MultiMSD: A Corpus for Multilingual Medical Text Simplification from Online Medical References

Koki Horiguchi[†] **Tomoyuki Kajiwara**[†] **Takashi Ninomiya**[†] **Shoko Wakamiya**[‡] **Eiji Aramaki**[‡]
[†] Ehime University [‡] Nara Institute of Science and Technology
{horiguchi@ai.cs., kajiwara@cs., ninomiya.takashi.mk@}ehime-u.ac.jp
{wakamiya, aramaki}@is.naist.jp

Abstract

We release a parallel corpus for medical text simplification, which paraphrases medical terms into expressions easily understood by patients. Medical texts written by medical practitioners contain a lot of technical terms, and patients who are non-experts are often unable to use the information effectively. Therefore, there is a strong social demand for medical text simplification that paraphrases input sentences without using medical terms. However, this task has not been sufficiently studied in non-English languages. We therefore developed parallel corpora for medical text simplification in nine languages: German, English, Spanish, French, Italian, Japanese, Portuguese, Russian, and Chinese, each with $10,000$ sentence pairs, by automatic sentence alignment to online medical references for professionals and consumers. We also propose a method for training text simplification models to actively paraphrase complex expressions, including medical terms. Experimental results show that the proposed method improves the performance of medical text simplification. In addition, we confirmed that training with a multilingual dataset is more effective than training with a monolingual dataset.

1 Introduction

Medical texts contain many technical terms (medical terms), and non-expert patients often cannot use the information effectively (Cheng and Dunn, 2015). Also, healthcare professionals frequently use technical terms when communicating with patients, even though they recognize that they should avoid them (Charpentier et al., 2021). Therefore, to facilitate patients' understanding of medical conditions and treatment methods, and to assist medical practitioners in explaining important medical information such as findings and diagnoses to patients, medical text simplification that paraphrases medical terms into expressions that are easily understood by patients is expected. In this background, while medical text simplification has been actively researched in English (Cao et al., 2020; Sakakini et al., 2020; Devaraj et al., 2021; Guo et al., 2021; Luo et al., 2022), it has not been sufficiently studied in non-English languages.

In medical text simplification, a parallel corpus consisting of pairs of complex sentences written for experts and simple sentences written for general audiences are used to train seq2seq models such as Transformer (Vaswani et al., 2017). In English, medical text simplification models are trained using large-scale parallel corpus for training (Luo et al., 2022; Bakker and Kamps, 2024). On the other hand, in Japanese, where only small-scale evaluation corpora for medical text simplification exist, text simplification models trained on other domains have been applied to medical text simplification (Horiguchi et al., 2024). For Spanish and French as well, only small-scale medical text simplification parallel corpora exist (Grabar and Cardon, 2018; Cardon and Grabar, 2020; Campillos-Llanos et al., 2022), which are not sufficient for training text simplification models. Therefore, a large-scale parallel corpus for medical text simplification is desired.

We utilize the online medical references of the MSD manual[1] to construct a large-scale parallel corpus for medical text simplification. This online medical references contains articles written for professionals such as medical practitioners, and articles written for the laypeople such as patients and their families, each of which is available in nine languages. Therefore, we performed embedding-based sentence alignment for those article pairs to automatically construct multilingual parallel corpora for text simplification, MultiMSD corpus. We then train medical text simplification models using our corpus and evaluate their perfor-

[1] https://www.msdmanuals.com

Findings of the Association for Computational Linguistics: ACL 2025, pages 9248–9258

	Reference	Sentence Pairs	Language
AutoMeTS	(Van et al., 2020)	3,300	English
-	(Sakakini et al., 2020)	4,554	English
MSD	(Cao et al., 2020)	930	English
MedLane	(Luo et al., 2022)	14,832	English
Med-EASi	(Basu et al., 2023)	1,979	English
PLABA	(Attal et al., 2023)	7,643	English
Cochrane-auto	(Bakker and Kamps, 2024)	35,800	English
CLARA-MeD	(Campillos-Llanos et al., 2022)	3,800	Spanish
CLEAR	(Cardon and Grabar, 2020)	4,596	French
JASMINE	(Horiguchi et al., 2024)	1,425	Japanese

Table 1: Parallel corpora for medical text simplification.

mance in each language. In addition, we propose a method to facilitate the paraphrasing of medical terms and complex expressions in medical texts.

