Leveraging Knowledge from Translation Memory for Globally and Locally Guiding Neural Machine Translation

Ruibo Hou **Hengjie Liu** **Yves Lepage**
Graduate School of Information, Production and Systems, Waseda University
houruiboc@akane.waseda.jp yo4c5ama@toki.waseda.jp yves.lepage@waseda.jp

Abstract

Neural Machine Translation (NMT) models augmented with Translation Memory (TM) have demonstrated success across various translation scenarios. In contrast to previous methodologies that primarily rely on either semantic or formally matched sentences from TM, or simply concatenate these augmented sentences together, our proposed approach aims to more effectively and explanatorily utilize both types of retrieved sentences from TM. Semantically matched sentences that cover the entire source sentence are used to guide the overall translation process, while formally matched sentences which cover source sentence partially are leveraged to guide the translation of specific segments. This refined methodology enables us to exploit knowledge from TM more effectively, thereby enhancing translation quality. Experimental results demonstrate that our framework not only achieves performance that is competitive with other strong baselines when applied to high-resource datasets, but also yields improvements over non-TM-augmented NMT systems in low-resource scenarios.

1 Introduction

Retrieval-Augmented Generation (RAG) methods (Khandelwal et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021) leverage non-parametric memory through retrieval to enhance parametric generative models, thereby enabling these models to effectively access and incorporate knowledge beyond their intrinsic parameters. Retrieval-augmented methods have numerous applications in the field of Natural Language Processing (NLP); For the Machine Translation (MT) task, Retrieval-Augmented Machine Translation (RAMT) aims to find relevant knowledge from a Translation Memory (TM) and leverages it to improve MT performance. A TM archives source sentences paired with their corresponding human translations. Upon retrieving a match, the translator is provided with similar source sentences and their translations. Early works (Utiyama et al., 2011; Liu et al., 2012) integrates TM with Statistical Machine Translation (SMT) systems to achieve better translation performance.

Recent research has demonstrated that integrating TM with Neural Machine Translation (NMT) can lead to significant improvements. This enhancement has been achieved through various approaches, including concatenating sentences retrieved from TM with the source input (Bulte and Tezcan, 2019; Xu et al., 2020), encoding retrieved sentences from TM and the source input separately (Gu et al., 2018; Xia et al., 2019; Cao et al., 2020; He et al., 2021), retrieving sentences from TM contrastively rather than greedily (Cheng et al., 2022), and leveraging fuzzy-matched sentences from TM by non-autoregressive machine translation models (Xu et al., 2023). The aforementioned works adopt non-trainable retrieval tools to retrieve similar sentences from the TM. In contrast, Cai et al. (2021) utilize a trainable retrieval model to retrieve relevant sentences from monolingual corpora.

However, previous research has two limitations. Firstly, some studies (Bulte and Tezcan, 2019; He et al., 2021; Xu et al., 2023), among others, focus on leveraging either semantically similar or formally similar sentences from the TM to enhance NMT. Other works, while utilizing both types of similar sentences from the TM, handle them identically and merely concatenate them with the source sentence. This leads to inefficient use of knowledge from the TM. Secondly, most existing works do not consider applications in low-resource settings, while other works require an external dataset beyond the training dataset to serve as a TM for retrieval. When utilizing only the training dataset as a TM for retrieval, it often fails to improve and may even harm translation performance compared to non-TM-augmented NMT models in low-resource scenarios.

