Can the Translation Memory Principle Benefit Neural Machine Translation? A Series of Extensive Experiments with Input Sentence Annotation

Yaling Wang
Graduate School of IPS
Waseda University
Kitakyushu, Japan
yaling.wang@moegi.waseda.jp

Yves Lepage
Graduate School of IPS
Waseda University
Kitakyushu, Japan
yves.lepage@waseda.jp

Abstract

Integrating translation memories (TM) into neural machine translation (NMT) has been shown to improve translation quality. We test various schemes to integrate translation suggestions into an NMT system without altering its architecture. We retrieve similar sentences covering the sentence to translate and examine various annotation schemes as input to the NMT system. Our results show that the method can outperform a baseline model in some cases. The improvements are mainly for the translation of sentences with a length ranging from 10 to 20 words.

1 Introduction

Translation Memories (TMs) are used daily by translators. They contain aligned parallel sentence pairs. Given a sentence to translate, a TM retrieves the most similar sentence in the source language that contains large common or similar parts. The corresponding sentence in the target language is returned to the translator. In this way, the translator needs only to modify the unmatched parts to complete the translation. A main advantage of translation memories is that they ensure consistency and *interpretability* across translations because common or similar parts in sentences can easily be identified.

Recently, with the development of neural machine translation (NMT), the quality of machine translation has significantly increased. Its main advantage is that it improves translation *efficiency* over previous machine translation techniques. However, the interpretability of NMT is poor: errors are difficult to interpret, i.e., to trace back to the training data.

Past research (Federico et al., 2012) already proposed to combine the advantages of TM (interpretability) with MT (efficiency). Recently, methods have been proposed to achieve closer integration with NMT. For example, an additional encoder can be added to an NMT architecture specifically for TM matches (Cao and Xiong, 2018). The decoding algorithm can be modified to incorporate retrieved strings (Gu et al., 2018). An easy-to-implement TM–NMT integration has been proposed by (Bulté and Tezcan, 2019): they concatenate the target-language side of matches retrieved from a TM with the sentence to translate. This only involves data pre-processing and augmentation. This is also compatible with different NMT architectures. All of the approaches above were shown to lead to a significant increase in the quality of MT outputs.

2 Method

In this paper, we propose to make use of the TM principle in conjunction with an NMT system, without altering the model architecture of the NMT system. In this way, the method can apply to any neural network architecture and can leverage pre-trained models. Figure 1 illustrates this method.

Suppose that we want to translate a sentence from English to German. We firstly retrieve English sentences from the parallel aligned data and obtain some similar English sentences which cover the input English sentence. For instance, the sentence to translate '*I want to go to school.*' is covered by the two following sentences '***I want to go to*** *hospital.*'

