the simplified Japanese sentence. Marchisio et al. (2019) have extracted a set of simple Spanish-English parallel sentences and a set of difficult ones from the Newsela corpus by undersampling and oversampling, as a resource for Spanish-English 2LCC-MT.

Complexity	English Sentence
8	However, she says that there are times when "you just need to get away."
5	Yet she says that there are times when "you just need to get away."
3	Still, **Bopp** says that there are times when "you just need to get away."
12	The White House said the U.S. will suspend participation in preparatory meetings for the G-8 economic summit planned.
7	The White House said the U.S. will stop participating in planning meetings for the G-8 economic summit.
5	The White House said the U.S. will stop participating in meetings about the G-8 summit **in Russia**.

Table 2: Examples of Proper Noun Insertion

Complexity	English Sentence
12	**A company called AquaBounty has been seeking for more than 20 years to win FDA approval to bring a genetically modified fast-growing salmon to supermarkets.**
9	A company called AquaBounty has been seeking for more than 20 years to win Food and Drug Administration (FDA) approval to bring a genetically modified fast-growing salmon to supermarkets.
7	**A company called AquaBounty has been seeking for more than 20 years to win FDA approval to bring a genetically modified fast-growing salmon to supermarkets.**
12	So few Indians drink brewed coffee that virtually all its best crop is exported to countries such as Italy, where the beans are used in name-brand espresso blends and sold at a huge markup.
9	**There the beans are used in name-brand espresso blends and sold at a huge price increase.**
7	**There, the beans are used in name-brand espresso blends and sold for a huge price increase.**

Table 3: Examples of Sentence Sets with Improper Complexity Levels

2.3. Multi-Level Complexity-Controllable NMT

To the best of our knowledge, so far, MLCC-NMT has been studied only by Agrawal and Carpuat (2019). The details of their NMT models are described in Section 4.1. This section describes the dataset used in their work.

Their dataset was derived from the Newsela corpus, in which the Spanish articles correspond to sections of the English articles. Each article is assigned a "grade level," which represents the article's level of complexity. The grade level value ranges from 2 to 12, with a higher value indicating a more complex article.

English and Spanish sentences in the Newsela corpus are not aligned across languages. Therefore, they used Spanish-to-English MT and MASSAlign (Paetzold et al., 2017), which is a tool to detect parallel sentences in the same language, to create Spanish-English parallel sentences for diverse complexity levels (i.e., a dataset for MLCC-MT between English and Spanish). The creation procedure is as follows:

Step 1 translates Spanish articles into English by using Google Translate.[4]

Step 2 aligns sentences of English articles in the Newsela corpus and the translated English sentences by using MASSAlign.

Step 3 groups the aligned sentences, where the translated English sentences are replaced with their original Spanish sentences. The complexity level of each sentence is set to the grade level of the article to which the sentence belongs.

Agrawal and Carpuat (2019) used an automatically generated dataset to train and evaluate their MLCC-NMT models, but the datasets have three flaws.

Problem 1. Incorrect translation pairs: A source language sentence and its target language sentence are not necessarily an exact translation pair because they are automatically detected.

Problem 2. Difference of granularity of information among target language sentences with different complexity levels: A target language sentence and its simplified target language sentence are not always perfectly parallel since they are automatically aligned. Inserting specific contents, such as proper nouns, into a simpler sentence, for example, is problematic because it is difficult to generate new contents when controlling the complexity level. Table 2 shows examples of the insertion of proper nouns with bold fonts. The proper

[4] https://translate.google.com/

nouns appear in the low complexity sentences, although they do not appear in the high complexity sentences.

Problem 3. Incorrect sentence-level complexity: A sentence-level complexity is not necessarily appropriate because an article-level complexity (i.e., grade level in the Newsela corpus) is tagged as a sentence-level complexity. The complexity of a sentence dose not correspond to the grade level of an article. For example, even if the grade levels attached as sentence-level complexities differ, the sentences could be identical, or the difference could be only symbols. Table 3 shows examples of sentences with incorrect complexity levels. The bold sentences are exactly or almost the same; however, different complexity levels are assigned to them.

