Simplifying Translations for Children: Iterative Simplification Considering Age of Acquisition with LLMs

Masashi Oshika[1] **Makoto Morishita**[2] **Tsutomu Hirao**[2]
Ryohei Sasano[1] **Koichi Takeda**[1]
[1] Graduate School of Informatics, Nagoya University
[2] NTT Communication Science Laboratories, NTT Corporation
oshika.masashi.f6@s.mail.nagoya-u.ac.jp
{makoto.morishita,tsutomu.hirao}@ntt.com
{sasano,takedasu}@i.nagoya-u.ac.jp

Abstract

In recent years, neural machine translation (NMT) has been widely used in everyday life. However, the current NMT lacks a mechanism to adjust the difficulty level of translations to match the user's language level. Additionally, due to the bias in the training data for NMT, translations of simple source sentences are often produced with complex words. In particular, this could pose a problem for children, who may not be able to understand the meaning of the translations correctly. In this study, we propose a method that replaces words with high Age of Acquisitions (AoA) in translations with simpler words to match the translations to the user's level. We achieve this by using large language models (LLMs), providing a triple of a source sentence, a translation, and a target word to be replaced. We create a benchmark dataset using back-translation on Simple English Wikipedia. The experimental results obtained from the dataset show that our method effectively replaces high-AoA words with lower-AoA words and, moreover, can iteratively replace most of the high-AoA words while still maintaining high BLEU and COMET scores.

1 Introduction

Neural machine translation (NMT) has seen significant progress in recent years, making it practical for everyday use. As a result, more and more people are using it to meet their translation needs. NMT systems can generate fluent and grammatically correct sentences in most cases. However, some people may have difficulty understanding the translations due to the complexity of the words used. This is an especially serious risk when children use NMT systems. For example, when using an NMT system to translate textbooks used in the primary schools of another country, the system may use words that are not appropriate for primary school students in the country of the target language. This is because NMT systems do not take into account the complexity of the words used to compose the translations. Additionally, the parallel corpus used for the training, which mostly comes from the web, is biased because words used on the web are generally more complex than those used for children.

Therefore, a mechanism is needed to control the complexity of words in NMT systems. One way to measure such complexity is through the Age of Acquisition (AoA), which is the average age at which a person learns a particular word. Words with higher AoA are more difficult to learn. By controlling the AoA of words used in translations, we can simplify the translated sentence to a difficulty level that is more appropriate to the user. However, merely replacing words based on their AoA may not sufficiently simplify the sentence. Another approach is using paraphrase models trained with a parallel corpus of complex and simple sentences, but this method cannot control the AoA of words.

This paper proposes a method for appropriately simplifying translations to particular user levels, such as children, through post-editing. Figure 1 shows an overview of our method. We replace a high-AoA word in a translation with a simple one by using a large language model (LLM) while giving the source sentence. This technique allows us to replace not only the target word but also the surrounding words within the context of the sentence, which helps to simplify the overall sentence while preserving its original meaning. Moreover, we can apply this process iteratively to simplify all high-AoA words in a translation or in a partially simplified sentence that still contains complex words. Furthermore, users are able to customize the editing process on their own by specifying particular words that need to be replaced. This feature allows a more personalized editing experience.

Since there is no adequate dataset available for

We have released our code and dataset at https://github.com/nttcslab-nlp/SimplifyingMT_ACL24

Findings of the Association for Computational Linguistics: ACL 2024, pages 8567–8577
August 11-16, 2024 ©2024 Association for Computational Linguistics

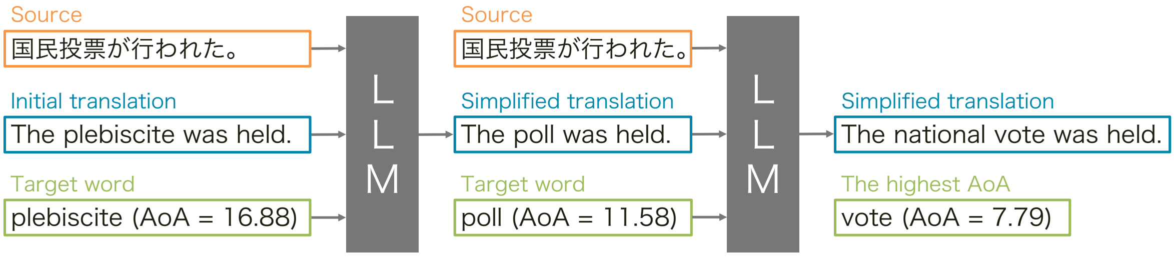

Figure 1: Overview of the proposed method, which generates a simplified translation given the source sentence, initial translation, and target word. If the highest AoA in the output sentence is higher than the target age, the model will iteratively revise the sentence. The target word is defined as the word with the highest AoA in the initial translation.

assessing the proposed method, we created a benchmark dataset using a back-translation approach. Initially, we translate a Simple English Wikipedia article into another language, and then we translate it back into English. In this way, the back-translations are treated as translations of simple sentences in the second language. The original sentences of the Simple English Wikipedia article serve as the reference simple sentences, and their corresponding back-translations serve as the sentences that need to be simplified.

