Target Language Monolingual Translation Memory based NMT by Cross-lingual Retrieval of Similar Translations and Reranking

Takuya Tamura s2120744_@_u.tsukuba.ac.jp
Xiaotian Wang s2320811_@_u.tsukuba.ac.jp
Takehito Utsuro utsuro_@_iit.tsukuba.ac.jp
Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba, Japan

Masaaki Nagata masaaki.nagata_@_ntt.com
NTT Communication Science Laboratories, NTT Corporation, Japan

Abstract

Retrieve-edit-rerank (Hossain et al., 2020) is a text generation framework composed of three steps: retrieving for sentences using the input sentence as a query, generating multiple output sentence candidates, and selecting the final output sentence from these candidates. This simple approach has outperformed other existing and more complex methods. This paper focuses on the retrieving and the reranking steps. In the retrieving step, we propose retrieving similar target language sentences from a target language monolingual translation memory using language-independent sentence embeddings generated by mSBERT or LaBSE. We demonstrate that this approach significantly outperforms existing methods that use monolingual inter-sentence similarity measures such as edit distance, which is only applicable to a parallel translation memory. In the reranking step, we propose a new reranking score for selecting the best sentences, which considers both the sentence length normalized log-likelihood of each candidate and the sentence embeddings based similarity between the input and the candidate. We evaluated the proposed method with English-to-Japanese translation of the ASPEC and English-to-French translation of the EU bookshop corpus. The proposed method significantly exceeded the baseline in BLEU score, especially observing a 1.4-point improvement in the EU bookshop dataset over the original retrieve-edit-rerank method.

1 Introduction

Many studies have incorporated translation memories (TM), a set of high-quality bilingual sentences, into the NMT model in recent years. Bulte and Tezcan (2019) and Tezcan et al. (2021) proposed a NFR (Neural Fuzzy Repair) model that improves translation accuracy by incorporating TM into NMT. The model retrieves a similar source sentence from the set of source language sentences in the TM based on edit distance and sent2vec (Pagliardini et al., 2018), and concatenates the translation of a similar source sentence with the input source sentence to the NMT model. Since this model only requires preprocessing of the input to the NMT model, TM can be incorporated without modifying the model's architecture. Therefore, it is highly compatible with existing NMT models and portable in terms of implementation. On the other hand, due to the limitation of input sentence length, the number of similar sentences available

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 313–323
September 4–8, 2023, Macau SAR, China.

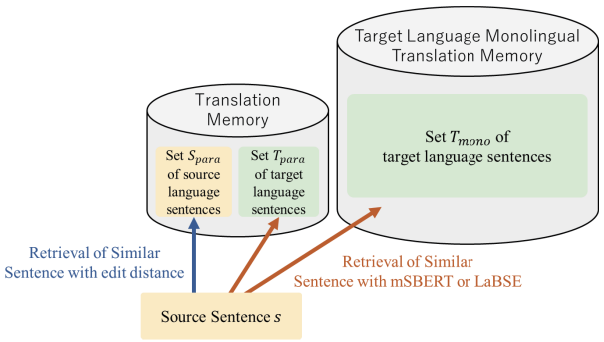


Figure 1: Retrieval of Similar Sentences from Translation Memory

for these methods is limited to one or two at most, and the retrieved similar sentences are not fully utilized. Also, if many informative similar sentences are obtained during inference, it is difficult to use all of them.

Hossain et al. (2020) proposed the retrieve-edit-rerank framework to overcome this limitation. They proposed a method that (1) retrieves multiple sentences from the training data using the input sentence as a query, (2) inputs the concatenation of the source and retrieved sentences into the model to generate multiple candidate sentences, and (3) extract the best sentence from the multiple candidates by choosing the sentence that maximizes the log-likelihood. In this paper, we focus on the (1) retrieval step and the (3) reranking step. As for the retrieval step, we compared monolingual inter-sentence similarity measures such as edit distance to cosine similarity based on language-independent sentence embedding with Multilingual Sentence-BERT (mSBERT) (Reimers and Gurevych, 2020) and LaBSE (Feng et al., 2022). Here, as shown in Figure 1, the edit distance requires the parallel corpus as the retrieval target, while the methods based on multilingual sentence embedding only requires a monolingual corpus of target language sentences. In the reranking step, we proposed a new reranking score for selecting the best sentences. This reranking score takes into account both the log-likelihood of each candidate with normalization by sentence length and the sentence embedding based similarity between the input and the candidate. We used the English-Japanese corpus of Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016) and the English-French corpus of EU bookshop corpus (EUbookshop) (Skadiņš et al., 2014; Tiedemann, 2012) to evaluate our method and found that the proposed method achieved significantly higher translation accuracy in all settings.