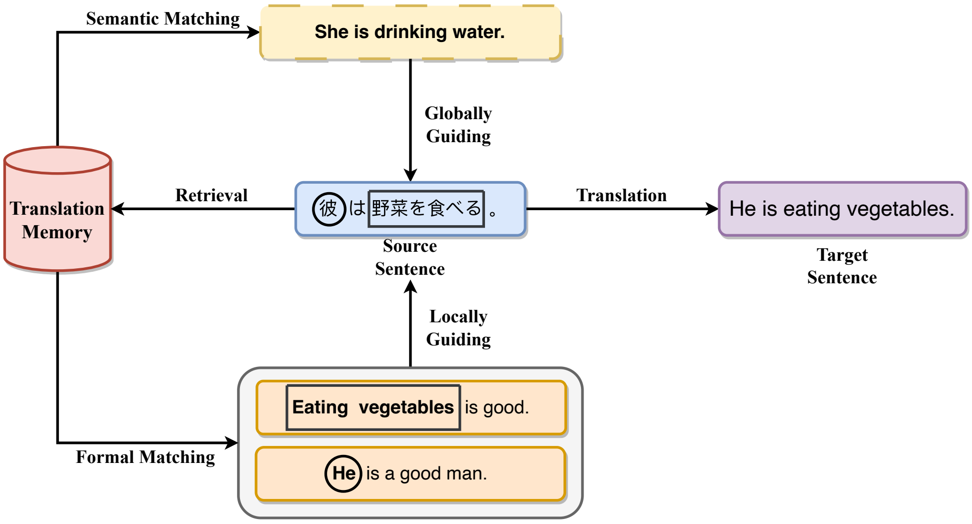


Figure 1: Overall sketch of our proposed method. Semantic matching (above) and formal matching (below) are performed separately, and are respectively guiding the translation at the global and local levels.

As Figure 1 illustrates, we propose globally guiding translation using semantically retrieved (entire match) sentences from TM, while leveraging formally retrieved (partial match) sentences for local guidance. We process the two types of retrieved sentences separately using different methods to emphasize their distinct roles in enhancing translation. In contrast, aiming to reduce reliance on external data, we emphasize maximizing acquisition of knowledge from the internal training set. Our main contributions are:

- By separately processing the semantically matched and formally matched sentences retrieved from TM, our approach globally and locally guides the translation process, enabling us to leverage the knowledge in the TM more effectively.
- Experimental results demonstrate that our model can achieve competitive performance compared to other strong baselines on high-resource datasets, and crucially, outperform non-TM-augmented NMT systems in low-resource scenarios without relying on external datasets beyond the training dataset.

2 Methodology

2.1 Overview

Our approach comprises two key components: retrieval from the TM (§2.2) and the integration of the retrieved sentences to guide translation (§2.3 - §2.5). Given a source sentence x, we perform semantic matching within the TM to retrieve a semantically matched sentence and obtain its corresponding target translation *smt*. Additionally, we conduct formal matching to retrieve a set of formally matched sentences and leverage word alignment to identify their related translated segments, denoted as the set of formally matched pieces $\{fml\}_{i=1}^{M}$. Our work employs a transformer architecture (Vaswani et al., 2017) with dual encoders to jointly capture global and local contextual information from the augmented sentences. Specifically, *smt* is concatenated with source sentence x and encoded by the global knowledge encoder (§2.3) to provide global guidance. The formally matched pieces are encoded by the local knowledge encoder (§2.4) for local guidance. The representations from both encoders are then fused with the decoder representations (§2.5). Here, to better utilize the local information contained in the formally matched pieces to assist with the translation, same as (Cai et al., 2021; Cheng et al., 2022), we employ a copy module (Gu et al., 2016; See et al., 2017) in the decoding process. The overview of the framework is illustrated in Figure 2. We first leverage the global knowledge contained in semantically matched sentences to enhance the overall translation process. Subsequently, formally matched pieces guide the translation of local segments within the sentence. In this context, the copy module in the decoder can be viewed as a post editor enriched with local knowledge. Through this design, we can effectively harness the knowledge from TM to facilitate and inform the translation task.