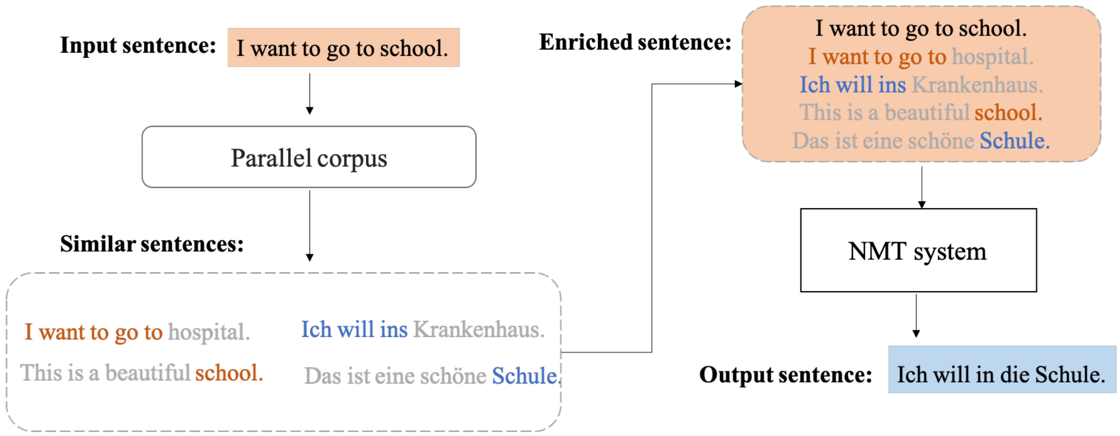

Figure 1: Overview of translation based on retrieval

and '*This is a beautiful* ***school.***'. Their corresponding German translations are obtained from the parallel data: '*Ich will ins Krankenhaus.*' and '*Das ist eine schöne Schule.*' That is, by retrieval, we acquire English–German sentence pairs, in which the English sentence is similar to the sentence to translate. The principle of translation memory postulates that the German sentences should also be similar to the German translation of the English sentence to translate. We use such translation pairs to enrich the input of an NMT system, i.e., we use them as annotations to the input sentence. We use such data to train an NMT system.

3 Enrichment Schemes

To emulate the principle of TM, we firstly retrieve a group of sentences in the source language that are similar to the sentence to translate. we use the tool introduced in (Liu and Lepage, 2021). Its goal is to cover a sentence in form and meaning with as few retrieved sentences as possible. It provides the possibility of retrieving sentences which are similar in form and meaning. Secondly, and to continue to emulate the TM principle, we obtain the corresponding sentences in the target part of the bilingual corpus.

However, we refine the principle of TM. The tool used for retrieval identifies the parts in the retrieved source sentences which are similar to the source sentence. Hence, based on these results, we use techniques in sub-sentential alignment from statistical machine translation, namely fast_align (Dyer et al., 2013), to obtain the corresponding translated parts in the sentences in the target language.

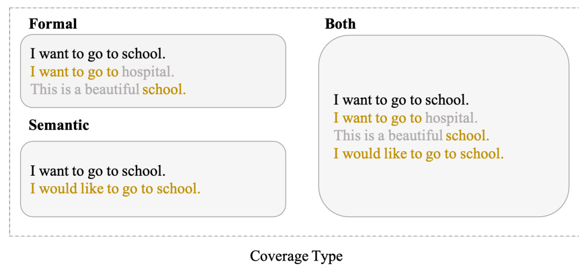

Figure 2: Illustration of different enrichment schemes that can be used according to the mode of retrieval (formal only, semantic only and both)

We now describe how we enrich the sentence to translate using results of retrieval and sub-sentential alignment. We propose different enrichment schemes to generate different possible inputs to the NMT system. Table 1 shows the different parameters which can be exploited under our settings. We describe them in details hereafter.

Coverage Type The tool used for retrieval provides two modes for similarity: formal and semantic similarity. Therefore, we can choose to retain sentences obtained by retrieval

- in form only;
- in meaning only;
- both.

Figure 2 illustrates results in these different modes.

Matched Parts Only or Whole Sentences From the retrieved sentences, we can choose to use:

Parameters	Options
Coverage Type	Formal only Semantic only Formal and Semantic
Matched parts	Matched parts only The whole sentences without markers The whole sentences with markers
Language Side	Source side only Target side only Source and Target sides
Order of Similar Sentences	All source sentences followed by all target sentences All target sentences followed by all source sentences Each source sentence followed by its corresponding target sentence, for all pairs of sentences Each target sentence followed by the source sentence it corresponds to, for all pairs of sentences

Table 1: List of different parameters that can be exploited to produce different enrichment schemes

Figure 3: Illustration of different enrichment schemes that can be used: with matched parts or whole sentences, with markers or not

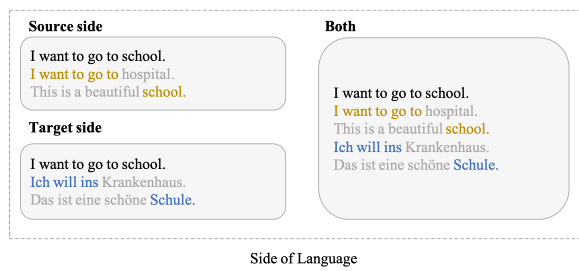

Figure 4: Illustration of different enrichment schemes that can be used depending on the language side used: source side only, target side only and both)

- only the matched parts, i.e., the parts which are similar to the sentence to translate (in form or in meaning, directly in the source language or by translation and sub-sentential alignment in the target language), and only these parts;
- the whole sentences retrieved with markers so as to identify the matched parts, in the source or the target language;
- the whole sentences retrieved without any markers to identify the matched parts.