In summary, our contributions are as follows

1. In the framework of NFR (Figure 2), the use of similar sentences retrieved by language-independent sentence embedding generation models such as mSBERT and LaBSE significantly improved translation accuracy compared to conventional edit distance based retrieval methods (Table 2).
2. In the reranking phase of retrieve-edit-rerank (Figure 3), which selects the best sentence from multiple candidate output sentences, translation accuracy significantly improved by using a reranking score that takes into account both the log-likelihood of output with normalization by sentence length and the sentence embedding based similarity between the input and output candidate sentences (Table 3).

2 Related Work

As an NMT using the retrieve-edit framework, Bulte and Tezcan (2019) and Tezcan et al. (2021) proposed NFR (Neural Fuzzy Repair), a method to incorporate translation memory (TM) into

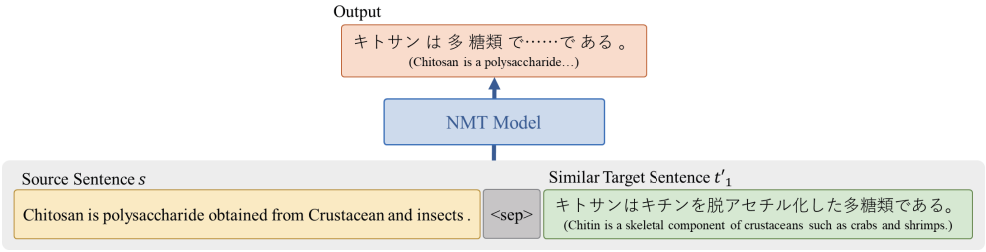


Figure 2: Framework of Translation with a Similar Target Sentence by NFR

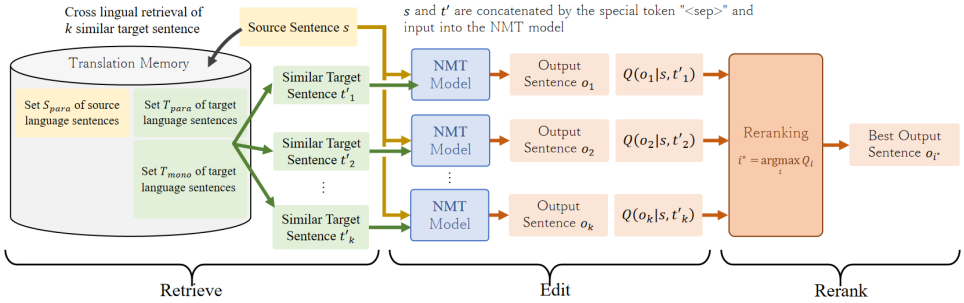


Figure 3: The Inference Framework of Retrieve-Edit-Rerank Model

NMT. They proposed a method that first retrieves similar source sentences based on edit distance using the input source sentence as the query, then concatenates the translation of the similar source sentence and the input source sentence and enters them into an LSTM-based NMT model. In this method, they achieved state-of-the-art in English-German and English-Hungarian translations. In addition, Xu et al. (2020) introduced word-by-word Fuzzy Matching to improve the accuracy of English-to-French translation using the Transformer model. They used cosine similarity of sentence embeddings as a similarity measure between an input sentence and the source language sentence. Among Bulte and Tezcan (2019), Tezcan et al. (2021), and Xu et al. (2020), both Tezcan et al. (2021) and Xu et al. (2020) introduced sentence embedding based similarity measure such as sent2vec for matching of the input sentence and similar source language sentences. However, they limit the search to the source language side of the training data. This information source is the same as the data used for training the baseline models, and the available quantity of data has not significantly increased. Our approach, on the other hand, is based on multilingual sentence embedding methods such as mSBERT and LaBSE and its information source for retrieving similar target sentences is the monolingual translation memory of target language sentences that is much larger than the training parallel data. Cai et al. (2021) also proposed a method that uses a monolingual corpus of the target language, rather than a bilingual corpus, as a target for retrieving similar sentences. They proposed a learnable retrieval model which is jointly optimized with the NMT model and performed similar sentence retrieval by MIPS (Maximum Inner Product Search). Although this approach achieves high performance, the model has to be built upon an architecture consisting of a Retrieval Model and a Translation Model. As a result, it eliminates the advantage of NFR where one can leverage the existing Transformer architecture and simply expand the input. Our approach, on the other hand, can be formalized as introducing the reranking phase of retrieve-edit-rerank into the architecture of the NFR framework, where it can be seen as leveraging the existing Transformer architecture in the edit phase of retrieve-edit-rerank framework.

For translation and summarization tasks, Hossain et al. (2020) proposed a method to gen-

erate multiple candidate output sentences and select the best output sentence by reranking them according to log-likelihood. They achieved a significant improvement in accuracy by combining NFR and retrieve-edit-rank frameworks. Despite the simplicity and versatility of this method, however, the improvement in translation accuracy due to the reranking is small.