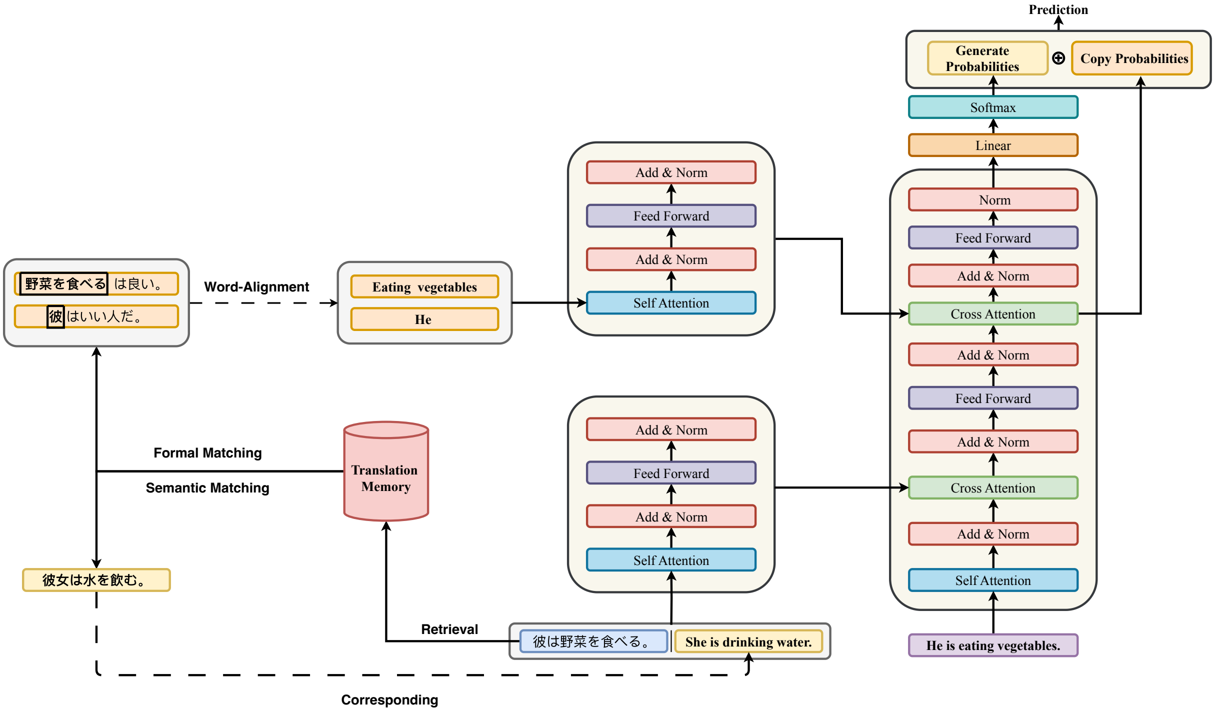


Figure 2: Overview of the architecture of the proposed model. The first cross-attention layer in the decoder incorporates global knowledge from semantically matched sentences to improve the translation process. Then, the second cross-attention layer uses formally matched pieces to guide the translation of specific parts within the sentence. Here we do not present specific layer configurations; the details of model layer settings are described in §3.2.

2.2 Retrieving Sentences from TM

Semantic Matching In our work, SBERT (Reimers and Gurevych, 2019) is used to generate distributed sentence representations. We define the semantic similarity between two sentences s_1 and s_2 as the cosine similarity in the sentence embedding space:

$$\mathrm{sim}(s_1, s_2) = \cos(\mathrm{Emb}(s_1), \mathrm{Emb}(s_2)) \quad (1)$$

where $\mathrm{Emb}(\cdot)$ denotes the SBERT encoder function. For given source sentence x, we retrieve sentences from the TM that have a semantic similarity exceeding a predetermined threshold θ. The corresponding translations of these retrieved sentences, referred to as *smt*, are then used to guide the translation process from a global perspective. To accelerate retrieval between the input vector representation and the corresponding vector of sentences in the TM, we utilize the FAISS toolkit (Johnson et al., 2019). After that, we concatenate the two sentences x and *smt*, using the token ' | ' to mark the boundary between them.

Formal Matching N-gram matching is utilized to find sentences in the TM that contain lexically overlapping pieces with the source input. We utilize the fscov toolkit (Liu and Lepage, 2021) for N-gram retrieval and use mask-align (Chen et al., 2021) to train a word-alignment model on each training set to generate word alignments between source and target phrases. Using word alignment allows us to avoid the need to use a threshold to filter sentences. For a source sentence x, we obtain several formally matched pieces $\{\mathit{fml}\}_{i=1}^{M}$, which are expected to appear in the target sentence y. Instances of fml_i are presented in Table 3.

2.3 Global Knowledge Encoder

Initially, we input the concatenation of the source sentence x and a a semantically matched sentence *smt* into the global knowledge encoder:

$$z^{x\&smt} = \mathrm{Enc}(\mathrm{Concat}(x, \mathit{smt})) \quad (2)$$

2.4 Local Knowledge Encoder

For formally matched pieces $\{\mathit{fml}\}_{i=1}^{M}$, each individual piece fml_i undergoes separate encoding within the local knowledge encoder. We obtain dense representations for all formally matched pieces, formulated as:

$$z^{fml} = \mathrm{Enc}(\{\mathit{fml}\}_{i=1}^{M}) \quad (3)$$

2.5 Decoder for Fusing Information

For a target sentence y, at each step t, we obtain a hidden representation h_t after token embedding layer and self-attention layer. Then the initial cross-attention layer incorporates information from the source sentence and semantically matched sentence:

$$\hat{h}_t = \mathrm{CrossAttn}(\mathrm{Add\&Norm}(h_t), z^{x\&smt}, z^{x\&smt}) \tag{4}$$

which is subsequently passed through a feed-forward network:

$$\widetilde{h}_t = \mathrm{FFN}(\mathrm{Add\&Norm}(\hat{h}_t)) \tag{5}$$

then passed through another add and normalization layer:

$$\overline{h}_t = \mathrm{Add\&Norm}(\widetilde{h}_t) \tag{6}$$

For the formally matched pieces, an additional cross-attention layer is employed, where a copy module (Gu et al., 2016; See et al., 2017) is applied, the implementation being the same as (Cai et al., 2021). Specifically, for each formally matched piece fml_i, there exists a sequence of contextualized tokens $\{\mathit{fml}_{i,k}\}_{k=1}^{L_i}$, where L_i denotes the length of the token sequence fml_i. In this cross-attention layer, we have:

$$c_t = W_c \sum_{i=1}^{M} \sum_{j=1}^{L_i} \alpha_{ij} z^{fml_{i,j}} \tag{7}$$

Here, α_{ij} denotes the attention score assigned to the j-th token in fml_i, $z^{fml_{i,j}}$ is the corresponding dense representation vector, c_t constitutes a weighted combination of embeddings of all tokens in formally matched pieces, and W_c is a trainable matrix.

The cross-attention mechanism is leveraged twice during the decoding phase. Initially, given the $t-1$ previously generated tokens and the corresponding hidden state $\overline{h}_t$, the decoder's hidden state is updated by incorporating the weighted sum c_t of token embeddings from the formally matched pieces, which can be formulated as: $\overline{h}_t = \overline{h}_t + c_t$. Subsequently, each attention score is interpreted as the probability of copying the corresponding token. The next-token probabilities are calculated as:

$$p(y_t|\cdot) = (1-\lambda_t)P_v(y_t) + \lambda_t \sum_{i=1}^{M} \sum_{j=1}^{L_i} \alpha_{ij} \delta_{fml_{i,j}, y_t} \tag{8}$$

In the above equation, δ represents the indicator function, and λ_t is a gating variable computed by another feed-forward network $\lambda_t = \mathrm{FFN}(\overline{h}_t, c_t)$. $P_v(y_t)$ is the probability distribution over the token y_t obtained from the final hidden state through a linear projection followed by a softmax function, representing the probability of generating the next token from a fixed vocabulary.

3 Experimental Setup

3.1 Dataset and Evaluation

High-Resource Dataset Settings For the task of enhancing NMT performance by incorporating TM in high-resource settings, we use the JRC-Acquis corpus (Steinberger et al., 2006), which is is a compilation of legislative texts from the European Union that are applied uniformly across EU member states. Following established practices, we split the dataset into training, development, and test subsets, in line with previous studies (Gu et al., 2018; Zhang et al., 2018; Xia et al., 2019; Cai et al., 2021; Cheng et al., 2022). In particular, we direct our empirical evaluation towards two language pairs, the translation from English to Spanish (en→es) and the translation from English to German (en→de).

Low-Resource Dataset Settings To assess the effectiveness of our approach in low-resource settings, we employ the WMT20 German to Upper Sorbian (de→hsb) dataset[1]. This corpus comprises 60,000 parallel sentences for training, accompanied by 2,000 sentences for each of the development and test sets. Additionally, we utilize the WMT22 German to Lower Sorbian (de→dsb) dataset[2], which contains 40,194 sentences for training and 1,353 sentences designated for development. Since only the development set is publicly available on the website, we perform a random shuffle and split it into two equal partitions to serve as validation and test sets, respectively.

Evaluation Following standard practice, we use SacreBLEU (Post, 2018) for evaluation, which is a standardized implementation of the widely adopted BLEU metric (Papineni et al., 2002).