Based on the results obtained using our benchmark dataset, we confirmed that our method outperforms the baseline methods, namely, MUSS (Martin et al., 2022), a simple post-editing approach, AoA-constrained NMT, an automatic post-editing approach, and LLM-based translation. Our method achieved the highest BLEU score while also maintaining the quality of simplification. Specifically, our method successfully replaced words with high AoA with those having lower AoA.

The contributions of this paper are as follows:

1. We automatically created a dataset based on Age of Acquisition (AoA) for text simplification from a monolingual corpus.

2. We propose a simplification technique using large language models (LLMs) to iteratively replace high-AoA words with lower-AoA ones. Furthermore, we demonstrate that our proposed method outperforms the baseline methods in both translation and simplification, as shown by the experimental results.

2 Related Work

2.1 Text Simplification

Text simplification is the task of transforming complex sentences into simple sentences that use basic vocabulary and grammar. This task shares the same objective as ours, which is to produce simplified sentences for children and beginner learners of a second or other language. The most common methods used to accomplish the text simplification task are adapting statistical machine translation (Wubben et al., 2012; Xu et al., 2016) and applying neural-based sequence-to-sequence models (Zhang and Lapata, 2017; Guo et al., 2018) trained with a parallel corpus consisting of complex and simple sentences (Coster and Kauchak, 2011). However, these methods cannot be controlled for specific simplification levels, including AoA. Our method can handle the more detailed aspects of simplification by iterating revisions with LLMs, permitting a more personalized simplification that would prove difficult for the previous methods.

2.2 Automatic Post-Editing

Automatic post-editing is a technique widely used to correct or improve machine translation outputs. Recently, transformer-based post-editing models have been widely used for this task (Bhattacharyya et al., 2022). In particular, methods that iteratively edit machine translation outputs have been proposed in recent years (Gupta et al., 2022; Chen et al., 2023). Gupta et al. (2022) proposed an iterative editing method using special tokens for keeping, deletion, controlling sentence length, and paraphrasing. Chen et al. (2023) used source sentences and mistranslated examples as prompts and then made iterative post-edits with LLMs. These methods are similar to ours, which edits machine translation outputs. However, automatic post-editing conventionally aims to correct errors in generated sentences to improve translation quality. In contrast, our work does not focus on correcting errors but attempts to control sentence difficulty, which is an important but challenging task.

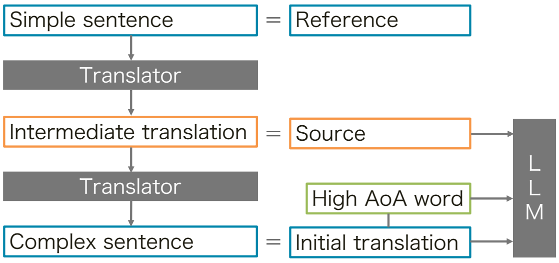

Figure 2: Procedure for creating a data set and the corresponding input to the LLM.

2.3 Complexity-control Method

There are several methods to generate sentences while controlling their difficulty. Agrawal and Carpuat (2019); Tani et al. (2022) proposed multi-level complexity controlling machine translation (MLCCMT) methods, which can control the difficulty of generated sentences at multiple levels. In addition, several methods have been proposed in the field of text simplification (Scarton and Specia, 2018; Nishihara et al., 2019; Chi et al., 2023; Agrawal and Carpuat, 2023). However, most of these methods define difficulty in sentence-level granularity, and thus they do not control difficulty at the level of individual words. In contrast, we achieve word-level control in our simplification by specifying the target word. In addition, since our method performs word-level simplification, each user can specify the word that he or she does not understand, which allows interactive adjustment of the sentence's difficulty level to match individual user's needs.