Figure 3 provides an illustration of such possible cases.

Language Side Following the principle of TM, we obtain similar sentences in the source language by retrieval. Now, the corresponding sentences in the target language should contribute to translation. In terms of language side, we can thus choose

- the sentences retrieved in the source language only;
- the corresponding translations in the target language only;
- both: the sentences in the source language and their corresponding translations in the target language.

Figure 4 illustrates the above three possibilities.

Order of Similar Sentences When both language sides are chosen, we can imagine several enrichment schemes for the ordering of the sentences in the source language and the target language.

- All source sentences followed by all target sentences;
- All target sentences followed by all source sentences;
- Each source sentence followed by its corre-

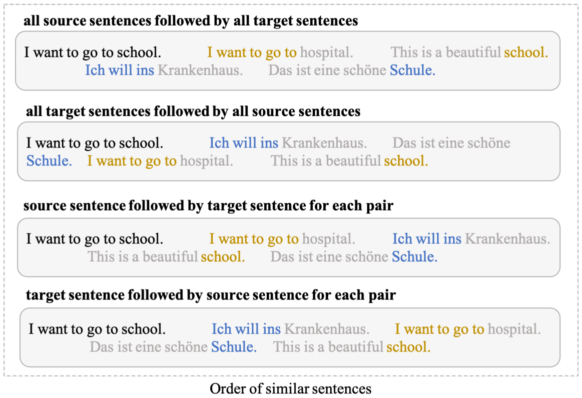

Figure 5: Illustration of different enrichment schemes that can be used for the ordering of similar sentences when both language sides are used

sponding target sentence for all pairs of sentences;

- Each target sentence followed by the source sentence it corresponds to, for all pairs of sentences.

Figure 5 shows an example.

List of all Possible Enrichment Schemes All possible choices for each of the parameters enumerated above lead to a list of 54 possible enrichment schemes for the exploitation of the information obtained following the TM principle. Table 5 lists them all.

4 Experimental Setup

4.1 Data

We use Multi30k (Elliott et al., 2016) as our parallel corpus. It contains multilingual image descriptions for multilingual and multimodal research. We use the German, English and French parts in our experiments. Some statistics are given in Table 2. We split the dataset into 3 parts, 80% for training, 10% for validation and 10% for testing. We perform translation experiments in all possible directions offered by the three languages, i.e., 6.

4.2 Evaluation

Following standard practice, evaluation of translation is done by computing the BLEU score (Papineni et al., 2002) on the test set. We report BLEU scores in the range of 0 to 100. BLEU scores indicate similarity to the reference translation in form only.

Hence, in addition to BLEU scores, for the purpose of measuring semantic similarity with the reference, we compute BERTScores (Zhang et al., 2020). BERTScores leverage pre-trained contextual embeddings from BERT and match words in the candidate and the reference sentences using cosine similarity. We report the F-measure, which ranges from 0 to 1. Higher BERTScores indicate higher similarity in meaning between the candidate and the reference sentences.

lang.	# sents.	vocab. size	avg. length (in tokens)
de	30,014	18,722	12.44
en	30,014	10,214	13.02
fr	30,014	11,794	13.62

Table 2: Statistics on the corpus (Multi30k)

Encoder	
Type	LSTM
Embedding Dimension	500
Number of layers	2
Size of hidden layer	500
Decoder	
Type	StackedLSTM
Embedding Dimension	500
Number of layers	2
Size of hidden layer	500
Total # of parameters	18,368,003
Optimizer	SGD
Learning rate	1.0

Table 3: Configuration for the NMT model

4.3 Baseline System

We compare our proposal to a baseline. Our baseline model is trained using the same NMT model but simply with the sentences to translate without anything else, as input.