When the dataset is used as training data, the problems described above may have little impact; however, when used as test data, these problems cannot be ignored owing to their interference with the accurate evaluation of model performance. In this study, we solve the three problems listed above to create a test dataset for more accurately evaluating MLCC-MT.

3. Benchmark Test Dataset

This section explains the procedure for creating our benchmark test dataset for MLCC-MT and the details of the created test dataset. This study focuses on Japanese-to-English MLCC-MT, and our test dataset includes triplets of a Japanese sentence (a source language sentence) and its counterpart English sentences (target language sentences) written at three or more complexity levels. Note that our test dataset does not contain inappropriate instances shown in Tables 2 and 3.

Section 3.1 discusses the selection of the source corpus for our test dataset, and Section 3.2 describes the details of the proposed procedure for creating our test dataset. Further, in Section 3.3, we will discuss why the automatic filtering process introduced in the proposed procedure is necessary to improve the quality of our test dataset.

3.1. Selection of Source Corpus

The Newsela corpus was used to create our test dataset. This section explains the reason for selecting the Newsela corpus as the source for our test dataset.

A test dataset for MLCC-MT should comprise the triplets of a source language sentence and its target language sentences written at three or more complexity levels. There are two possible approaches to constructing such a dataset based on existing corpora: one is to translate sentences of an existing monolingual simplification corpus with multiple references, and the other is to simplify target language sentences of a bilingual parallel corpus at multiple levels. The former approach translates one sentence from an existing corpus to generate one test instance. The latter approach, in contrast, requires the creation of two or more new sentences written at different complexity levels for one test instance, which is more costly than the former approach. Therefore, in this study, we adopted the former approach, i.e., translation of an existing monolingual simplification corpus with multiple references.

Furthermore, there are crowdsourced monolingual simplification corpora, apart from the Newsela corpus, with multiple references, such as the Turk Corpus (Xu et al., 2016) and ASSET Corpus (Alva-Manchego et al., 2020). We selected the Newsela corpus as the source corpus for the following reasons. Since the Newsela corpus was created by professionals who write news articles rather than by crowdsourcing, the simplified texts of the Newsela corpus may be of higher quality and reliability than those of the other corpora. The Newsela corpus is larger than the other corpora. The multiple references of the corpora other than the Newsela corpus do not have a complexity level; and therefore, they are not straightforwardly converted into a dataset with multiple complexity levels.

3.2. Proposed Creation Procedure

The following two steps are used to create our test dataset:

Step 1 generates the sets of English sentences with the same content written at multiple complexity levels.
1-1: Extraction of English aligned sentences from the Newsela-auto corpus
1-2: Removal of inappropriate data by an automatic filtering
1-3: Manual check

Step 2 translates English sentences into Japanese.

Step 1 creates English sentence groups with the same content written at multiple complexity levels from English articles of the Newsela corpus. As in the previous study (Agrawal and Carpuat, 2019), we find such groups based on automatically aligned English sentence pairs. In particular, we use the Newsela-auto corpus (Jiang et al., 2020). It consists of 813,972 pairs of English sentences that appear Newsela articles of various grade levels and are automatically aligned by an aligner trained on the Newsela-manual dataset, a corpus of manually aligned English sentences. It is worth mentioning that the aligner used to generate the Newsela-auto corpus has been revealed to achieve better alignment performance than MASSAlign, used in the previous study (Agrawal and Carpuat, 2019). We create the sets of three or more aligned English sentences from the Newsela-auto corpus since our aim is to control multilevel (three or more level) complexities. As a result of Step 1-1, we obtained 603,785 English

# of Grade Levels	3	4	5
# of Instances	906	97	11
Percentage (%)	89.3	9.6	1.1

Table 4: The Number of Target-Side Grade Levels in a Test Instance of Our Dataset

sentence sets.[5]

Since the sentence pairs in the Newsela-auto corpus are automatically aligned, the English sentence sets, the results of Step 1-1, have the Problem 2 described in Section 2.3 (i.e., the difference in granularity of information between target language sentences with different complexity levels). To address this issue, in Step 1-3, we manually check the English sentence sets whether new contents, such as proper nouns, do not pop up and remove the sets where new content appears in a simpler sentence.