Experimental results showed that large language models outperformed seq2seq models pretrained on multilingual data. We also confirmed that the proposed method improves the performance of medical text simplification. In addition, we confirmed the effectiveness of training on a multilingual dataset compared to a monolingual dataset. We release our corpus on GitHub.[2]

2 Related Work

Medical text simplification is the task of removing medical terms from input sentences and paraphrasing them into expressions that are easier for patients to understand. Table 1 shows sentence-level medical text simplification corpora. In English, parallel corpora ranging from 1,000 to 35,000 sentence pairs have been released, such as the MedLane dataset (Luo et al., 2022), which was constructed by manually annotating sentences collected from the MIMIC-III database, and AutoMeTS (Van et al., 2020), which was constructed by automatically extracting medical sentence pairs from Wikipedia-derived text simplification parallel corpora. In Japanese, a small-scale medical text simplification corpus for evaluation, called JASMINE (Horiguchi et al., 2024), has been released. This is a parallel corpus where the text from patients' weblog is paraphrased using medical terms. In French, CLEAR (Grabar and Cardon, 2018) was constructed by manually aligning professional and simplified texts collected from encyclopedias, pharmaceutical leaflets, and scientific summaries. Cardon and Grabar (2020) subsequently expanded the parallel corpus by extracting 4,596 sentence pairs from similar sources. Also, In Spanish, CLARA-MeD (Campillos-Llanos et al., 2022) was constructed by experts manually aligning against professional and simplified text collected from systematic review summaries and drug leaflets. However, large-scale sentence-level medical text simplification corpora for training seq2seq models based on deep learning do not exist for any language other than English.

As in our study, Cao et al. (2020) determined that the MSD manual, online medical references, is a useful language resource in medical text simplification. They have constructed a parallel corpus by collecting professional and consumer article pairs from the MSD manual and manually aligning them by experts. However, while this method provides high-quality alignment, it is costly and limited in scale. Furthermore, they use only English articles and do not focus on other languages. In this study, we collect professional and consumer article pairs from the MSD manual across nine languages and conduct embedding-based sentence alignment to construct a medical text simplification parallel corpus with a scale of 10,000 sentence pairs in each language.

3 MultiMSD Corpus

We focus on the MSD Manual[1] of the online medical reference to construct a large-scale parallel corpus. MSD Manual is the world's most widely used source of medical information on all medical topics, with a professional version written for specialists such as doctors and healthcare workers, and a consumer version written for the general public

[2] https://github.com/EhimeNLP/MultiMSDcorpus

en	Deep venous thrombosis (DVT) is clotting of blood in a deep vein of an extremity (usually calf or thigh) or the pelvis. Deep vein thrombosis is the formation of blood clots (thrombi) in the deep veins, usually in the legs.
fr	La thrombose veineuse profonde correspond à la formation d'un caillot sanguin dans une veine profonde d'un membre (habituellement le mollet ou les cuisses) ou le petit bassin. La thrombose veineuse profonde est la formation de caillots sanguins (thrombi) à l' intérieur des veines profondes, généralement dans les jambes.
it	La trombosi venosa profonda consiste nella formazione di un coagulo di sangue in una vena profonda di un arto (solitamente a livello del polpaccio o della coscia) o della pelvi. La trombosi venosa profonda consiste nella formazione di coaguli di sangue (trombi) all' interno delle vene profonde, in genere delle gambe.
ja	深部静脈血栓症(DVT)とは,四肢(通常は腓腹部または大腿部)または骨盤の深部静脈で血液が凝固する病態である。 深部静脈血栓症は、深部静脈に血栓(血液のかたまり)が形成される病気で、通常は脚で発生します。

Table 2: Corpus examples in English (en), French (fr), Italian (it), and Japanese (ja). The examples in each language are semantically aligned, with the first sentence being complex and the second sentence being simple.

such as patients and their families. The articles on each topic are available in 11 languages in the professional version and 12 in the consumer version. This study focuses on nine languages (German, English, Spanish, French, Italian, Japanese, Portuguese, Russian, and Chinese) supported in both the professional and consumer versions. We then automatically construct a parallel corpus of medical text simplification (MultiMSD corpus) by conducting embedding-based sentence alignment for professional and consumer article pairs. Table 2 shows examples from the MultiMSD corpus.