3 Dataset Creation

3.1 Procedure using Back-Translation

In order to evaluate our simplification technique for machine translation systems with a focus on Age of Acquisition (AoA), we need a dataset of translations that contain words with high AoA. We also need corresponding reference sentences that only include appropriate AoA words. Unfortunately, datasets of such pairs of sentences are not publicly available. Therefore, we created a dataset by automatically paraphrasing a monolingual corpus. Figure 2 shows the procedure for creating this dataset. To obtain translations including high-AoA words automatically, we use a back-translation approach, which translates a reference, a simple sentence, into another language (an intermediate translation) and then back into the original language using machine translation models (Sennrich et al., 2016; Edunov et al., 2018; Zhang et al., 2018). Current machine translation systems often produce translations with high-AoA words due to biases in a training parallel corpora, even when the original sentences are intended for beginners. This means that when translating simple English sentences into another language, the translations often contain complex words in the target language. As a result, back-translations from the second language may end up having more complex words than the original sentences from which they were derived. Note that our method has the advantage of not requiring a simplification corpus that consists of complex-simple sentence pairs. Therefore, this method can be used for any language with sufficient monolingual data and an AoA list[1].

When training or testing models to simplify translations for children of a certain age, we choose pairs of source sentences that contain words with an AoA greater than that age and the corresponding reference sentences that contain words with an AoA less than that age.

3.2 Creating Dataset from Simple English Wikipedia

We used sentences from Simple English Wikipedia[2] as a reference in our experiment. This was because it is available to the public and written in easy English for beginners. As part of the pre-processing step, we excluded the titles and section titles of entries. Consequently, 1,754,964 sentences were extracted.

We then performed back-translation on the dataset. We applied an English-to-Japanese translator to Simple English Wikipedia articles. After that, we applied a Japanese-to-English translator to the translated Japanese articles to obtain alternatives to the original English articles. Finally, we selected pairs of English sentences that showed a difference greater than 0.5 between the words with highest AoA in the reference and source sentences, respectively.

In Figure 3, we show the distribution of the differences between the highest AoA of the reference and that of the corresponding back-translation. From the figure, we found that the highest AoA of the words in the back-translation is sometimes greater than those in the original reference sen-

[1] For example, AoA for Spanish words was estimated by Alonso et al. (2014).

[2] https://huggingface.co/datasets/wikipedia/viewer/20220301.simple

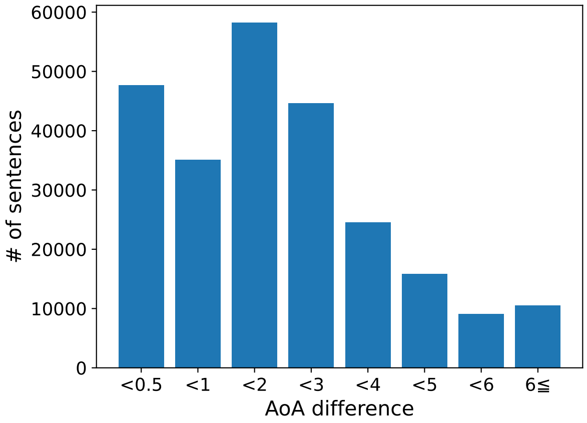

Figure 3: Distribution of the AoA difference in the created dataset, showing only the sentences where the AoA difference is greater than 0. AoA difference was 0 for 1,164,870 sentence pairs for 66% of the dataset.

tences. This finding appears to support our motivation, i.e., the need for a mechanism to replace complex words with simple ones for children.

After applying the filter, 235,331 sentences remained, which we divided into three sets: training, development, and test, at a ratio of 8:1:1. To provide back-translation, we trained two MT systems, English-to-Japanese and Japanese-to-English, using a transformer-based encoder-decoder model. Our training used JParaCrawl v3.0 (Morishita et al., 2022), the largest parallel corpus for English-to-Japanese/Japanese-to-English translation. Accordingly, our back-translation achieved a BLEU score of 48.3 and a COMET score of 90.0, indicating sufficient translation performance.

4 Proposed Method

Post-editing of initial translations is a promising approach to providing child-appropriate translations, given the difficulty of controlling the AoA of words in MT systems. One approach to achieve this is by simply replacing high AoA words in translations with lower AoA ones using a thesaurus. However, word-to-word replacement may sometimes change the meaning of the original sentences due to mismatched collocations. On the other hand, existing text simplification techniques are also available, but they do not guarantee that simplified translations consist of only appropriate AoA words.

To address these issues, we propose a method for iteratively applying LLMs to explicitly replace a high-AoA word with a lower one. LLMs are advanced language models that can reconstruct phrases containing high-AoA words, meaning they can significantly alter translations without changing their meanings.