In Step 1, as in the previous study, the complexity level of each sentence is set to the grade level of the article to which the sentence belongs. As a result, the Problem 3 (i.e., incorrect sentence-level complexity) (described in Section 2.3) arises. To address the issues, the proposed creation procedure includes an automatic filtering (Step 1-2) and manual check (Step 1-3) to obtain three or more English sentences written at different levels of readability.

The automatic filtering in Step 1-2 removes sentence pairs that are the same after symbols[6] are removed as well as sentence pairs where the grade level difference is less than or equal to 1. The algorithm of automatic filtering is depicted in Algorithm 1. In the algorithm, $edit\_distance$ indicates Levenshtein edit-distance.

The manual check in Step 1-3 confirms whether the sentence with a higher complexity level is more complex. We used the sets of three or more English sentences after the filtering and manual check as the sets of target language sentences.

In Step 2, for each English sentence set generated in Step 1, the English sentence with the highest complexity level (i.e., the most complex sentence) is manually translated into Japanese by professional translators from a translation company, and the Japanese sentences are used as the source language sentences in our test dataset. As a result, we created a dataset for evaluating Japanese-to-English MLCC-MT using the proposed procedure, which consisted of 1,014 sets of a Japanese sentence and its English counterpart (i.e., English sentences written at multiple complexity levels).

Figure 1 and Table 4 show the statistics of the grade levels of our dataset's target language sentences (i.e., English sentences). Table 5 describes the statistics of the sentence lengths in our dataset. The length of a Japanese sentence in Table 5 is the number of characters in the sentence, whereas the length of an English sentence is the number of words in the sentence.

[5]We identified aligned English sentences on the basis of "paragraph index" and "sentence index within the paragraph" attaced to the Newsela-auto corpus.

[6]",", ":", ";", "-" and etc.

Algorithm 1 Automatic Filtering to Remove Inappropriate Data

function MAKE_CLEAN_LIST($lines$)
  $list \leftarrow []$
  for all $sent1, sent2 \leftarrow lines$ **do**
    $diff \leftarrow (sent1.level - sent2.level)$
    $val \leftarrow edit\_distance(sent1, sent2)$
    if $(val >= 1)$ and $(diff >= 2)$ **then**
      $list \leftarrow sent1, sent2$
    end if
  end for
  return $list$
end function

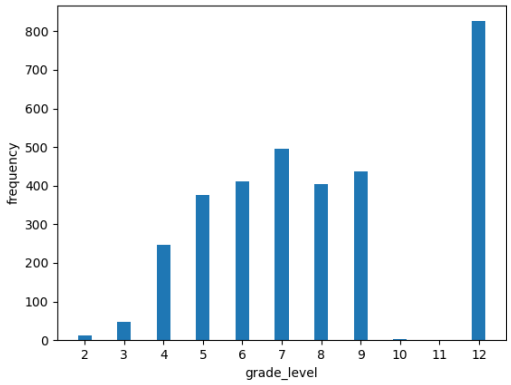

Figure 1: Target-Side Grade Levels in Our Dataset

3.3. Discussion on Automatic Filtering

The proposed creation procedure introduces an automatic filtering (Step 1-2) to improve the quality of our test dataset. This section discusses the necessity of automatic filtering.

We examined inappropriate sets in which the complexities of sentences with different grade levels do not change for the English sentence sets before automatic filtering (i.e., the sets simply based on the Newsela-auto corpus). In particular, we sampled 100 sets at random from the sets created by Step 1-1 and then assessed whether complexities change in each sampled set.