Our NMT model follows the Seq2seq architecture (Bahdanau et al., 2015) implemented in the OpenNMT-py toolkit (Klein et al., 2017). The model configuration is shown in Table 3.

5 Experiment Results

5.1 Results for Retrieval

For each of the sentences in the English, German and French test sets, we apply the TM principle and retrieve similar sentences.

Retrieval in Form For the results of retrieval in form, we focus on the number of similar sentences retrieved per input sentence. This is because the retrieval method used aims at maximal coverage with the least possible number of retrieved sentences. A lesser number of similar sentences means that the common parts are longer.

Table 4 gives some statistics on the results of retrieval. The number of retrieved sentences in the 3 languages is similar. The average number of retrieved sentences is about 5, which means that 5 n-grams in the retrieved sentences almost cover the input sentence. The value of the standard deviation is also relatively small. The most frequent number of retrieved sentences is 4, 5 or 6. There are only few cases where the number of retrieved sentences is greater than 10. This means that, in general, the retrieval method used covers the sentence to translate with a relatively small number of similar sentences.

Retrieval in Meaning Table 4 gives some statistics on the results of semantic retrieval. Compared to retrieval in form, the number of retrieved sentences is less. This is because the method used selects the top k sentences that contribute to the increase in coverage of the input sentence. These sentences are a supplement.

5.2 Translation Results with Different Enrichment Schemes

We measure the performance of different enrichment schemes and select the scheme that performs the best. The translation task that we consider is from German to English, so that we use the German sentences for the retrieval step. Figure 8 shows examples of translations obtained using different enrichment schemes and Table 5 gives the results of evaluation for all possible different enrichment schemes.

To analyze the results, we draw box plots by groups of four parameters (coverage type, matched parts, side and ordering), in Figures 6 and 7. The

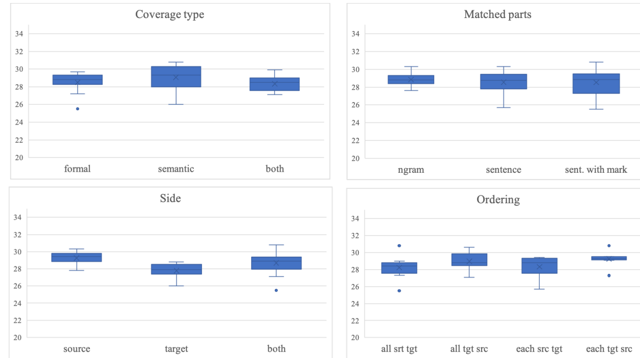

Figure 6: Box plots by coverage type, matched part/whole sentence, language side and ordering on BLEU scores.

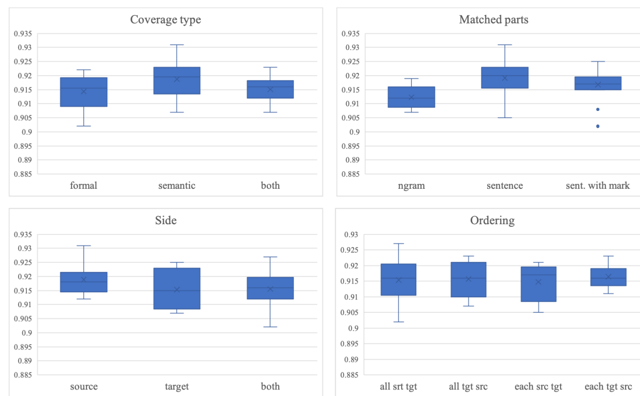

Figure 7: Box plots by coverage type, matched part/whole sentence, language side and ordering on BERTScores.

box plot show six ranges for the results: upper edge, upper quartile, median, lower quartile, lower edge, and outliers.