3.1 Pre-processing of Article Pairs

In the MSD Manual, professional and consumer articles on the same topic are linked. The same articles in each language corresponding to each other are also linked. Based on the structures of these web pages, we collected article pairs of complex and simple articles in nine languages and used them for corpus construction. Then, we applied sentence segmentation using Stanza[3] (Qi et al., 2020) for each article to split into sentence units. For Japanese, where many errors in sentence segmentation were observed, such as line breaks in the middle of sentences, we applied rule-based sentence segmentation[4].

3.2 Embedding-based Sentence Alignment

For a given article pair, let sentences in the professional article D^c be denoted as S_i^c $(1 \leq i \leq |D^c|)$ and sentences in the consumer article D^s be denoted as S_j^s $(1 \leq j \leq |D^s|)$. We also consider converting the sentence into a d-dimensional vector by the sentence embedding model $\varepsilon(\cdot)$.

We formulate the problem of sentence alignment for a given article pair as a weighted matching problem on a complete bipartite graph based on sentence embeddings. That is, the bipartite graph consists of the complex side D^c and the simple side D^s, with sentence embedding $\varepsilon(S_i^c)$ and $\varepsilon(S_j^s)$ as nodes. Moreover, the edges between nodes have weights $\phi(\varepsilon(S_i^c), \varepsilon(S_j^s))$. This weight is represented as the matrix $\Phi \in [0,1]^{|D^c| \times |D^s|}$. In this study, cosine similarity between vectors is used as the weight $\Phi(\cdot)$.

Sentence alignment $A \in \{0,1\}^{|D^c| \times |D^s|}$ is obtained by selecting the most similar sentence from the opposite side for each sentence on one side. We use two methods: asymmetric sentence alignment, which emphasizes recall, and symmetric sentence alignment, which emphasizes precision.

Asymmetric Sentence Alignment. In asymmetric sentence alignment, for each sentence S_i^c on the complex side, the most similar sentence S_j^s from the simple side is selected as $j = \arg\max_k \Phi_{i,k}$ and the sentences are aligned as $A_{i,j} = 1$. Similarly, for each sentence S_j^s on the simple side, the most similar sentence S_i^c from the complex side is selected as $i = \arg\max_k \Phi_{k,j}$, and $A_{i,j} = 1$. However, to prevent alignment errors, if $\phi(\varepsilon(S_i^c), \varepsilon(S_j^s)) < \theta$, then $A_{i,j} = 0$. All other sentence pairs are assigned $A_{i,j} = 0$, meaning that those sentences are not aligned.

Symmetric Sentence Alignment. In symmetric sentence alignment, a sentence is aligned only when the most similar sentence from the complex side and the most similar sentence from the simple side match. That is, $A_{i,j} = 1$ is set for sen-

[3] https://github.com/stanfordnlp/stanza

[4] https://github.com/wwwcojp/ja_sentence_segmenter

	English				Japanese				Avg.
	θ	Precision	Recall	F1	θ	Precision	Recall	F1	F1
mBERT ($\leftrightarrow$)	0.90	0.812	0.728	0.768	0.90	0.567	0.250	0.347	0.558
mBERT ($\rightarrow$)	0.95	0.877	0.564	0.687	0.55	0.174	0.586	0.269	0.478
XLM-R ($\leftrightarrow$)	0.95	0.716	0.673	0.694	0.95	0.575	0.400	0.472	0.583
XLM-R ($\rightarrow$)	0.95	0.281	0.772	0.412	0.95	0.173	0.536	0.262	0.337
LaBSE ($\leftrightarrow$)	0.60	0.833	0.866	0.850	0.70	0.871	0.864	**0.867**	**0.859**
LaBSE ($\rightarrow$)	0.70	0.883	0.782	0.829	0.70	0.816	0.873	0.844	0.837
mE5 ($\leftrightarrow$)	0.90	0.840	**0.911**	**0.874**	0.90	0.725	**0.877**	0.794	0.834
mE5 ($\rightarrow$)	0.95	**0.949**	0.649	0.771	0.95	**0.888**	0.723	0.797	0.784

Table 3: Evaluation of sentence alignment in English and Japanese. $\leftrightarrow$ represents symmetric sentence alignment, while $\rightarrow$ represents asymmetric sentence alignment. Threshold θ is the cosine similarity that achieved the highest F-score on the validation data, and the average value is the mean F-score in English and Japanese.

tence pairs such that $(i = \arg\max_k \Phi_{k,j}) \wedge (j = \arg\max_k \Phi_{i,k})$, and $A_{i,j} = 0$ for all other sentence pairs. Furthermore, similar to asymmetric sentence alignment, if $\phi(\varepsilon(S_i^c), \varepsilon(S_j^s)) < \theta$, we set $A_{i,j} = 0$ to prevent alignment errors.