As a result, suitable sets consisting of sentences with different complexities are only 37 sets, and the remaining 63 sets contain inappropriate sentence pairs where the complexities are the same even though the attached complexity levels (i.e., grade levels) are different. For example, the only difference between some inappropriate sentence pairs is difference in symbols. Such data should be excluded from a test dataset since they disturb accurate evaluation of complexity controlling

	Japanese	English
Minimum Length	11	3
Maximum Length	188	71
Average Length	55.6	18.7

Table 5: Statistics of Sentence Length in Our Dataset

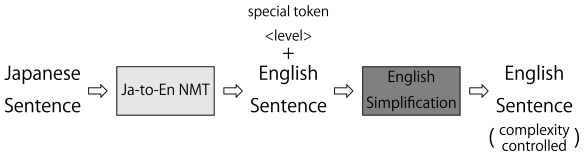

Figure 2: Overview of Our Pipeline Model

the performance of an MT model. However, manually removing them is impractical since they account for about 60% of the total, resulting in time-consuming and costly labor. Therefore, our filtering process, which can automatically remove such inappropriate data, should be useful.

4. Benchmark Experiments

In this section, we use a Transformer architecture to implement two MLCC-NMT frameworks, pipeline framework and multi-task framework, proposed in the previous study (Agrawal and Carpuat, 2019), and evaluate their performance on our test data to serve as baselines for future MLCC-MT research. Our models' implementation is based on Fairseq (Ott et al., 2019).[7] Note that while the previous study used LSTM-based sequence-to-sequence models (Bahdanau et al., 2015) in both frameworks, we used Transformer models (Vaswani et al., 2017), which have recently become the de facto standard for various NLP tasks, including MT and text simplification. It should also be noted that while the Newsela corpus' original Spanish and English sentences can be used as training data for MT between Spanish and English, as studied previously, the Newsela corpus does not contain Japanese sentences, which are required for training of Japanese-to-English MT.

4.1. Benchmark Models

4.1.1. Pipeline Model

The pipeline model is a sequential combination model of an MT model and a multilevel simplification model. While Agrawal and Capuat (2019) implemented two pipeline models, one that first translates and then simplifies an input ("Translate-then-Simplify") and the other that first simplifies and then translates an input ("Simplify-then-Translate"), we only implement and evaluate the Translate-then-Simplify model because previous work has shown that the Translate-then-Simplify model outperforms the Simplify-then-Translate model, and moreover, it is not straightforward to create training data for a Japanese multilevel simplification model from the Newsela corpus. Our pipeline model is depicted in Figure 2.

We used the Transformer model (Vaswani et al., 2017) both for a Japanese-to-English NMT model and an English multilevel simplification model.

Our Japanese-to-English NMT model follows Kiyono et al. (2020)'s Japanese-to-English NMT model, which has achieved state-of-the-art performance on the WMT'20 Japanese-to-English news translation task. We used the JParaCrawl (Morishita et al., 2020) and News Commentary datasets as the training data. Note that we applied language identification filtering to the training data using langid,[8] keeping only the sentence pairs where the source language is Japanese and the target language is English. As a result, 9.7M sentence pairs were used in training. Trucasing was performed on only English sentences, and SentencePiece (Kudo and Richardson, 2018) was used for subword segmentation, with the subword size set to 32,000. Table 6 shows the other experimental settings for our Japanese-to-English NMT model.

According to Scarton and Specia (2018), our English multilevel simplification model is a sequence-to-sequence model that converts the token sequence of an input English sentence with a special token representing a target complexity level into an English sentence at that level. We used 150K data randomly sampled from the Newsela-auto corpus as the training data. Note that we removed the data existed in our test dataset. We also used clean-corpus-n.perl[9] to remove sentences with more than 100 words and sentence pairs with a source/target length ratio greater than 2.0. FastBPE[10] was used for subword segmentation, with the subword size set to 8,000. Table 6 shows the other experimental settings for our English multilevel simplification model.