In terms of coverage type, among three coverage types, the median BLEU scores of formal coverage and semantic are similar. The BERTScore with semantic coverage is the highest. This is expected because for semantic retrieval, retrieves sentences that have similar meanings by definition.

The results of only using matched parts for translation is found to be the most stable with the smallest standard deviations. However, the performance across the three possible choices (only matched part, sentence, sentence with marker) is close in scores.

In terms of language side, among the three possible options, using source information only performs the best in BLEU with an average score of 29.26.

Retrieval	Language	# of retrieved sentences			Average length	
		mean ± stdev.	median	mode	in tokens	in char.
Formal	de	5.03 ± 2.22	5	4	62	352
	en	5.50 ± 2.22	5	5	71	332
	fr	5.40 ± 2.32	5	4	76	395
Semantic	de	2.81 ± 1.42	3	2	36	209
	en	2.22 ± 1.14	2	2	29	137
	fr	2.69 ± 1.33	2	2	38	199

Table 4: Statistics of results for formal (top) and semantic retrieval (bottom)

No.	Retrieval	Matched parts	Side	Ordering	BLEU	BERTScore
1	Formal only	n-gram	Source only	-	28.8	0.923
2		sentence		-	29.5	0.926
3		sent. with markers		-	29.7	0.927
4		n-gram	Target only	-	28.3	0.919
5		sentence		-	28.7	0.920
6		sent. with markers		-	27.2	0.917
7		n-gram	Source and Target	all src. tgt.	28.2	0.918
8		n-gram		all tgt. src.	28.4	0.921
9		n-gram		each src. tgt.	29.3	0.925
10		n-gram		each tgt. src.	29.1	0.923
11		sentence		all src. tgt.	28.5	0.927
12		sentence		all tgt. src.	29.4	0.926
13		sentence		each src. tgt.	25.7	0.914
14		sentence		each tgt. src.	29.3	0.926
15		sent. with markers		all src. tgt.	25.2	0.912
16		sent. with markers		all tgt. src.	28.8	0.925
17		sent. with markers		each src. tgt.	29.4	0.925
18		sent. with markers		each tgt. src.	29.3	0.923
19	Semantic only	n-gram	Source only	-	30.3	0.923
20		sentence		-	29.4	0.932
21		sent. with markers		-	28.9	0.926
22		n-gram	Target only	-	27.9	0.924
23		sentence		-	28.0	0.927
24		sent. with markers		-	26.0	0.927
25		n-gram	Source and Target	all src. tgt.	29.0	0.921
26		n-gram		all tgt. src.	28.5	0.918
27		n-gram		each src. tgt.	29.3	0.920
28		n-gram		each tgt. src.	29.3	0.921
29		sentence		all src. tgt.	27.8	0.929
30		sentence		all tgt. src.	30.3	0.928
31		sentence		each src. tgt.	27.6	0.926
32		sentence		each tgt. src.	29.5	0.927
33		sent. with markers		all src. tgt.	30.8	0.925
34		sent. with markers		all tgt. src.	30.6	0.928
35		sent. with markers		each src. tgt.	29.4	0.924
36		sent. with markers		each tgt. src.	30.8	0.925

No.	Retrieval	Matched parts	Side	Ordering	BLEU	BERTScore
37	Formal and Semantic	n-gram	Source only	-	29.9	0.926
38		sentence		-	27.8	0.921
39		sent. with markers		-	29.0	0.925
40		n-gram	Target only	-	27.6	0.917
41		sentence		-	28.8	0.927
42		sent. with markers		-	27.6	0.923
43		n-gram	Source and Target	all src. tgt.	28.6	0.923
44		n-gram		all tgt. src.	28.6	0.918
45		n-gram		each src. tgt.	28.8	0.919
46		n-gram		each tgt. src.	29.2	0.921
47		sentence		all src. tgt.	27.3	0.921
48		sentence		all tgt. src.	29.0	0.924
49		sentence		each src. tgt.	27.9	0.925
50		sentence		each tgt. src.	29.5	0.925
51		sent. with markers		all src. tgt.	28.4	0.925
52		sent. with markers		all tgt. src.	27.1	0.923
53		sent. with markers		each src. tgt.	27.5	0.923
54		sent. with markers		each tgt. src.	27.3	0.924

Table 5: All possibilities of formats with results of evaluation. All confidence intervals for the BLEU scores are between 0.75 and 0.85.