3.3 Construction of Parallel Corpora

In this study, four sentence embedding models based on BERT (Devlin et al., 2019) corresponding to the nine languages are used for sentence alignment. The settings of each sentence embedding method followed the original paper's settings.

- mBERT[5] (Devlin et al., 2019): Multilingual sentence embedding pre-trained with the masked language modeling task using Wikipedia in 104 languages. The special token [CLS] at the beginning of the sentence was used.

- XLM-R[6] (Conneau et al., 2020): Multilingual sentence embedding pre-trained with the masked language modeling task using Common Crawl in 100 languages. The [CLS] token was used.

- LaBSE[7] (Feng et al., 2022): Multilingual sentence embedding obtained by fine-tuning the multilingual masked language model with the translation ranking task (Guo et al., 2018). The [CLS] token was used.

- mE5[8] (Wang et al., 2024): Multilingual sentence embedding obtained by contrastive learning (Wang et al., 2022) of the multilingual masked language model followed by fine-tuning with tasks such as question answering (Bajaj et al., 2018), information retrieval (Fan et al., 2019) and natural language inference (Gao et al., 2021). Average pooling of token embeddings was used.

To evaluate the performance of sentence alignment, we randomly extracted 20 article pairs with correspondences in English and Japanese. For these article pairs, the authors manually annotated sentence alignment, and obtained 373 and 401 sentence pairs, respectively. Then, we divided them into 10 article pairs each for validation and evaluation. Since the distribution of cosine similarity is different for each embedding model, we used the validation data to adjust alignment threshold $\theta \in \{0.50, 0.55, ..., 0.95\}$ for each model.

Table 3 shows the performance on the evaluation data with the threshold that achieved the highest F-score on the validation data. The table shows that mE5 ($\leftrightarrow$) achieved the highest performance in English and LaBSE ($\leftrightarrow$) in Japanese. However, when comparing the average performance of mE5 and LaBSE in both languages, LaBSE performs better than mE5. Therefore, we adopted symmetric sentence alignment based on LaBSE's multilingual sentence embedding and set the threshold $\theta = 0.7$, which achieved the highest F-score on the validation set. We applied this method to the article pairs in Section 3.1 and automatically con-

[5] https://huggingface.co/google-bert/bert-base-multilingual-cased

[6] https://huggingface.co/FacebookAI/xlm-roberta-base

[7] https://huggingface.co/sentence-transformers/LaBSE

[8] https://huggingface.co/intfloat/multilingual-e5-base

	de	en	es	fr	it	ja	pt	ru	zh
Article pairs	1,540	1,544	1,556	1,542	1,550	1,580	1,562	1,561	1,544
Sentence pairs	16,163	8,871	15,743	16,546	14,973	14,349	17,384	17,757	12,834
Vocab size	34,436	15,429	22,267	21,906	22,726	14,452	24,588	46,119	17,400
	27,132	12,939	18,346	17,942	19,369	12,619	19,415	36,078	15,019
Avg. characters	144.35	124.52	150.29	145.86	149.13	50.78	136.50	148.56	35.83
	141.91	121.44	148.79	146.65	144.59	50.41	135.00	138.73	34.95
Avg. words	22.19	22.17	26.69	26.02	26.91	30.67	24.90	22.12	22.57
	22.26	22.34	26.84	26.45	26.34	30.98	25.09	21.04	22.17
Train	13,343	7,249	13,025	13,616	12,361	11,793	14,314	14,610	10,532
Valid	1,304	720	1,268	1,363	1,222	1,147	1,413	1,472	1,059
Test	1,516	812	1,450	1,567	1,390	1,409	1,657	1,675	1,243

Table 4: Statistics of MultiMSD. Vocab size, Avg. characters, and Avg. words are shown, with upper values for complex and lower for simple sentences.

structed a medical text simplification parallel corpus in multiple languages.