We fine-tuned an LLM to generate a simplified sentence from a given initial translation, the source sentence[3], and a word whose AoA is greater than the specific value in the translation (Figure 1). We assumed that using the source sentences to simplify the translations would maintain the original meaning despite replacing the high-AoA words. High-AoA words in translations are enclosed in `<edit>` tags. Since our method determines the words to be edited based on their AoA, it allows for easy tailoring of the MT system based on the user's age. In other words, when the MT system users are n-year-old children, our system generates translations by replacing words with an AoA greater than n with words with an AoA less than or equal to n. In addition, since users themselves can specify words, it is possible to simplify a sentence by specifying words that a particular user does not understand, regardless of their age.

This method can be used iteratively, and thus if a translation contains multiple words that need to be replaced, all of these words can be replaced by repeatedly specifying words. In addition, if the AoA does not decrease after one round of simplification, it is possible to iteratively continue simplification until the AoA is below the criterion by again specifying the target words. The prompts used in the experiments are shown in Table 1.

5 Experiments

5.1 Settings

As an LLM, we used the pre-trained English-Japanese bilingual GPT-NeoX, which consists of 3.8 billion parameters, provided by Rinna[4]. To adapt the model to our task, we fine-tuned it using LoRA (Hu et al., 2022). LoRA was applied to all linear layers in the model, resulting in 25,952,256 trainable parameters, which is 0.68% of all model parameters. The hyperparameters of LoRA were set to $r = 16$ and $\alpha = 32$. The learning rate for fine-tuning was linearly decayed, with an initial rate of $1e$-5. We set the batch size to 16 and used AdamW as the optimization method. Fine-tuning was performed for 10 epochs, and the model with the smallest loss in the dev set was used to generate

[3] The source sentence is the intermediate translation of the reference sentence, as shown in Figure 2.

[4] `https://huggingface.co/rinna/bilingual-gpt-neox-4b-instruction-sft`

Proposed method	Instruction: Translate the following source language sentence based on the machine translation by simplifying the words surrounded by <edit>. ### source language sentence: この用語は、アルバム上の特定の曲を数字で表すためによく使用されます。 ### machine-translation: This term is often used to <edit>denote<edit> certain songs on the album by numbers. ### translation:
Direct Translation	You are a Japanese-English translator who only generates words that ten-year-old children can understand. Output the translation only. ### Source Japanese この用語は、アルバム上の特定の曲を数字で表すためによく使用されます。 ### Translated English:
Multi-word	Instruction: Translate the following source language sentence based on the machine translation by simplifying the words surrounded by <edit>. ### source language sentence: この用語は、アルバム上の特定の曲を数字で表すためによく使用されます。 ### machine-translation: This term is often used to <edit>denote<edit> <edit>certain<edit> songs on the album by numbers. ### translation:

Table 1: Prompts used in experiments. The original prompts are in Japanese, but here English-translated prompts are displayed for readability.

the test set.

In this study, the target age was set to 10 years old. In other words, the test set consists of data in which the highest AoA in the back-translation to be simplified was greater than 10, and the highest AoA in the reference sentences was less than 10. The total number of sentences in the test set was 8,289.

5.2 Evaluation Metrics

To evaluate how well the model succeeds in generating a simplified translation, we need to evaluate it from two viewpoints: machine translation accuracy and simplification quality. Machine translation accuracy indicates how fully the source sentence's meaning is maintained in the simplified sentence, while simplification quality indicates how well the model generates easy sentences. As machine translation metrics, we employed BLEU (Papineni et al., 2002), which evaluates text based on n-gram agreement, and COMET (Rei et al., 2020), which evaluates text based on embedded similarity. For the COMET model, we used `Unbabel/wmt22-comet-da`[5]. As the text simplification metric, we used SARI (Xu et al., 2016), which is evaluated based on the number of paraphrases and other factors from a triple of source, reference, and generated sentences. We also used FKGL (Kincaid et al., 1975), which is evaluated based on the number of words per sentence and syllables per word in the sentence. In addition, we employed Dale-Chall Readability (Chall and Dale, 1995), which assesses the readability of English text based on the average sentence length and the percentage of difficult words not found in a list of pre-defined common words. The Average AoA is the calculated average of the highest AoA for each sentence generated by each method. The success rate of simplification is the percentage of sentences in which the highest AoA in the sentence is lower than the target age.

5.3 Compared Methods

Here, we compare our method with the following baseline methods:

MUSS (Martin et al., 2022) is an unsupervised simplification model based on BART and trained with a parallel dataset for paraphrasing. To obtain simplified translations, we apply MUSS to initial translations, i.e., we used it as a post-editor.