In recent years, dense retrieval methods enabled us to retrieve semantically similar sentences with high accuracy and speed due to the development of Transformer-based language models. Reimers and Gurevych (2019) proposed a sentence embedding generation model, Sentence-BERT, to embed semantically similar sentences close to each other in vector space based on the pre-trained BERT (Devlin et al., 2019). More recently, Feng et al. (2022) proposed a multilingual sentence embedding generation model, LaBSE. These multilingual sentence embeddings can be retrieved quickly using approximate nearest neighbor search methods such as FAISS (Johnson et al., 2019).

3 Retrieval of Similar Sentences from Translation Memory (Retrieve)

3.1 Translation Memory

A translation memory (TM) is a set of high-quality bilingual sentence pairs that have been manually translated in the past. Computer-Aided Translation (CAT) is used as a tool to assist manual translation. If the source language sentence is already stored in the TM, it can be translated without error simply by replacing it with the target language sentence. Even when there is no exact match, a sentence with a certain degree of similarity (similar sentence) may be helpful during translation. In recent years, incorporating TM into NMT has been studied. In this paper, we define "similar target sentence" as a target language translation of a source language sentence similar to the input source language sentence ("similar source sentence").

Hereafter in this paper, a translation memory is defined as a set of pairs of a source language sentence s and a target language sentence t. Also, let S_{para} be a set of input source language sentences, T_{para} be a set of target language sentences, and T_{mono} be a monolingual translation memory of target language sentences. As shown in Figure 1, the original NFR requires the parallel corpus as the retrieval target and similar source sentences in the source language side S_{para} of the parallel translation memory are retrieved based on the edit distance. The proposed method, on the other hand, is based on multilingual sentence embedding methods such as mSBERT and LaBSE and only requires a monolingual corpus of target language sentences, where similar target sentences are retrieved not only from the target language side T_{para} of the parallel translation memory but also from the monolingual translation memory T_{mono} of target language sentences.

3.2 Similarity Measure based on Edit Distance

The edit distance is defined as the minimum number of operations required to convert one string into another string by inserting, deleting, or replacing. This paper followed Bulte and Tezcan (2019) and adopted the following similarity score of Vanallemeersch and Vandeghinste (2015),

$$sim_{ed}(x, y) = 1 - \frac{\Delta_{ed}(x, y)}{\max(|x|, |y|)}$$

where $\Delta_{ed}(x, y)$ is the edit distance between two sentences x, y, and $|x|$ is the number of tokens in x. When x and y perfectly match, the similarity score takes the maximum value $sim_{ed}(x, y) = 1$. Since the edit distance can only be calculated between two sentences of the same language, the retrieval is limited to the source sentences S_{para} in the TM. Therefore, the translation of "similar source sentences" is considered to be "similar target sentences". In addition, the computational cost during retrieval for large translation memories is significantly high

because similarity must be calculated and compared on a brute-force basis when retrieving similar sentences by edit distance. Therefore, following Bulte and Tezcan (2019), we also adopted a method to calculate edit distance only for candidate set[1] of similar sentences retrieved using the similarity measure $containment_{max}$ provided by a Python library $SetSimilaritySearch$ (sss). The $containment_{max}$ is defined for the set of unique tokens v_x and v_y contained in each source sentence x and y respectively as follows:

$$containment_{max}(v_x, v_y) = \frac{||v_x \cap v_y||}{\max(||v_x||, ||v_y||)}$$

3.3 Similarity Measure based on Multilingual Sentence Embeddings

In this section, we describe a similarity measure based on multilingual sentence embedding. Sentence embedding is a mapping of a sentence to a vector of real numbers, which is used for document classification, sentiment analysis and bilingual sentence retrieval. In this paper, we used Multilingual Sentence-BERT[2][3] (Reimers and Gurevych, 2020) and LaBSE (Feng et al., 2022) as the sentence embedding generation model. Sentence-BERT (SBERT) was trained on NLI datasets and achieved high accuracy in STS tasks. It is extended to Multilingual SBERT by knowledge distillation using monolingual English SBERT and parallel sentences. LaBSE is also a sentence embedding generation model trained on large-scale monolingual and bilingual texts and achieved state-of-the-art accuracy in the BUCC task of bilingual sentence retrieval. We defined the similarity measure sim_{se} based on multilingual sentence embeddings between two sentences x and y as follows, where $E(x)$ is the sentence embedding for the sentence x[4]:

$$sim_{se}(x, y) = \frac{E(x) \cdot E(y)}{|E(x)||E(y)|}$$

4 Generation with NMT Model (Edit)

4.1 Training

As shown in Figure 2[5], we trained the translation model using the same procedure as Bulte and Tezcan (2019) and Tezcan et al. (2021). Specifically, we first retrieve k-best similar target sentences $t'_1, t'_2, \ldots, t'_k$ from the TM by edit distance or sentence embeddings, using the source language sentence s as a query. As in NFR model, for each of t'_i ($i = 1, \ldots, k$), we concatenated s and t'_i with a special token "$\langle$sep$\rangle$" and entered them to the translation model as below together with the reference target language translation t.