[1] https://statmt.org/wmt20/unsup_and_very_low_res/

[2] https://statmt.org/wmt22/unsup_and_very_low_res.html

Configuration	en→de	en→es	de→hsb	de→dsb
Base	55.15 ± 1.40	61.31 ± 1.08	40.91 ± 1.27	27.02 ± 2.37
+Semantic	57.46 ± 1.53	62.77 ± 1.09	41.85 ± 1.25	27.65 ± 2.50
+Formal	57.55 ± 1.49	62.78 ± 1.08	39.53 ± 1.23	24.74 ± 2.36
+Semantic+Formal	**58.45 ± 1.48**	**63.19 ± 1.04**	**42.66 ± 1.27**	**28.28 ± 2.39**

Table 1: Experimental results (BLEU scores) on each test set with different TM-integrating configurations.

3.2 Implementation Details

We employ byte pair encoding (BPE) (Sennrich et al., 2016) for word segmentation in our work. For the high-resource machine translation task, the vocabulary size is capped at 20,000 subword units per language, while in low-resource scenarios, it is limited to 8,000 subword units per language. The threshold for semantic similarity, denoted as θ, is set to 0.8 for high-resource tasks as in (Xu et al., 2020) and lowered to 0.5 for the low-resource setting to accommodate the data scarcity. For a given source sentence, we retrieve up to 5 semantically most relevant sentences from the TM, effectively setting the top-k retrieval size to 5. During validation and testing, there is no threshold for semantic matching; only the most semantically similar sentence is concatenated with the source sentence. Regarding the number of formally matched pieces leveraged for augmenting the translation, denoted as $|M|$, for the high-resource tasks, we employ the two longest pieces, while for the low-resource setting, this number is reduced to the single longest piece. Regarding the setting of the number of layers, consistent with (Cai et al., 2021), the global knowledge encoder and decoder have 6 layers, while the local knowledge encoder has 4 layers. In all our experiments, we adopt the learning rate schedule, label smoothing settings and optimizer configurations as outlined in (Vaswani et al., 2017).

3.3 Ablation Study

To systematically investigate the effects of incorporating TM sentences through different retrieving methods, and analyze the contribution of each component in our proposed model, we conduct a series of ablation studies with the following TM-integrating configurations:

- **Base**: A base transformer model without access to any augmented sentences from a TM.
- +**Semantic**: A base transformer model, where the encoder takes as input the concatenation of the source sentence and a sentence retrieved from a TM via semantic matching.
- +**Formal**: A dual-source transformer model, where one encoder takes the source sentence as input, and the other encoder takes formally matched pieces retrieved from a TM as input.
- +**Semantic**+**Formal**: The proposal of this paper, i.e., a dual-source transformer model, where one encoder inputs a concatenation of source sentence and a semantically matched sentence from a TM, the other takes formally matched pieces as input.

4 Experimental Results and Analysis

4.1 Results

Comparison with Ablation Studies Based on the results of the ablation studies (Table 1), we observe that augmenting translation models with both semantically and formally matched sentences retrieved from the TM is the optimal configuration across both high-resource and low-resource datasets. In high-resource scenarios, our method achieves up to a 3.30 BLEU improvement over the non-TM baseline on the test set (en→de). Notably, our proposed approach outperforms non-TM-augmented NMT systems on the two low-resource datasets without reliance on external datasets. Our method outperforms the non-TM baseline by up to 1.75 BLEU points on the test set (de→hsb). This finding demonstrates that our method effectively leverages the knowledge encapsulated within the TM to enhance NMT translation, delivering improvements in scenarios spanning from high-resource to low-resource settings.

Comparison with Other Methods As shown in Table 2, we compare our approach with other TM-augmented NMT systems on the high-resource JRC-Acquis dataset. In both English to German and English to Spanish translation tasks, our system achieves competitive results that closely approach the state-of-the-art (Cai et al., 2021; Cheng et al.,

System	en→de	en→es
(Gu et al., 2018)	48.80	57.27
(Zhang et al., 2018)*	55.14	61.56
(Xia et al., 2019)	56.88	62.76
(Cai et al., 2021)	58.42	63.86
(Cheng et al., 2022)	58.69	64.04
Ours	58.45	63.19

Table 2: Comparing with results of other methods on JRC-Acquis dataset. *The results for (Zhang et al., 2018) are given in (Xia et al., 2019). All other results are from the respective paper.