Constraint Generation is a translation model that restricts the generation of words whose AoA is more than ten. Specifically, if a hypothesis contains a word with an AoA of ten or more at each generation time-step, the score of the hypothesis is set to $-\infty$. The beam size for search is set to 6 and 20. This method does not output anything if the translation failed in the restricted vocabulary. Therefore, if the translation fails, the initial translation in the test set is treated as the generated sentences. Note that Constraint Generation is incorporated in a machine translation model, i.e., it is not a post-editing approach. This method uses the same neural machine translation model that was used for dataset creation (Morishita et al., 2022).

APE is an automatic post-editing model trained on our dataset. As the APE model, we trained the transformer-big model that generates the reference

[5] https://huggingface.co/Unbabel/wmt22-comet-da

Method	Initial translation	Source language sentence	Target word
MUSS	✓	×	×
Constraint Generation	×	✓	×
APE	✓	✓	×
Direct Translation	×	✓	×
Multi-Word	✓	✓	✓
Proposed mmethod	✓	✓	✓

Table 2: Input information used by each baseline method and proposed method in generating sentences.

	Initial translation	MUSS	Constraint Generation	APE	Direct Translation	Multi-word		Proposed method				
						1	5	1	2	3	4	5
# of generated sentence	8,289	8,289	8,289	8,289	8,289	8,289	601	8,289	1,232	599	387	287
BLEU↑	38.4	26.9	39.6 / 39.7	44.8	30.5	44.3	44.6	44.9	45.0	45.0	45.0	**45.1**
COMET↑	87.3	83.4	86.4 / 86.2	87.9	84.8	**88.3**	88.2	88.2	88.2	88.2	88.2	88.2
SARI↑	53.2	42.6	53.9 / 53.5	58.6	49.5	59.3	59.8	59.8	60.0	**60.1**	**60.1**	**60.1**
FKGL↓	9.26	**6.49**	8.75 / 8.66	8.78	8.37	8.85	8.78	8.69	8.65	8.64	8.64	8.64
Dale-Chall↓	10.0	8.82	9.59 / 9.55	9.70	**8.17**	9.58	9.52	9.50	9.47	9.46	9.46	9.45
Average AoA	11.58	9.18	8.86 / 8.62	8.87	9.14	8.76	8.23	8.42	8.18	8.09	8.05	8.03
Success Rate↑	0.0	0.62	0.82 / 0.89	0.71	0.66	0.78	0.94	0.85	0.93	0.95	**0.97**	**0.97**

Table 3: Experimental results. The second row of the "Multi-word" and "Proposed method" columns indicate the number of iterations. The scores on the left side were obtained by the beam size of 6, while those on the right side were obtained by the beam size of 20 in the "Constraint Generation" column. "# of generated sentences" indicates the number of sentences generated in each iteration. If the highest AoA in the generated sentence is less than 10 (target age), the sentence will not be included in the next iteration for Multi-word and Proposed methods. During evaluation, the entire test set (8,289 sentences) is used, not just the generated sentences.

simple translation given a pair of source language sentences and initial translations, which are concatenated with a special token (i.e., <SEP>). The model architecture is the same as the original Transformer, and it is trained with a cross-entropy loss. This method is one of the simple but strong baselines in the APE task, as seen by participants in the recent WMT APE task (Bhattacharyya et al., 2022) also employing similar Transformer-based approaches.

Direct Translation is an LLM-based translation model that directly translates source language sentences into simple target sentences. We utilized GPT-3.5-turbo as the LLM and provided it with the prompt shown in the second row of Table 1.

Multi-word is a variant of our proposed method. We provide LLMs with all words having an AoA above ten in a translation within a single iteration, instead of providing words iteratively. The prompt for this model is shown in the third row of Table 1. The input data utilized for each method is shown in Table 2.

6 Results and Discussion

6.1 Main Results

Table 3 shows the experimental results. Although MUSS achieved the best FKGL and Dale-Chall scores, it scored the worst in BLEU and COMET, indicating that MUSS simplifies translations, but this simplification leads to inaccuracies. In addition, the low success rate implies that this level of simplification does not align with our goal of creating simplified translations for ten-year-old children.

Constraint Generation yielded better results in all metrics except COMET than the initial translation for both beam sizes by effectively suppressing the generation of words with an AoA greater than 10. When comparing beam sizes 6 and 20, the latter beam size generally yielded better results, with the exception of COMET. However, its success rate was only 0.89 with the beam size of 20, indicating that Constraint Generation often failed to generate words with an AoA of less than 10. It was also unsuccessful in generating translations for about 13% of the sentences due to the limitations of beam search, such as the repetition problem. These findings suggest the limitations of constraint generation with simple beam search, making it difficult to satisfy the constraints during decoding.