$$\text{Input} \quad : \quad s \ \langle\text{sep}\rangle \ t'_i, \qquad\qquad \text{Reference} \quad : \quad t$$

Thus, for each source language sentence s, we entered k parallel sentences to the translation model for training.

4.2 Inference

Figure 3 shows the inference procedure for the retrieve-edit-rerank model. First, we search for k-best similar target sentences $t'_1, t'_2, \ldots, t'_k$ in the TM using edit distance or sentence embeddings. We then decode k times using the translation model to obtain the k output candidates o_i

[1] Candidates are limited to those satisfying the similarity lower bound of 0.5.

[2] `https://github.com/UKPLab/ sentence-transformers`

[3] In the implementation of this paper, we used `paraphrase-multilingual-mpnet-base-v2`.

[4] For the retrieve-edit-rerank machine translation, we have to extract k similar sentences from $T_{para} \cup T_{mono}$ using the input source sentence s as a query. We used FAISS (Johnson et al., 2019), a library for approximate nearest neighbor search on GPUs, to extract k-best similar sentences.

[5] Figure 2 illustrates the inference procedure by Bulte and Tezcan (2019), where only the translation of the source sentence with the highest similarity is used as the "similar target sentence".

	ASPEC (En→Ja)	EUbookshop (En→Fr)
Train	100,000	100,000 1,000,000
Dev	1,790	2,000
Test	1,812	2,000
Target Language Monolingual TM (including the target language side of Train)	2,000,000 (Ja)	8,421,120 (Fr)

Table 1: Statistics of the Datasets

($i = 1, \ldots, k$) and calculate the reranking score Q_i of o_i ($i = 1, \ldots, k$) based on the decoder's output probability p_{MT}.

5 Reranking Outputs by Reranking Scores (Rerank)

In the reranking step, out of the k output candidates o_i ($i = 1, \ldots, k$), we select the i^*-th output candidate o_{i^*} whose score Q_{i^*} is the largest among the k output candidates:

$$i^* = \underset{i=1,2,\ldots,k}{\arg\max}\, Q_i$$

We compared three reranking scores in this paper. The first is a reranking score based on the log-likelihood of the output candidate (Hossain et al., 2020).

$$Q_i^{(\text{Hossain})} = Q(s, t'_i, o_i) = \log_2 p_{MT}(o_i|s, t'_i)$$

Here, the p_{MT} represents the output probability of o_i when s, t'_i is input to the trained NMT model. It is calculated as follows:

$$p_{MT}(o_i|s, t'_i) = \prod_l p_{MT}(o_i^{(l)}|s, t'_i, o_i^{(<l)})$$

where, supposing that $o_i^{(<l)}$ represents the token sequence already output at the l-th step and $o_i^{(l)}$ represents the token output by the decoder at the l-th step, $p_{MT}(o_i^{(l)}|s, t'_i, o_i^{(<l)})$ represents the output probability at the l-th step of decoding.

The second is the proposed method, which is based on the average log-likelihood with normalization by sentence length. Here, let $|deSW(o_i)|$ be the number of words after detokenizing the subwords of the output candidate o_i.

$$Q_i^{(\text{proposed1})} = Q(s, t'_i, o_i) = \frac{\log_2 p_{MT}(o_i|s, t'_i)}{|deSW(o_i)|}$$

The third is another proposed method, which takes into account the average log-likelihood normalized by sentence length and the similarity between input and output candidates using multilingual sentence embeddings. In the subsequent experiments, we chose $\alpha = 0.4$ as the optimal value based on the development data. Furthermore, the similarity measure sim_{se} employed in this context is derived from LaBSE.

$$Q_i^{(\text{proposed2})} = Q(s, t'_i, o_i) = \alpha \frac{\log_2 p_{MT}(o_i|s, t'_i)}{|deSW(o_i)|} + (1 - \alpha)\, sim_{se}(s, o_i)$$

6 Experiments

6.1 Datasets

In this paper, to evaluate the proposed method, we used the English-Japanese corpus of Asian Scientific Paper Excerpt Corpus (ASPEC)[6] (Nakazawa et al., 2016) and the English-French corpus of the EU bookshop corpus (EUbookshop)[7] (Skadiņš et al., 2014; Tiedemann, 2012), which

[6] https://jipsti.jst.go.jp/aspec/

[7] https://opus.nlpl.eu/EUbookshop.php

is based on publications from various European institutions. The translation direction was from English to Japanese and from English to French, respectively. Only 100,000 or 1,000,000 randomly sampled sentences from each corpus were used as training data for the translation models, while the rest and the target language side of the training data were used as the monolingual translation memories. Table 1 shows the detailed numbers of sentences in these datasets. We tokenize the corpus using Moses tokenizer[8] for both English and French sentences and using MeCab[9] for Japanese. We then split it into sub-words using byte pair encoding BPE[10] (Sennrich et al., 2016) with applying 32,000 merge operations.