As for the order of similar sentences, the second order (all target sentences followed by all source sentences) and the fourth order (each target sentence followed by the source sentence it corresponds to, for all pairs of sentences) perform better than the other ones in BLEU. This shows that giving target information before source information is a better choice.

All in all, to select the best combination of four parameters among all the possible formats through BLEU score and BERTScore, we notice that a higher BLEU score is not always accompanied by a higher BERTScore. We want the translations to be close to the reference translations not only in form but also in meaning. Hence, we sort all configurations using the average of the BLEU scores (recast from 0 to 1) and BERTScores and select the configuration ranked the highest. It is configuration No. 36. We apply this best configuration in all other translation directions.

5.3 Translations in Different Languages

We perform machine translation experiments in all directions of all languages pairs between German, English and French. This is 6 directions in total. We use enrichment scheme No. 36, i.e., results of semantic retrieval only, using whole sentences with matching parts indicated with markers, each target sentence followed immediately by the source sentence it corresponds to, for all pairs of retrieved sentences. Table 6 summarizes the translation results.

When using formal coverage retrieval results, our models outperform the baseline model in three translation tasks: de→en, de→fr and fr→en. In the other cases, although our models do not exceed the baseline system, confidence intervals, as shown in Table 6, indicate that the baseline model and our models perform similarly. For instance, for the direction en→de, confidence intervals of $\pm$ 0.8 do not allow to say that a baseline of 27.4 is really better than our model with 27.1. As the main difference is the language of query sentences, i.e., the source language, we might think that the differences in BLEU observed by the difference in morphology of the source and target languages explain the results. In general, the result shows that the formal coverage retrieval method contributes to improving the translation quality or performs similarly compared to the baseline system.

When using semantic coverage retrieval, our

Translation direction			Baseline	Proposed method	
				Formal coverage	Semantic coverage
de	→	en	29.6 ± 0.8	**30.5** ± 0.9	**30.6** ± 0.8
de	→	fr	30.6 ± 0.8	**31.8** ± 0.8	29.7 ± 0.8
en	→	de	27.4 ± 0.8	27.1 ± 0.8	26.1 ± 1.0
en	→	fr	42.2 ± 1.2	41.8 ± 1.2	**47.2** ± 1.0
fr	→	de	24.3 ± 0.8	24.1 ± 0.8	23.6 ± 0.8
fr	→	en	38.8 ± 0.9	**39.6** ± 0.9	**42.5** ± 1.2

Table 6: Translation results (in BLEU) for each different translation directions

No.	Sentence to translate	Output translations	Reference translation
3	ein mann in einem gelben oberteil macht eine inspektion an einem schwinn-fahrrad neben einem picknicktisch .	a man in a yellow top is making a inspektion at a schwinn-fahrrad .	man in yellow shirt performing maintenance on schwinn bicycle near a picnic table .
9		a man in a yellow top is taking a break by a picnic table next to a picnic table .	
15		a man in a yellow top is taking a trick by a picnic table .	
41	ein mann auf einem motorrad und zwei weitere männer auf einem wagen fahren auf einer staubigen zweispurigen straße .	a man on a motorcycle and two other men on a wagen on a sunny road .	a very man on a motorcycle and 2 men on a cart are traveling down a dusty two lane road .
35		a man on a motorcycle and two other men riding on a dusty bike .	
38		man on a motorcycle and two more men on a dusty road .	
42	ein ball befindet sich zwischen einem werfer und einem fänger auf dem baseballfeld .	a ball is in between a werfer and batter on the baseball field .	a pitcher and catcher on a baseball field with the ball in between them .
41		a ball is between a werfer and baseball on the baseball .	
40		a ball is between a werfer and a fänger on the baseball .	