On the other hand, we found that APE produced better results than Constraint Generation. However, the Success Rate was the third worst, which suggests that it was challenging to replace high-AoA words. In addition, this also suggests that simplification might be achieved by ignoring high-AoA words. We also tested APE without source sentences (namely, intermediate translations) but did not observe any improvement in its performance compared to APE with intermediate translations.

The direct translation method improved the FKGL and Dale-Chall scores, but it led to a notable decrease in both BLEU and COMET scores. Additionally, the success rate of the method is only 0.66. These findings suggest that direct translation, which involves translating the source language sentence into a simple target sentence, cannot fully ensure the accuracy and complexity of translations.

After the first round of iterations, the proposed method showed improvements in all metrics when compared to the initial translations. This indicates that our method can simplify a translation while still retaining the original meaning. In addition, our method achieved significantly better scores than MUSS, Constraint Generation, APE, and Direct Translation, with the exceptions being FKGL and Dale-Chall.

With further iterations, our method showed slight improvements in BLEU and COMET but significant improvements in Average AoA and Success Rate. These results clearly demonstrate the effectiveness of our iterative simplification approach.

When comparing one-word replacement per iteration with multi-word replacement per iteration, it was found that the former achieved a better Success Rate, while both approaches obtained similar BLEU and COMET scores. On further analysis, it was observed that after five iterations, the one-word replacement approach outperformed multi-word replacement on BLEU and Success Rate. Another interesting finding was that the Average AoA was lower in one-word replacement than in multi-word replacement. These results suggest that multi-word replacement is a more challenging task for LLMs than one-word replacement.

Figure 4 shows the distribution of the highest AoA for each sentence and for different methods. The red line represents the reference, while the blue line represents the initial translation. After the first round of iterations, our method generated words with an AoA of 10 or higher. However, with further iterations, these words were replaced with low-AoA words. Although the Constrained Generation method and APE use words with low AoA, in some cases, they fail to generate low AoA sentences. As a result, sentences with an AoA of 10 or more are generated. Furthermore, MUSS generates a large number of words with an AoA of 10 or higher.

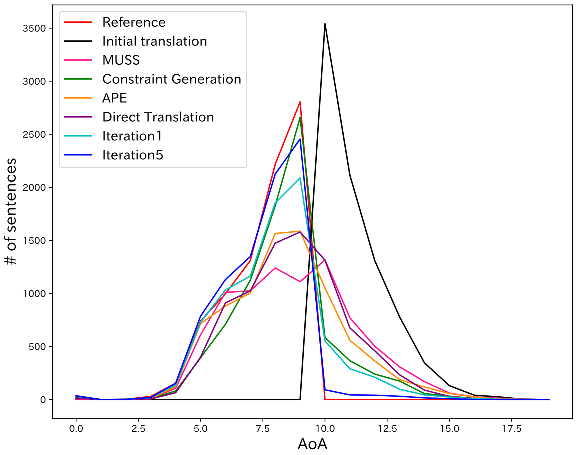

Figure 4: Statistical plots of highest AoA of the generated sentence by each method.

In short, our method simplifies given translations for a certain age group of children without reducing translation performance, as evidenced by these findings.

6.2 Ablation Study

In order to demonstrate the effectiveness of the proposed method, we conducted ablation studies using three different models. The first model, called "w/o intermediate," only provides translations and target words to LLMs using prompts. The prompts used in this model are shown in the first row of Table 4. The second model, called "w/o word," only provides translations and intermediate translations to LLMs, and the prompts this model uses are shown in the second row of Table 4. The third model, called "w/o intermediate and word," only provides translations to LLMs, and the prompts it uses in this model are shown in the third row of Table 4.

Table 5 exhibits the results obtained from the first and fifth iterations of each model. It is evident from the table that when intermediate translations were excluded, the translation performance showed a significant degradation. This is particularly clear in the results obtained from the "w/o intermediate" and "w/o intermediate and word" columns.

w/o intermediate	Instruction: Simplify the following machine translation by simplifying the words surrounded by <edit>. ### machine-translation: This term is often used to <edit>denote<edit> certain songs on the album by numbers. ### simplified sentence:
w/o word	Instruction: Translate the following source language sentence based on the machine translation. ### source language sentence: この用語は、アルバム上の特定の曲を数字で表すためによく使用されます。 ### machine-translation: This term is often used to denote certain songs on the album by numbers. ### translation:
w/o intermediate and word	Instruction: Simplify the following machine translation. ### machine-translation: This term is often used to denote certain songs on the album by numbers. ### simplified sentence:

Table 4: Prompts used in ablation study.