x	2. The decision to impose surveillance shall be taken by the Commission according to the procedure laid down in Article 16 (7) and (8).	BLEU
y	(2) Der Beschluss über die Einführung einer Überwachung wird von der Kommission nach dem Verfahren des Artikels 16 Absätze 7 und 8 gefasst.	
smt	**Die Verfahren für die Durchführung von Kommissionsinspektionen werden nach dem in Artikel 16 Absatz 2 genannten Verfahren beschlossen.**	
fml_1	von der Kommission Nach dem Verfahren des Artikels	
fml_2	Beschlüsse *über die* Einführung einer Überwachung	
y^{Base}	(2) Die Kommission beschließt über die Einführung einer Überwachung nach dem Verfahren des Artikels 16 Absätze 7 und 8.	49.40
$y^{+Semantic}$	**(2) Der Beschluss zur Einführung einer Überwachung wird von der Kommission nach dem Verfahren des Artikels 16 Absätze 7 und 8 gefasst.**	85.46
$y^{+Semantic+Formal}$	**(2) Der Beschluss** *über die* **Einführung einer Überwachung wird von der Kommission nach dem Verfahren des Artikels 16 Absätze 7 und 8 gefasst.**	100.00

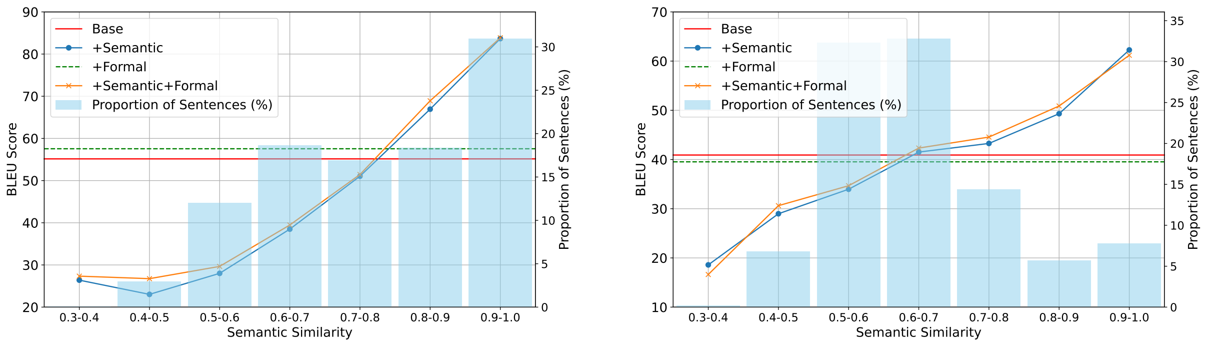
Table 3: The following are translation examples from experiments done on the English to German (en→de) dataset. The semantically similar sentences guide the translation globally, resulting in a better translation compared to the base model. By jointly using formally similar sentences to guide the sentence translation at a local level by the copy module, we achieve an even better translation. For clarity of presentation, all sentences are in untokenized form.

2022), with particularly strong performance on the English to German dataset.

4.2 Analysis

Could Our Method Guide the Translating Process Globally and Locally? Table 3 demonstrates how the global and local information contained in the sentences retrieved from the TM enhances translation performance. As previously mentioned, the source sentence is denoted as x and the target sentence as y. The semantically matched sentences and each formally matched piece are represented by *smt* and fml_i, respectively. Additionally, we denote the translation results of the non-TM-augmented base model as y^{Base}, the results of the model augmented with only semantically matched sentences as $y^{+Semantic}$, and the translation results of our proposed method as $y^{+Semantic+Formal}$. Here, we perform a comparison to show how, as a sequential model, our approach first encodes global knowledge followed by local knowledge. Therefore, we focus on how formally matched sentences enhance translation at the local level after semantically matched sentences have guided the translation globally. Our results indicate that our method effectively integrates these two levels of knowledge, leading to the enhancement in translation performance. This suggests that our approach combines broad contextual understanding with precise local details, improving overall translation accuracy.

In particular, by building upon $y^{+Semantic}$, which already provides a strong foundation for translation, we further leverage the model's copy mechanism to copy '*über die*' from the second formally matched piece fml_2, to replace the word '**zur**' in $y^{+Semantic}$, guiding the translation of the local phrase, thereby enhancing the overall sentence translation, even to a perfect one. With the guidance from both global and local levels of knowledge, $y^{+Semantic+Formal}$ results in a more accurate and contextually appropriate translation, showcasing the effectiveness of leveraging both global and local knowledge from TM in the translation process.