6.2 Setting

For the retrieval of similar sentences, we compared three different methods: $SetSimilaritySearch$ + edit distance (sss+ed), mSBERT, and LaBSE. With sss+ed, only the similar source sentences in the source language side of the training data (i.e., only 100,000 or 1,000,000 sentences shown in Table 1) are retrieved, while with the proposed methods with mSBERT and LaBSE, the similar target sentences not only in the target language side of the training data but also in the monolingual translation memory of target language sentences (i.e., 2,000,000 or 8,421,120 sentences shown in Table 1) are retrieved. During training, we compared the normal method without similar sentence retrieval (w/o retrieval) with a method that uses up to four similar sentences (top 1 to top 4). During inference, we compared three methods: a method that does not use similar sentences (w/o retrieval), a method that uses only the similar translation of the topmost 1 sentence as in the original NFR (top 1), a method that reranks based on $Q^{(\mathrm{Hossain})}$, and two proposed methods that rerank based on $Q^{(\mathrm{proposed1})}$ and $Q^{(\mathrm{proposed2})}$. In those reranking methods, we use the number k of output candidates as $k = 32$. In addition, we define the oracle as selecting the one with the highest Sentence-BLEU out of the output candidates for each input sentence to investigate the upper bound of translation accuracy improvement due to reranking. In the comparison of retrieval methods in Table 2, we consider sss+ed as the baseline. In the comparison of reranking methods in Table 3, on the other hand, for each retrieval method, we consider the method that uses only the similar translation of the topmost 1 sentence (top 1) as the first baseline (baseline 1) and that based on $Q^{(\mathrm{Hossain})}$ as the second baseline (baseline 2)[11].

6.3 Results

The results of training the translation model by retrieving similar translations using each retrieval method are shown in Table 2. The number of similar sentences used for training is set to $k = 1, 2, 3, 4$, and the number of similar sentences used for inference is set to $k = 1$. Without the retrieval of similar translations, the ASPEC, EUbookshop (100K), and EUbookshop (1M) BLEUs were 26.2, 20.2, and 26.9 points, respectively, whereas the sss+ed BLEUs were up to 26.4, 20.2, and 28.6 points, respectively, and significantly improved only for EUbookshop (1M). On the other hand, LaBSE showed significantly higher BLEU than sss+ed in all cases, with maximums of 27.1, 21.0, and 30.6 points. The highest BLEUs were obtained for both mSBERT and LaBSE when the topmost two or three sentences were used, and it can be confirmed that the accuracy conversely decreases when the topmost four sentences are used.

[8] https://www.statmt.org/moses/

[9] https://github.com/neologd/mecab-ipadic-neologd

[10] https://github.com/rsennrich/subword-nmt

[11] The encoder and decoder were 6 layers each, with 512 hidden dimensions, 2,048 dimensions in the FF layer and 8 multi-heads. We also adopted a warm-up of 6,000 steps and trained 30 epochs with a batch size of 32 sentences. Then, the BLEU score was measured against the test data at the number of epochs with the highest BLEU score against the development data.

	# of Similar Sentences		ASPEC (En→Ja)	EUbookshop (En→Fr)	
			# Training Data		
	Training	Inference	100,000	100,000	1,000,000
w/o retrieval	-	-	26.2	20.2	26.9
sss+ed (baseline)	top 1	top 1	26.4	20.2	28.6
	top 2		26.2	19.5	28.2
	top 3		26.1	18.3	27.6
	top 4		25.7	16.4	27.0
mSBERT	top 1	top 1	25.8	20.5	29.9†
	top 2		26.5	20.8	29.9†
	top 3		26.4	19.9	29.6†
	top 4		26.2	19.0	29.4†
LaBSE	top 1	top 1	25.8	20.9	30.2†
	top 2		**27.1**†	**21.0**†	30.3†
	top 3		26.5	20.4	**30.6**†
	top 4		26.3	19.3	30.0†