	Proposed method		w/o intermediate		w/o word		w/o intermediate and word	
	1	5	1	5	1	5	1	5
# of generated sentences	8,289	287	8,289	326	8,289	779	8,289	585
BLEU↑	44.9	**45.1**	40.0	39.4	44.6	45.0	40.2	39.4
COMET↑	88.2	88.2	86.9	86.4	**88.3**	**88.3**	87.0	86.5
SARI↑	59.8	**60.1**	55.5	55.4	59.3	60.0	55.5	55.4
FKGL↓	8.69	8.64	8.67	8.60	8.60	8.58	8.58	**8.46**
Dale-Chall↓	9.50	9.45	9.51	9.44	9.59	9.53	9.48	**9.39**
Average AoA	8.42	8.03	8.46	7.99	8.43	8.17	8.51	8.00
Success Rate↑	0.85	**0.97**	0.83	**0.97**	0.74	0.91	0.80	0.94

Table 5: Experimental results of ablation study. Numbers under the method names indicate the number of iterations.

The findings suggest that the input of intermediate translations is necessary to preserve the meaning of translations while simplifying them. Additionally, excluding target words resulted in a decrease in Success Rates, which is evident from the "w/o word" and "w/o intermediate and word" columns. These results suggest that specifying words for editing is essential for good simplification with fewer iterations. In summary, the presence of intermediate translations and that of target words are key indicators of successful simplification.

6.3 Examples of Simplification

Successful Simplification

In Table 6, we can see an example of successful simplifications in terms of AoA. In MUSS, the high-AoA word "denote" was replaced with a lower one, "show," resulting in all words having an AoA of less than 10. However, this change altered the meaning of the original translation. Constraint Decoding and APE were able to replace "denote" with "describe" while maintaining an AoA of less than 10 for all words. Our method, on the other hand, was initially unable to replace "denote" with a lower AoA word, but with more iterations, it eventually succeeded in doing this while keeping all words at an AoA of less than 10.

Erroneous Simplification

Table 7 shows an example of erroneous generation. After the second iteration, our approach was successful in replacing the high-AoA word "foreigners" with a simpler phrase, "foreign people." However, the word "origins" remains another complex word. Unfortunately, in the subsequent attempt, the word "foreigners" was generated again. This indicates that our method sometimes fails to simplify translations even after five iterations. To overcome this limitation, providing previous simplified sentences in each iteration could be an effective solution.

7 Conclusion

We proposed a method to provide easy-to-understand translations for children. Our method involves iteratively replacing high-AoA words in a translation with lower-AoA ones using LLMs. This is done by providing a triple of the source sentence, the translation, and the target word to be replaced. We also automatically generated a dataset from a monolingual corpus to evaluate simplification based on the back-translation technique. Moreover, we compared our method with baseline methods, namely, MUSS, Constraint Generation, Automatic Post-Editing, and Direct Translation, and we tested them on ten-year-old children (AoA=10). The re-

	Sentence	Highest AoA word
Source	この用語は、アルバム上の特定の曲を数字で表すためによく使用されます。	
Initial translation	This term is often used to denote certain songs on the album by numbers.	denote (11.24)
MUSS	This term is often used to show how many songs are on the album.	term (8.28)
Constraint Generation	The term is often used to describe a specific song on an album in numbers.	specific (9.28)
APE	The term is often used to describe certain songs on the album by numbers.	term (8.28)
Direct Translation	This word is often used to represent a specific song on an album with numbers.	represent (10.33)
Multi-word	The term is often used to describe certain songs on an album by numbers.	term (8.28)
Iteration 1	The term is often used to represent a particular song on a given album by numbers.	represent (10.33)
Iteration 2	The term is often used to describe certain songs on a given album by numbers.	term (8.28)
Reference	The term is often used to mean a specific song on the album by number.	specific (9.28)

Table 6: Successful generation example. Proposed method could iteratively decrease the AoA to below the target age.