(a) en→de

(b) de→hsb

Figure 3: The relationship between BLEU scores and semantic similarity intervals for high-resource (en→de) and low-resource (de→hsb) translation tasks. The line charts illustrate the BLEU scores of different models across varying levels of semantic similarity, while the overlaid histograms represent the proportion of sentences falling within each semantic similarity interval. It can be observed that it is more feasible to obtain sentences with higher semantic similarity from the high-resource dataset in comparison to the low-resource dataset.

How to Select Sentences from TM to Maximize Translation Enhancement? Through Figure 3, we can observe two points. First, essentially, for both types of TM-integration configurations that leverage semantically matched sentences, the higher the semantic similarity of the integrated semantically matched sentences, the greater the improvement observed in translation quality. Moreover, when the retrieved semantically matched sentences are of low semantic similarity to the source sentences, it in fact hurts the performance of the model, causing it to under-perform compared to the non-TM baseline.

Second, leveraging both semantically and formally matched sentences to guide translation, compared with just utilizing semantically matched sentences, provides additional benefits at nearly all similarity levels. As Figure 3 shows, the trends in translation quality for '+Semantic' and '+Semantic+Formal' with varying semantic similarity are highly consistent, while our method shows improvement over '+Semantic' across most semantic similarity intervals. Although the improvement is not particularly pronounced, when combined with Table 1 and Figure 3, we can still observe a degree of enhancement. This aligns perfectly with our previous proposition of treating formally matched sentences, combined with a copy module, as a post-editing mechanism. This suggests that while globally augmenting translation effectiveness by selecting sentences with higher semantic similarity, achieving optimal translation performance involves further enhancing translation locally through the use of formally matched sentences.

How Does Our Method Enhance NMT in Low-Resource Scenarios? Combining Table 1 and Figure 3, we can analyze the reasons behind the superior performance of our method on low-resource tasks. First, on these two low-resource datasets, using semantically matched sentences to enhance sentence translation from a global perspective outperforms the non-TM baseline, which may be attributed to: 1) the translation improvement brought by global knowledge, and 2) the increase in the quantity and diversity of training samples through concatenation with semantically matched sentences. Building upon this, introducing local knowledge via formally matched sentences further enhances translation without compromising the existing advantages, leading to better translation quality. This aligns with our goal of leveraging both global and local knowledge to maximize translation improvement, especially in low-resource scenarios.

Moreover, according to Table 1, using only formally matched sentences in TM-integration to enhance translation in low-resource scenarios can actually degrade the performance of the NMT system. This could be due to the limited number of training samples, causing the dual-source transformer to overfit the data. Our approach, on the other hand, avoids this drawback and instead improves performance of the model in low-resource settings through the design of concatenating semantically matched sentences.

5 Conclusion

Recently, many studies have focused on leveraging non-parameterized knowledge to enhance parameterized models. We propose an effective approach to strengthen NMT by exploiting Translation Memory (TM) knowledge. By utilizing semantically similar sentences for global translation guidance and formally matched sentences for local guidance, our method achieves promising results on both high-resource and low-resource datasets, strongly demonstrating the effectiveness of leveraging TM knowledge. Particularly in low-resource scenarios, incorporating TM knowledge can improve translation quality without relying on external datasets beyond the training dataset.

However, our work still has some limitations: since we employ semantic retrieval based on pretrained sentence embeddings, the semantic matching accuracy may be impacted if both languages are low-resource. Secondly, as we use word-alignments to obtain formally matched pieces, our translations are inevitably affected by the alignment accuracy. Addressing these two limitations presents a challenge.

References

Bram Bulte 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.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7307–7318, Online. Association for Computational Linguistics.

Qian Cao, Shaohui Kuang, and Deyi Xiong. 2020. Learning to reuse translations: Guiding neural machine translation with examples. In *ECAI 2020*, pages 1982–1989. IOS Press.

Chi Chen, Maosong Sun, and Yang Liu. 2021. Mask-align: Self-supervised neural word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4781–4791, Online. Association for Computational Linguistics.