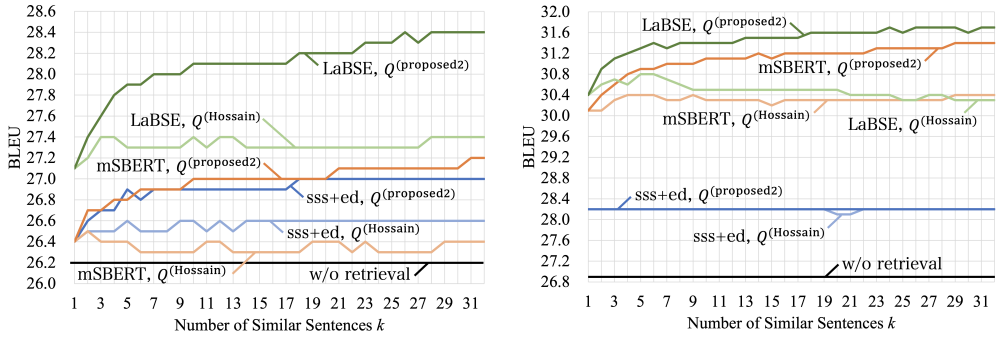
Table 2: Results of Comparing Retrieval Methods by the Translation Accuracies in BLEU (Topmost 1 similar sentence to be used during inference. "w/o retrieval" for vanilla Transformer without using similar sentences, sss+ed for a method using edit distance as NFR. † for significant ($p<0.05$) difference with the BLEU of sss+ed (baseline) when # of similar sentences in training is the same.)

Dataset	**Retrieval Method**	**w/o reranking** w/o retrieval	**w/o reranking** top 1 (baseline 1)	**w/ reranking** ($k=32$) $Q^{(\mathrm{Hossain})}$ (baseline 2)	$Q^{(\mathrm{proposed1})}$	$Q^{(\mathrm{proposed2})}$	oracle
ASPEC (En→Ja)	w/o retrieval	26.2	-	-	-	-	-
	sss+ed	-	26.2	26.6	26.8	27.0	28.5†‡
	mSBERT	-	26.5	26.4	26.9	27.2	29.7†‡
	LaBSE	-	27.1	27.4	28.1†	**28.3**†‡	31.8†‡
EUbookshop (En→Fr, 100k)	w/o retrieval	20.2	-	-	-	-	-
	sss+ed	-	20.2	20.3	20.3	20.3	20.3
	mSBERT	-	20.8	19.9	22.1†‡	22.4†‡	25.2†‡
	LaBSE	-	21.0	19.6	21.7‡	**22.5**†‡	25.6†‡
EUbookshop (En→Fr, 1M)	w/o retrieval	26.9	-	-	-	-	-
	sss+ed	-	28.2	28.2	28.2	28.2	28.3
	mSBERT	-	29.9	30.4	31.0†	31.4†‡	34.0†‡
	LaBSE	-	30.3	30.3	31.0	**31.7**†‡	34.2†‡

Table 3: Results of Comparing Retrieval/Reranking Methods by the Translation Accuracies in BLEU (Topmost 2 similar sentences to be used during training. "w/o retrieval" for vanilla Transformer without using similar sentences, "top 1" for a method using the most similar target sentence as NFR. $Q^{(\mathrm{Hossain})}$ for the reranking score based on log-likelihood of the output candidate, $Q^{(\mathrm{proposed1})}$ for the reranking score with length normalization of $Q^{(\mathrm{Hossain})}$, $Q^{(\mathrm{proposed2})}$ for the reranking score with $Q^{(\mathrm{proposed1})}$ and the similarity between input and output candidates. Oracle for selecting the sentence with the highest Sentence-BLEU from output candidates. † for significant (p<0.05) difference with the BLEU of "top 1" (baseline 1) when the retrieval method is the same, ‡ for significant (p<0.05) difference with the BLEU of $Q^{(\mathrm{Hossain})}$ (baseline 2) when the retrieval method is the same.)

Then, the results of reranking following the framework of retrieve-edit-rerank are shown in Table 3. First, when we focus on the reranking method using $Q^{(\mathrm{Hossain})}$, no significant improvement in BLEU was obtained for any of the reranking methods. On the other hand, the reranking method using $Q^{(\mathrm{proposed1,2})}$ did not improve BLEU significantly for sss+ed, but significantly improved BLEU in many cases when using mSBERT and LaBSE. The oracle that retrieves the sentence with the highest Sentence-BLEU shows an upper bound for reranking, but it is lower for sss+ed than for mSBERT and LaBSE, suggesting that there is little room for further improvement[12].