3.4 Post-processing

The sentence pairs obtained in the previous section contained the following types of noise, which were automatically removed:

- Sentence pairs containing sentences that are too short (5 characters or fewer)
- Duplicate sentence pairs
- Sentence pairs where the complex sentence and the simple sentence are identical

After this post-processing, we constructed a medical text simplification parallel corpus for training, consisting of the nine languages.

3.5 Analysis of Corpus

Table 4 shows the statistics of the parallel corpus constructed in this study. We tokenized the sentences using Stanza[5] (Qi et al., 2020) and calculated the vocabulary size and the average number of words per sentence for each language.

Difference in Number of Sentence Pairs The number of article pairs collected from the MSD Manual is almost uniform across each language, but the number of sentence pairs varies significantly. For example, the number of sentence pairs in English is $8,871$, while in Russian, it reaches $17,757$, a nearly two-fold difference. The difference in the number of sentence pairs across languages may be attributed to variations in the cosine similarity distribution of multilingual sentence embedding.

Average Sizes The articles in the consumer versions of MSD Manual are written in expressions that are easy to understand for patients and the general public. As a result, the vocabulary size of simple sentences is consistently smaller than that of complex sentences in all languages. However, the average number of characters and words are almost the same between complex and simple sentences. This suggests that although technical terms and complex expressions are simplified, they are supplemented by simple expressions to retain the meaning of the sentences.

4 Proposed Method

The conservative behavior of text simplification models is one of their traditional challenges (Kajiwara, 2019). To address this problem, we proposed a method to actively paraphrase terminology in this section.

We propose a method that weights the cross-entropy loss of the correct words in the reference sentence not included in the input sentence to promote the paraphrasing of medical terms and complex words in the input sentence that are written for experts. When the loss of a word is increased, the system is trained to actively output that word, so that more simple words contained in reference sentences written for the general public are expected to be output. In this method, if the correct word w in the reference sentence is not included in

		de	en	es	fr	it	ja	pt	ru	zh
Input		8.66	18.43	11.89	10.48	9.91	11.13	10.53	7.95	13.93
mBART	mono	36.26	32.11	34.30	40.20	37.26	46.65	30.42	35.56	39.59
	multi	34.83	40.73	39.57	39.72	39.01	45.71	35.69	37.78	41.44
	mono*	41.63	38.84	39.09	43.53	40.54	46.56	37.46	39.40	41.99
	multi*	41.69	42.28	42.79	43.69	43.24	48.70	38.73	40.84	41.94
Llama	mono	35.78	36.83	36.37	41.61	36.77	46.15	36.26	38.49	41.03
	multi	37.64	39.33	39.59	39.78	38.38	47.60	38.97	38.77	42.16
	mono*	**43.88**	42.40	42.91	**45.64**	**44.91**	48.95	**45.57**	42.41	46.94
	multi*	42.35	**44.68**	**44.18**	44.79	43.04	**50.02**	43.31	**43.25**	**47.00**
	0-shot	40.16	42.17	41.95	40.81	41.24	45.21	41.09	39.29	40.73
	5-shot	40.30	42.37	43.32	40.65	41.68	47.12	42.35	40.48	43.58

Table 5: Results of SARI scores (* indicates proposed method, **bold** indicates best performance for each language).

the input sentence X, its cross-entropy loss $L(w)$ is multiplied by the weight θ. Finally, the word loss $L'(w)$ is as follows.

$$L'(w) = \begin{cases} \theta \cdot L(w), & w \notin X, \\ L(w), & w \in X. \end{cases}$$

5 Experiments

Through experiments on text simplification in the medical domain, we validate the usefulness of the MultiMSD corpus constructed in this study.

5.1 Model

We used mBART[9] (Tang et al., 2020), pre-trained on multilingual data, and Llama[10] (Grattafiori et al., 2024), an open-source large language model. We fine-tuned the models in a monolingual (mono) setting, where the dataset for each language was used, and in a multilingual (multi) setting, the dataset for all nine languages was used. In the multilingual setting, we combined the datasets for each language and then randomly shuffled them to avoid biasing the data in the batch toward a specific language. Fine-tuning was stopped after 3 epochs using early stopping based on the cross-entropy loss on the validation data. We used two RTX A6000 GPUs for both training and inference.