Figure 8: Examples of translation using different formats

Input sentence	**Translation**	**Reference**
one lady in a plaid coat eating cotton candy .	une femme en manteau à carreaux mange de la barbe .	une femme en veste écossaise mangeant de la barbe à papa .
two men and a woman are inspecting the front tire of a bicycle .	deux hommes et une femme untersuchen le vorderrad d' un vélo .	deux hommes et une femme inspectent le pneu avant d' un vélo .
un petit chien avec un ruban rouge sur sa tête marche dans l' herbe .	ein kleiner hund mit einer roten ruban auf seinem kopf .	ein kleiner hund mit einem roten band auf dem kopf läuft durch das gras .
trois femmes en rouge de l' équipe de basket russe suivant le ballon .	drei frauen in roter équipe suivant suivant .	drei frauen in roten trikots aus der russischen basketballmannschaft laufen dem basketball hinterher .
ein thaiboxer übt zum aufwärmen vor dem kampf einen beinhochtritt .	a thaiboxer band is practicing for the aufwärmen in front of the net .	this thai boxer is practicing a high leg kick as a warm up before his fight .
ein mann mit einem rucksack springt von einem pier .	a man with a backpack jumps off a pier .	a man wearing a backpack is jumping off a pier .

Figure 9: Random examples in different translation directions

models outperform the baseline model in three translation tasks: de→en, en→fr and fr→en. This is the same number as for formal coverage, but one language direction is different: en→fr instead of de→fr. A large improvement is obtained in the direction: fr→en. In this translation task, the model using semantic coverage retrieval outperforms the baseline model by 3.7 BLEU points, which is largely more than the model using formal coverage retrieval. Our method leads to an even larger improvement in the translation task en→fr using semantic coverage retrieval. The BLEU score increases by 5.0 points over the baseline model, whereas the model using formal coverage retrieval does not exceed the baseline system. We conclude that our proposed method with semantic coverage is especially efficient for the language pair en–fr, in both directions.

Figure 9 shows some examples of translation results. (input sentence is just source sentence without enrichment)

5.4 Length of the sentence to translate

Based on some samples, we found that our model delivers similar performance as the baseline model for shorter sentences (length less than ten words). However, our model offers better translations for sentences between 10 and 20 words due to the information found in similar sentence pairs. In order to confirm the impression left by this observation, we split the test set into three parts by the length of the sentence to translate, and we compare the performance on these three separate subsets.

Length of sentences	# of sentences	BLEU score	
		Baseline	Ours
<10	448	31.3	31.0
10–20	2,207	30.1	**31.1**
>20	247	25.9	25.5

Table 7: Translation results for different sentence lengths (in BLEU, de→en)

Table 7 shows the results for the three separate subsets containing sentences with different lengths. The sentences of a length between 10 and 20 words account for the most part of the test set. Our model outperforms the baseline model on this subset by 1.0 BLEU point. However, for sentences of length more than 20, both models cannot perform well.

6 Conclusion

We proposed to test whether the principle of translation memory (TM) can benefit results in neural machine translation (NMT). We enriched the input of the NMT system with such sentences retrieved. We studied different annotation schemes, and found that the scheme which delivers the best translation accuracy consists in providing the target sentence immediately before its corresponding source sentence, for all sentence pairs, and identifying matching parts with markers.

Such enrichment schemes can contribute to the interpretability of the results obtained by neural machine translation systems. The results of our translation experiments show that, for some translation tasks, our system performs better than a standard NMT system without retrieval. Increases in translation accuracy are mainly obtained for sentences with a length in the range of 10 to 20 words.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bram Bulté and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Qian Cao and Deyi Xiong. 2018. Encoding gated translation memory into neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3047, Brussels, Belgium. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648. The Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2018. Search engine guided neural machine translation. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pages 5133–5140. AAAI press.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Yuan Liu and Yves Lepage. 2021. Covering a sentence in form and meaning with fewer retrieved sentences. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (PACLIC 35)*, pages 1–10.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.