[12]For $Q^{(\mathrm{proposed2})}$ with mSBERT/LaBSE, the percentages of similar target sentences retrieved from target language monolingual TM (excluding the target language side of the training data) that give the largest score through reranking are 97.6/95.5, 99.0/98.9, and 87.9/87.6 (ASPEC, EUbookshop 100k and 1M),



(a) for ASPEC (En→Ja)

(b) for EUbookshop (En→Fr, 1M)

Figure 4: The Changes in BLEU Scores for the Number of Similar Sentences k for Each Retrieval Method

6.4 Impact of the Number of Similar Target Sentences used for Reranking

Figure 4 shows the changes in BLEU scores when the number of similar translations k used for reranking is changed. As an overall trend, the reranking method based on $Q^{(\text{proposed2})}$ yields significantly higher BLEU than the method based on $Q^{(\text{Hossain})}$. In particular, for EUbookshop (1M) in Figure 4b, using LaBSE and $Q^{(\text{proposed2})}$, BLEU improves almost monotonically as k increases, reaching 31.7 points at $k = 32$. On the other hand, using $Q^{(\text{Hossain})}$ improves to 30.8 points at $k = 5$ and 6, but drops to 30.3 points at $k = 32$. In terms of corpus differences, only the method using LaBSE achieves significantly higher BLEUs than the baseline of sss+ed in ASPEC in Figure 4a, while both mSBERT and LaBSE achieve significantly higher BLEU scores than sss+ed in EUbookshop (1M) in Figure 4b. This difference may be derived from the difference of the training of the models of mSBERT and LaBSE. While mSBERT is multilingualized by distilling the model to measure the similarity of English sentences, LaBSE is more suitable for bilingual sentence retrieval because LaBSE was originally trained using bilingual data. In addition, in terms of the number of sentences per language included in the LaBSE's training data, the Japanese language ranks at third following English and Russian, suggesting that it is more suitable for tasks involving the Japanese language than mSBERT. Finally, focusing on the differences in retrieval methods, sss+ed has the smallest range of change in BLEU due to reranking compared to the other retrieval methods, with little or no effect from reranking. This is mainly because sss+ed's retrieval target was limited to 100K/1M sentences of the source language side of the parallel translation memory and did not find high-quality similar target sentences. This suggests that the cross-language retrieval method based on mSBERT and LaBSE sentence embedding can find a much larger number of high-quality similar sentences than sss+ed.

6.5 Example

Table 4 shows a concrete example of the results of an evaluation experiment using ASPEC. This example describes "the absorption of glucose in the small intestine of a mouse". In the table, "Similar Target Sentences selected through Reranking" shows the similar sentences with which the largest score Q_{i^*} is obtained through reranking among the $k = 32$ output candidates, and "Output" shows the output results from the NMT model. The Sentence-BLEU for each sentence

while they are 98.0/94.2, 99.3/98.9, and 86.7/87.8 for oracle. These percentages are mostly close to the percentages of the numbers of the sentences within the target language monolingual TMs (excluding the target language side of the training data) against the total numbers of the target language sentences to be retrieved (i.e., 95.0, 98.8, and 88.1 for ASPEC, EUbookshop 100k and 1M, respectively).

Input	Study of the effect on the glucose absorption power , the TCDD exposed C57BL / 6J mouse increased the glucose absorption power in the intestinum tenue .		
Reference	グルコース吸収能に対する影響を検討した結果，ＴＣＤＤ暴露Ｃ５７ＢＬ／６Ｊマウスは小腸におけるグルコース吸収能が増加した．		
	Similar Target Sentence selected through Reranking	**Output**	**Sentence-BLEU**
w/o Similar Sentence	N/A	グルコース吸収能に対する効果の研究では，ＴＣＤＤがＰＢＬ／６Ｊマウスに暴露されたものは小腸のグルコース吸収能を増加させた．	21.0
sss+ed + top 1	また，ポンプ光と信号光との間の偏光角を変化させて，スイッチングに与える影響を実験により確認した．	耐糖能吸収能への影響の研究では，ＴＣＤＣＤが５７ＢＬ／６Ｊを搭載したが，小腸では耐糖吸収能は増加していた．	18.8
sss+ed +$Q^{(\text{Hossain})}$	また，ハタケシメジ投与は糞中胆汁酸排せつ量と，コレステロールを胆汁酸に変換するコレステロール７α－ヒドロキシラーゼの活性を上昇させた．	耐糖能吸収能への影響の研究では，ＴＣＤＣＤが５７ＢＬ／６Ｊを搭載したＴＣＤＤは小腸中での耐糖吸収能を増加した．	24.5
sss+ed +$Q^{(\text{proposed2})}$	また，ハタケシメジ投与は糞中胆汁酸排せつ量と，コレステロールを胆汁酸に変換するコレステロール７α－ヒドロキシラーゼの活性を上昇させた．	耐糖能吸収能への影響の研究では，ＴＣＤＣＤが５７ＢＬ／６Ｊを搭載したＴＣＤＤは小腸中での耐糖吸収能を増加した．	24.5
LaBSE + top 1	グルコースとしょ糖間の相互転化を触媒する細菌性（大腸菌）キシロースイソメラーゼの発現によって，発育中のジャガイモ塊茎の代謝におけるヘキソースの役割について調べた．	グルコース吸収電力に対する影響の研究，ＴＣＤＤがＣＭＳ／６Ｊマウスに暴露されたＣＣＤは小腸のグルコース吸収電力を増加させた．	14.0
LaBSE +$Q^{(\text{Hossain})}$	ヤギにおけるコレシストキニン（ＣＣＫ）の食欲減退効果を評価する目的で，２２時間粗食させた５頭の去勢ヤギに，塊状注入もしくは連続注入で第三脳室内にＣＣＫ（硫酸エステルＣＣＫ－８）を投与した．	グルコース吸収電力に対する効果の研究，ＴＣＤＤがＣｕｐ５７ＢＬ／６Ｊマウスに暴露された．	24.0
LaBSE +$Q^{(\text{proposed2})}$	正常マウスやＫＫ－Ａｙマウスを用いて，小腸におけるα－グルコシル二糖類の消化／吸収能力を比較した．	グルコース吸収電力，ＴＣＤＤ露光ＣＭＳ／６Ｊマウスのグルコース吸収電力に対する影響を研究した結果，小腸におけるグルコース吸収電力を増加させた．	**27.3**