5.1.1 mBART

The batch size was set to 32, dropout rate to 0.1, learning rate to 10^{-4}, and maximum sequence length to 256, and the optimization method was AdamW (Loshchilov and Hutter, 2019).

5.1.2 Llama

We used LoRA (Low-Rank Adaptation) (Hu et al., 2022) for fine-tuning. The LoRA rank was set to $r = 4$, the scaling factor to $\alpha = 8$, the dropout rate to 0.05, and the target modules were [q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj]. We then set the batch size to 8, gradient accumulation to 4 steps, learning rate to $5e-5$, and maximum sequence length to 256.

Furthermore, we conducted text simplification using in-context learning in both zero-shot and few-shot settings. For few-shot, five random examples were selected from the validation data of each language. We used the following sentence as instructions, translated into each language:

> *You are an expert in paraphrasing complex sentence into simple sentence. Please rephrase the following sentence into simple sentence while keeping their original meaning.*

5.2 Dataset

In the MSD manual, each language article is a translation of the English article. Therefore, when randomly splitting the dataset for each language into sentence pair units for training, validation, and test set, sentence pairs in a translation relationship may be mixed, making it difficult to evaluate fairly in a multilingual (multi) setting. In the MultiMSD corpus, articles in a translation relationship between different languages are linked (Section 3.1). We use this property to split the data by article pair units in a translation relationship rather

[9] https://huggingface.co/facebook/mbart-large-50

[10] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Input	入浴および浸漬は,広範な接触皮膚炎やアトピー性皮膚炎など,広い部位の治療が必要な場合に使用される。 (Baths and soaks are used when therapy must be applied to large areas, such as with extensive contact dermatitis or atopic dermatitis.)
Llama$_{multi*}$	入浴や浸漬は、広い範囲の皮膚炎(例えば、アトピー性皮膚炎)の治療に使用されます。 (Baths and soaks are used to treat large areas of dermatitis (e.g., atopic dermatitis).)
Llama$_{5\text{-}shot}$	入浴や浸けるときは、広い体の部分に炎症が広がっている場合に使います。 (Baths and when soaking are used when inflammation has spread to large parts of the body.)

Table 6: Examples of text simplification output with "接触皮膚炎 (contact dermatitis)" and "アトピー性皮膚炎 (atopic dermatitis)" as technical terms.

than by random sentence pair units. In this study, we split 128 article pairs in each language for validation and another 128 for test set (Table 4).

5.3 Inference and Evaluation

In mBART, we generated output sentences using beam search with beam size 5 and $max\_length = 256$. In Llama, we generated output sentences using greedy search with $max\_new\_tokens = 256$. Text simplification performance was automatically evaluated using SARI (Xu et al., 2016) with EASSE[11] (Alva-Manchego et al., 2019) package.

5.4 Application of the Proposed Method

In monolingual (mono) and multilingual (multi) settings of mBART and Llama, we applied the proposed method in Section 4 and experimented with weights $\theta = \{2, 4, 8\}$. Then, we selected results for the threshold that maximized the average of SARI and BLEU (Papineni et al., 2002).

5.5 Results

Proposed method significantly improves simplification performance Experimental results are shown in Table 5. The models with the proposed method significantly improved SARI scores, with Llama (mono*) and Llama (multi*) achieving the highest performance for all languages.

Large language models achieve high performance Comparing mBART and Llama, Llama showed higher performance in monolingual and multilingual settings, confirming the effectiveness of fine-tuning large language models. In the setting of in-context learning without parameter updates, Llama with five examples (5-shot) outperformed Llama without examples (0-shot).

Training with multilingual data is effective When analyzing the detailed results for each language, it was observed that, in most languages, models trained with the multilingual dataset (multi) tended to outperform models trained with the monolingual dataset (mono). This result suggests that fine-tuning with multilingual data shares linguistic knowledge of other languages and improves text simplification models' performance.

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

		Grammar	Meaning	Simplicity
English	mBART$_{multi*}$	4.7	4.2	4.0
	Llama$_{mono}$	4.7	4.4	3.9
	Llama$_{multi}$	4.7	4.3	4.0
	Llama$_{multi*}$	4.6	4.2	4.0
	Reference	4.5	4.1	4.2
Japanese	mBART$_{multi*}$	4.7	4.4	3.5
	Llama$_{mono}$	4.8	4.6	3.6
	Llama$_{multi}$	4.9	4.5	3.7
	Llama$_{multi*}$	4.9	4.4	3.8
	Reference	4.9	4.4	4.2

Table 7: Results of human evaluation.