4.1.2. Multi-task Model

The multi-task framework trains one encoder-decoder model based on the three losses: the loss for conventional MT (L_{MT}), the loss for text simplification ($L_{Simplify}$), and the loss for complexity-controllable MT (L_{CMT}). The loss function of the multi-task model ($loss$) is as follows:

$$loss = L_{MT} + L_{Simplify} + L_{CMT}, \quad (1)$$

$$L_{MT} = \sum_{(s_i, s_o) \in D_{MT}} \log P(s_o | s_i; \theta), \quad (2)$$

$$L_{Simplify} = \sum_{(s_o, c_{o'}, s_{o'}) \in D_S} \log P(s_{o'} | s_o, c_{o'}; \theta), \quad (3)$$

[7] https://github.com/pytorch/fairseq

[8] https://github.com/saffsd/langid.py

[9] https://github.com/moses-smt/mosesdecoder.git

[10] https://github.com/glample/fastBPE.git

	Pipeline Model		Multi-Task Model
	NMT Model	Simplification Model	
arch	transformer	transformer	transformer
share-decoder-input-output-embed	True	True	True
activation-fn	relu	relu	relu
optimizer	adam	adam	adam
adam-betas	'(0.9, 0.98)'	'(0.9, 0.98)'	'(0.9, 0.98)'
clip-norm	1.0	0.0	1.0
lr	7e-4	7e-4	5e-4
lr-scheduler	inverse_sqrt	inverse_sqrt	inverse_sqrt
warmup-updates	4000	4000	4000
warmup-init-lr	1e-7	1e-7	1e-7
weight-decay	0.0001	0.0001	1e-5
dropout	0.3	0.1	0.1
criterion	label_smoothed_cross_entropy	label_smoothed_cross_entropy	label_smoothed_cross_entropy
label-smoothing	0.1	0.1	0.1
max-tokens	40000	80000	4096
patience	-	-	5
fp 16	True	True	True
max-epoch	100	100	100,000,000

Table 6: Experimental Settings with Fairseq

$$L_{CMT} = \sum_{(s_i, c_{o'}, s_{o'}) \in D_{CMT}} \log P(s_{o'} | s_i, c_{o'}; \theta), \quad (4)$$

where θ, s_i, s_o, $s_{o'}$, and $c_{o'}$ are the shared parameters of the model, a source language sentence, a target language sentence, a simplified sentence of s_o, and the complexity of $s_{o'}$, respectively. For each set of aligned English sentences in the Newsela-auto corpus, we used the English-to-Japanese Google Translate to translate English sentence with the highest complexity level, and then built triplets of the translated Japanese sentence, an English sentence with a lower complexity level than the highest one, and its complexity level. Note that we excluded the data that existed in our test dataset. We used 200K triplets randomly sampled from the constructed triplets as D_{CMT}. We used the 200K triplets where translated Japanese sentences of D_{CMT} are replaced with their original English sentences, as D_S. We used 3,000K Japanese-English translation pairs randomly sampled from the JParaCrawl dataset as D_{MT}. Moreover, the preprocessing for each training dataset is the same as in the pipeline model. Table 6 shows the experimental settings for our multi-task model.

4.2. Evaluation Metrics

Following Agrawal and Carpuat (2019), we used BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) as evaluation metrics. Following Nishihara et al. (2019) and Scarton et al. (2018), we also report MAE_{fkgl}, which is the absolute mean error between a target complexity level (reference) and FKGL (Kincaid et al., 1975) computed from the set of sentences with the same target complexity level. We used EASSE[11] (Alva-Manchego et al., 2019) to compute these scores.

When computing SARI scores, three types of sentences in the same language are required (i.e., an input sentence, its simplified sentence, and its reference sentence), but the existing test dataset does not always provide a complexity-uncontrolled target language sentence that should be used as an input sentence for computing SARI. Agrawal and Carpuat (2019) used the target language sentence translated from a source language sentence by MT, as an input sentence for computing SARI. This evaluation assumes that MT performance is high enough, but actual MT outputs contain many errors. As a result, the SARI scores reported in the previous study cannot accurately assess the simplification performance of MLCC-MT.

In contrast, the English sentence with the highest complexity level in our test dataset corresponds to the source language sentence. Therefore, by using the English sentence with the highest complexity level as an input sentence for SARI calculation, our SARI scores do not depend on MT performance and can accurately measure simplification performance.