Table 4: Examples of Results of Experiments with ASPEC

is calculated. Focusing on the similar sentences retrieved by sss+ed, the target language translation of the most similar source language sentence (i.e., top 1) is not a sentence in the biological field, and the similar target sentences selected through reranking by $Q^{(\text{Hossain})}$ and $Q^{(\text{proposed2})}$, although they are sentences in the biological field, do not have much information relevant to the reference sentence in terms of the content. On the other hand, when we focus on the similar sentences retrieved by LaBSE, we find that even the "top 1" sentence describes "glucose(グルコース)" and the similar target sentence selected through reranking by $Q^{(\text{proposed2})}$ describes "the absorption of sugars in mice (マウスにおける糖類の吸収)", which is the most relevant to the content of the reference translation. The highest value of Sentence-BLEU of the output candidate is also obtained by LaBSE+$Q^{(\text{proposed2})}$.

7 Conclusion

In this study, within the retrieve-edit-rerank framework, we introduced a method for cross-lingual retrieval of similar translations through multilingual sentence embedding, along with an enhanced reranking method. We demonstrated that utilizing vector neighborhood search, based on language-agnostic sentence embedding generation models like mSBERT and LaBSE, contributed to a significant improvement in translation accuracy within this framework. This proved more effective than the retrieval technique based on edit distance employed in the previous research. Moreover, we applied multiple similar sentences to generate various candidate translations, subsequently selecting the optimal translation through an automatic reranking process. The reranking score considered both the output log-likelihood normalized for the length of the reconstituted subword sentences, and the cosine similarity between the input and output candidate sentences through sentence embeddings. This methodology has led to a significant enhancement in translation accuracy.

References

Bulte, B. and Tezcan, A. (2019). Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proc. 57th ACL*, pages 1800–1809.

Cai, D., Wang, Y., Li, H., Lam, W., and Liu, L. (2021). Neural machine translation with monolingual translation memory. In *Proc. 59th ACL and 11th IJCNLP*, pages 7307–7318.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proc. 60th ACL*, pages 878–891.

Hossain, N., Ghazvininejad, M., and Zettlemoyer, L. (2020). Simple and effective retrieve-edit-rerank text generation. In *Proc. 58th ACL*, pages 2532–2538.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). ASPEC: Asian scientific paper excerpt corpus. In *Proc. 10th LREC*, pages 2204–2208.

Pagliardini, M., Gupta, P., and Jaggi, M. (2018). Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proc. NAACL-HLT*, pages 528–540.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. EMNLP and 9th IJCNLP*, pages 3982–3992.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proc. EMNLP*, pages 4512–4525.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proc. 54th ACL*, pages 1715–1725.

Skadiņš, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proc. 9th LREC*, pages 1850–1855.

Tezcan, A., Bulté, B., and Vanroy, B. (2021). Towards a better integration of fuzzy matches in neural machine translation through data augmentation. *Informatics*, 8(1):1–27.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proc. 8th LREC*, pages 2214–2218.

Vanallemeersch, T. and Vandeghinste, V. (2015). Assessing linguistically aware fuzzy matching in translation memories. In *Proc. 18th EAMT*, pages 153–160.

Xu, J., Crego, J., and Senellart, J. (2020). Boosting neural machine translation with similar translations. In *Proc. 58th ACL*, pages 1580–1590.