6 Analysis

6.1 Examples of Medical Text Simplification

Table 6 shows examples of medical text simplification in Japanese. This example includes technical terms such as "接触皮膚炎 (contact dermatitis)" and "アトピー性皮膚炎 (atopic dermatitis)." Llama (multi*), applying the proposed method, paraphrases these into the general term "皮膚炎 (dermatitis)", followed by the specific examples, "例えば、アトピー性皮膚炎 (e.g., atopic dermatitis)." On the other hand, Llama (5-shot), which does not update the parameters, uses the very generalized term "炎症 (inflammation)." However, "炎症 (inflammation)" is a term for broad symptoms that is not limited to the skin, which may lead to a reduction in understanding or potential misinterpretations.

	English	Japanese
Input	513	432
Reference	485	328
mBART$_{multi*}$	459	367
Llama$_{mono}$	498	372
Llama$_{multi}$	486	358
Llama$_{multi*}$	490	320
Llama$_{0\text{-}shot}$	463	223
Llama$_{5\text{-}shot}$	452	228

Table 8: Number of medical term types in output sentences of medical text simplification model.

6.2 Human Evaluation

We conducted human evaluation of the grammar, meaning, and simplicity of output sentences in order to assess the quality of the medical text simplification model. In addition, we conducted human evaluation of reference sentences. English and Japanese were the target languages, and 100 sentences were randomly selected from the test set and evaluated. In English, we employed three annotators using Amazon Mechanical Turk[12], who have a background as a healthcare professional and hold Master certification with a past approval rate of 95%. In Japanese, we employed three university students who are native Japanese speakers.

Table 7 shows the average scores of the three annotators. Llama (multi*), applying the proposed method, achieved high simplicity in both languages, consistent with the automatic evaluation results. For grammar and meaning, all models scored satisfactorily compared to the Reference.

6.3 Analysis of Types of Technical Terms

To evaluate the medical text simplification model's ability to paraphrase technical terms, we counted medical term types in the input, reference, and output sentences. For English, we used the Medical Subject Headings (MeSH)[13] as the medical terms dictionary and tokenized each sentence using Scispacy (Neumann et al., 2019). For Japanese, we used J-MedDic[14] (Ito et al., 2018) and tokenized each sentence with MeCab[15] (Kudo et al., 2004), which loaded the medical dictionary[14].

From Table 8, we confirm that Llama (multi*) effectively reduces the number of medical term types in both English (490) and Japanese (320), with results similar to the reference sentences (English: 485, Japanese: 328). This shows the proposed method's effectiveness in promoting the output of correct words in reference sentences that are not included in the input sentences. In addition, Llama (0-shot) and Llama (5-shot) have fewer types of technical terms than other models. In-context learning without parameter updates actively paraphrases technical terms according to the given instructions, but it should be noted that over-simplification is possible, as in Table 6.

	SNOW	MultiMSD
BART	32.88	**35.68**
SimpleBART	34.72	**35.80**

Table 9: Experimental results in Japanese. These scores are SARI, which evaluates text simplification models trained on our MultiMSD in the medical domain or SNOW in other domains, on the JASMINE corpus in the medical domain.

6.4 Comparison with Other Corpora

Since there is no training parallel corpus for medical text simplification in Japanese, previous research (Horiguchi et al., 2024) has evaluated text simplification models trained in other domains, SNOW (Maruyama and Yamamoto, 2018; Katsuta and Yamamoto, 2018) for the medical domain. In this section, we improve the performance of medical text simplification in Japanese by training with our MultiMSD corpus. Pre-trained seq2seq models and hyperparameters followed the previous research (Horiguchi et al., 2024) and automatically evaluate SARI (Xu et al., 2016) on JASMINE[16] (Horiguchi et al., 2024), an existing evaluation corpus for medical text simplification in Japanese. Table 9 shows that for both BART (Lewis et al., 2020) and SimpleBART (Sun et al., 2023) models, training on the Japanese portion of our MultiMSD corpus achieved higher performance than training on SNOW in other domains for Japanese medical text simplification.