	Sentence	Highest AoA word
Source	しかし、その起源は、453年後の1951年に外国人によって最初に調査されました。	
Initial translation	But its origin was first investigated by foreigners in 1951, 453 years later.	foreigners (10.39)
MUSS	But foreigners first looked at its origin in 1951, 453 years later.	foreigners (10.39)
Constraint Generation	However, its roots were first investigated by a foreign citizen in 1951, 453 years later.	investigated (9.0)
APE	However, its origin was first investigated by foreign people in 1951, 453 years later.	investigated (9.0)
Direct Translation	But, its origin was first investigated by foreigners in 1951, 453 years later.	foreigners (10.39)
Multi-word	Its roots, however, were first explored by outsiders in 1951, 453 years later.	outsiders(9.75)
Iteration 1	But its origin was first explored by foreigners in 1951 after 453 years.	foreigners (10.39)
Iteration 2	However, its origins were first investigated by foreign people in 1951 after 453 years.	origins (10.25)
Iteration 3	However, its origins were first explored in 1951 by foreigners 453 years later.	foreigners (10.39)
Iteration 4	But its origins were first examined in 1951 by foreign people 453 years later.	origins (10.25)
Iteration 5	Its origins, however, were first looked at in 1951 by foreign researchers, after 453 years.	origins (10.25)
Reference	Its source, however, was first explored by non-native people in 1951, 453 years later.	native (9.20)

Table 7: Example of erroneous generation. Proposed method generated “foreigners” and “origins,” in which AoA is not less than the target age.

sults show that our method successfully replaced high-AoA words with lower-AoA ones while maintaining the highest BLEU and COMET. In particular, our method simplified 97% of complex translations after five iterations. Furthermore, ablation studies identified the best choice as the approach of replacing one complex word with a simpler one in each iteration.

Limitations

One serious limitation of the proposed method is its computational cost. In our experiments, our proposed method required an average of 0.5 seconds to generate a sentence in each iteration. However, note that this is not the slowest speed among the compared methods. The constraint generation method, which is the slowest, requires an average of 1.8 seconds to generate a sentence, since it requires checking whether the hypothesis includes high-AoA words in each time step. We assume the computation could be faster by distilling the LLMs or utilizing smaller LLMs.

Another limitation is that our method sometimes fails to simplify the sentence to meet the target age, as described in Section 6.3. In that case, the model became stuck in the loop of two high-AoA words, 'foreigners' and 'origins.' We hypothesized that this is because the model does not know the history of the previous iterations, and thus the results could be improved by feeding this history to the model.

We also plan to conduct extensive experiments with other experimental settings. For example, This study carried out experiments with a target age of 10, but it would be interesting to see the effect of varying the target age. This paper focused on simplifying English, but we plan to extend it to other languages, such as Spanish, which already has an AoA estimation (Alonso et al., 2014). We believe we can extend this method to a language with no AoA list by estimating the word complexity using unigram probability in the corpus, since a rare unigram would be difficult while and a frequently used one would be an easy word. However, we left these pursuits for future work.

References

Sweta Agrawal and Marine Carpuat. 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.

Sweta Agrawal and Marine Carpuat. 2023. Controlling Pre-trained Language Models for Grade-Specific Text Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12807–12819.

María A. Alonso, Angel Fernandez, and Emiliano Díez. 2014. Subjective age-of-acquisition norms for 7,039 Spanish words. *Behavior Research Methods*, 47:268–274.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2022. Findings of the WMT 2022 Shared Task on Automatic Post-Editing. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 109–117.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative Translation Refinement with Large Language Models. arXiv preprint arXiv:2306.03856.

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. 2023. Learning to Paraphrase Sentences to Different Complexity Levels. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1332–1354.

William Coster and David Kauchak. 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 (ACL)*, pages 665–669.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 489–500.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic Multi-Level Multi-Task Learning for Sentence Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 462–476.

Prabhakar Gupta, Anil Nelakanti, Grant M. Berry, and Abhishek Sharma. 2022. Interactive Post-Editing for Verbosity Controlled Translation. In *Proceedings of the 29th International Conference on Computational Linguistic (COLING)*, pages 5119–5128.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations (ICLR)*.

J. Peter Kincaid, Robert P. Fishburne, R L Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Institute for Simulation and Training. 56.*

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 1651–1664.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 6704–6710.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable Text Simplification with Lexical Constraint Loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL): Student Research Workshop*, pages 260–266.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Carolina Scarton and Lucia Specia. 2018. Learning Simplifications for Specific Target Audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 712–718.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.

Kazuki Tani, Ryoya Yuasa, Kazuki Takikawa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2022. A Benchmark Dataset for Multi-Level Complexity-Controllable Machine Translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 6744–6752.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence Simplification by Monolingual Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1015–1024.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics (TACL)*, 4:401–415.

Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 584–594.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 555–562.