4.3. Results

Table 7 shows the performance of our pipeline model and our multi-task model described in Section 4.1, and Table 8 indicates AE_{fkgl}, the absolute error between a target complexity level and FKGL, for each target

[11] https://github.com/feralvam/easse

Model	BLEU	SARI	MAE_{fkgl}
Pipeline	15.12	23.89	2.084
Multi-Task	20.17	26.78	0.600

Table 7: Performance of Our Pipeline Model and Multi-Task Model

Target Complexity	Pipeline	Multi-task	Simplification Part of Pipeline
2	2.647	0.945	2.474
3	2.205	0.569	1.628
4	2.600	0.257	2.357
5	2.463	0.158	2.365
6	2.442	0.455	1.940
7	1.543	0.194	1.157
8	0.295	0.272	0.037
9	0.006	0.166	0.558
10	3.161	1.212	2.347
12	3.477	1.767	3.841

Table 8: AE_{fkgl} for Each Target Complexity Level

complexity level. Table 9 and Table 10 show the performance of the NMT model and the multilevel simplification model on our test dataset, respectively, for the pipeline model. On the most complex Japanese-English sentences, the performance of an NMT model was evaluated. We also report the performance of Japanese-to-English Google Translate on our dataset for comparison. We used a reference English sentence with the highest complexity level as an input and measured the simplification performance at other complexity levels when evaluating the performance of multilevel simplification model. Table 7 demonstrates that the multi-task model outperforms the pipeline model in terms of BLEU and SARI, which is consistent with the previously reported results (Agrawal and Carpuat, 2019). Table 7 and Table 8 show that MAE_{fkgl} and almost all AE_{fkgl} of the multi-task model are smaller than those of the pipeline model, which also supports the observation that the multi-task model more properly controls the complexity of a target language sentence.

5. Conclusion

We created a new benchmark test dataset for Japanese-to-English MLCC-MT. The proposed creation method includes automatic filtering of data with inappropriate sentence-level complexity, manual check for parallel target language sentences with different complexity levels, and manual translation to make our test dataset more appropriate than existing test datasets. We also implemented two Transformer-based MLCC-NMT models, a pipeline model and a multi-task model, and evaluated their performance on our test dataset, which can be used as benchmark performance for future research.

	BLEU
NMT of our Pipeline Model	17.51
Google Translate	13.45

Table 9: Performance of NMT Part of Our Pipeline Model

BLEU	SARI	MAE_{fkgl}
68.40	37.55	1.870

Table 10: Performance of Simplification Part of Our Pipeline Model

In future work, we would like to increase the size of our dataset and create a multi-lingual dataset for MLCC-MT.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP22K12177. These research results were partially obtained from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), JAPAN.

References

Agrawal, S. and Carpuat, M. (2019). Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564.

Alva-Manchego, F., Martin, L., Scarton, C., and Specia, L. (2019). EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54.

Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., and Specia, L. (2020). ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Jiang, C., Maddela, M., Lan, W., Zhong, Y., and Xu, W. (2020). Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960.

Kato, T., Miyata, R., and Sato, S. (2020). BERT-based simplification of Japanese sentence-ending predicates in descriptive text. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 242–251.

Katsuta, A. and Yamamoto, K. (2018). Crowdsourced corpus of sentence simplification with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Kiyono, S., Ito, T., Konno, R., Morishita, M., and Suzuki, J. (2020). Tohoku-AIP-NTT at WMT 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 145–155.

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.

Kuczmarski, J. and Johnson, M. (2018). Gender-aware natural language translation. In *Technical Disclosure Commons, (October 08, 2018)*.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Marchisio, K., Guo, J., Lai, C.-I., and Koehn, P. (2019). Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203.

Maruyama, T. and Yamamoto, K. (2018). Simplified corpus with core vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Morishita, M., Suzuki, J., and Nagata, M. (2020). JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3603–3609.

Nishihara, D., Kajiwara, T., and Arase, Y. (2019). Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Paetzold, G., Alva-Manchego, F., and Specia, L. (2017). MASSAlign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4.

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

Scarton, C. and Specia, L. (2018). Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.

Schioppa, A., Vilar, D., Sokolov, A., and Filippova, K. (2021). Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696.

Sennrich, R., Haddow, B., and Birch, A. (2016). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

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.

Surya, S., Mishra, A., Laha, A., Jain, P., and Sankaranarayanan, K. (2019). Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

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.