7 Conclusion

In this study, we automatically constructed parallel corpora for training consisting of nine languages by embedding-based sentence alignment

[12] https://www.mturk.com/
[13] https://www.nlm.nih.gov/mesh/meshhome.html
[14] https://sociocom.naist.jp/manbyou-dic/
[15] https://taku910.github.io/mecab/
[16] https://github.com/EhimeNLP/JASMINE

from professional and consumer article pairs in the online medical references to facilitate research on medical text simplification in non-English. We compared 4 types of multilingual sentence embeddings: mBERT, XLM-R, LaBSE, and mE5. We found that mE5 was useful for sentence alignment in English medical texts, while LaBSE was useful for Japanese. Evaluating the performance of the text simplification models using the corpus constructed in this study, we found that the large language model Llama outperformed mBART, which was pre-trained on multilingual data. Furthermore, we confirmed that the proposed method of weights the loss of correct words in reference sentences not included in input sentences improves the performance of medical text simplification.

Limitations

The output sentences of automatic text simplification may contain hallucinations, which could lead to the risk of misinterpretation if patients take the information into account. Therefore, the outcomes of this research should be used appropriately under the supervision of a healthcare professional.

Acknowledgments

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425.

References

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 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: System Demonstrations*, pages 49–54.

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268*.

Jan Bakker and Jaap Kamps. 2024. Cochrane-auto: An Aligned Dataset for the Simplification of Biomedical Abstracts. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability*, pages 41–51.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-EASi: Finely Annotated Dataset and Models for Controllable Simplification of Medical Texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14093–14101.

Leonardo Campillos-Llanos, Ana R. Terroba Reinares, Sofía Zakhir Puig, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for Spanish medical text simplification. *Procesamiento del Lenguaje Natural*, 69(0):189–196.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071.

Rémi Cardon and Natalia Grabar. 2020. French Biomedical Text Simplification: When Small and Precise Helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716.

Victoria Charpentier, Rachael Gotlieb, Corinne E. Praska, Marissa Hendrickson, Michael B. Pitt, and Jordan Marmet. 2021. Say What? Quantifying and Classifying Jargon Use During Inpatient Rounds. *Hospital Pediatrics*, 11(4):406–410.

Christina Cheng and Matthew Dunn. 2015. Health Literacy and the Internet: A Study on the Readability of Australian Online Health Information. *Australian and New Zealand Journal of Public Health*, 39(4):309–314.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – Simple Corpus for Medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation*, pages 3–9.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective Parallel Corpus Mining using Bilingual Sentence Embeddings. In *Proceedings of the Third Conference on Machine Translation*, pages 165–176.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 160–168.

Koki Horiguchi, Tomoyuki Kajiwara, Yuki Arase, and Takashi Ninomiya. 2024. Evaluation Dataset for Japanese Medical Text Simplification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 219–225.

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 *International Conference on Learning Representations*.

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. 2018. J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Tomoyuki Kajiwara. 2019. Negative Lexically Constrained Decoding for Paraphrase Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 461–466.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the 2018 International Conference on Learning Representations*.

Junyu Luo, Junxian Lin, Chi Lin, Cao Xiao, Xinning Gui, and Fenglong Ma. 2022. Benchmarking Automated Clinical Language Simplification: Dataset, Algorithm, and Evaluation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3550–3562.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified Corpus with Core Vocabulary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1153–1160.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.

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*, pages 311–318.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Tarek Sakakini, Jong Yoon Lee, Aditya Duri, Renato F.L. Azevedo, Victor Sadauskas, Kuangxiao Gu, Suma Bhat, Dan Morrow, James Graumlich, Saqib Walayat, Mark Hasegawa-Johnson, Thomas Huang, Ann Willemsen-Dunlap, and Donald Halpin. 2020. Context-Aware Automatic Text Simplification of Health Materials in Low-Resource Domains. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 115–126.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification. In *Proceedings of the 2023 Findings of the Association for Computational Linguistics*, pages 9345–9355.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *arXiv:2008.00401*.

Hoang Van, David Kauchak, and Gondy Leroy. 2020. AutoMeTS: The Autocomplete for Medical Text Simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1424–1434.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pretraining. *arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv:2402.05672*.

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*, 4:401–415.