Xin Cheng, Shen Gao, Lemao Liu, Dongyan Zhao, and Rui Yan. 2022. Neural machine translation with contrastive translation memories. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180, Online. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

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

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lemao Liu, Hailong Cao, Taro Watanabe, Tiejun Zhao, Mo Yu, and Conghui Zhu. 2012. Locally training the log-linear model for SMT. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 402–411, Jeju Island, Korea. 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*, pages 513–522, Shanghai, China. Association for Computational Lingustics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Masao Utiyama, Graham Neubig, Takashi Onishi, and Eiichiro Sumita. 2011. Searching translation memories for paraphrases. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Mengzhou Xia, Guoping Huang, Lemao Liu, and Shuming Shi. 2019. Graph based translation memory for neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7297–7304.

Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.

Jitao Xu, Josep Crego, and François Yvon. 2023. Integrating translation memories into non-autoregressive machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1326–1338, Dubrovnik, Croatia. Association for Computational Linguistics.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

A Sample Translation Examples

We provide genuine translation examples similar to those illustrated in Table 3, extracted from our experiments conducted across all four utilized datasets, spanning both high-resource and low-resource settings. All sentences are presented in untokenized form for clarity. These authentic instances effectively demonstrate the effectiveness and interpretability of our approach.

x	(5) Directive 92 / 105 / EEC should therefore be amended accordingly .	BLEU
y	(5) Die Richtlinie 92 / 105 / EWG ist daher entsprechend zu ändern .	
smt	**(5) Die Richtlinie 92 / 118 / EWG sollte daher entsprechend geändert werden .**	
fml$_1$	/ EWG ist daher entsprechend *zu ändern* .	
fml$_2$	(5) Die Richtlinie 92 /	
y^{Base}	(5) Die Richtlinie 92 / 105 / EWG sollte daher entsprechend geändert werden .	63.89
$y^{+Semantic}$	**(5) Die Richtlinie 92 / 105 / EWG ist daher entsprechend geändert werden .**	80.65
$y^{+Semantic+Formal}$	**(5) Die Richtlinie 92 / 105 / EWG ist daher entsprechend** *zu ändern* **.**	100.00

Table 4: The translation examples are from experiments done on the English to German (en→de) dataset.

x	Special provisions as regards additional payments	BLEU
y	Disposiciones especiales referentes a los pagos adicionales	
smt	**Disposiciones especiales relativas a las asignaciones**	
fml$_1$	Disposiciones especiales *referentes*	
fml$_2$	para pagos adicionales	
y^{Base}	Disposiciones particulares en materia de pagos adicionales	7.73
$y^{+Semantic}$	**Disposiciones especiales relativas a los pagos adicionales**	48.89
$y^{+Semantic+Formal}$	**Disposiciones especiales** *referentes* **a los pagos adicionales**	100.00

Table 5: The translation examples are from experiments done on the English to Spanish (en→es) dataset.

x	(a) the additional guarantees set out in the model veterinary certificate in Annex III ; and	BLEU
y	a) las garantías adicionales previstas en el modelo de certificado veterinario del anexo III ; y	
smt	**c) el envío cumpla las garantías establecidas en el certificado veterinario elaborado de conformidad con el modelo del anexo V , teniendo en cuenta las notas explicativas del anexo III . &amp; quot ; .**	
fml$_1$	las garantías adicionales *previstas* en el modelo de certificado veterinario del anexo	
fml$_2$	establecidos en el modelo de certificado veterinario del anexo III	
y^{Base}	a) las garantías suplementarias establecidas en el modelo de certificado veterinario que figura en el anexo III , y	39.42
$y^{+Semantic}$	**a) las garantías adicionales establecidas en el modelo de certificado veterinario del anexo III , y**	70.86
$y^{+Semantic+Formal}$	**a) las garantías adicionales** *previstas* **en el modelo de certificado veterinario del anexo III ; y**	100.00

Table 6: The translation examples are from experiments done on the English to Spanish (en→es) dataset.

x	Wir bewegen uns nach vorn, nach hinten, nach rechts und nach links!	BLEU
y	Schylamy se dopředka, naslědk, napšawo a nalewo!	
smt	**Musalej smej se rozsuźiś, lec napšawo, nalewo abo narowno dalej ganjamej.**	
fml$_1$	*dopředka,* naslědk,	
y^{Base}	Wobgranicujomy se dopředka hyś, naslědk a nalewo!	7.73
$y^{+Semantic}$	**Smy se wupórali, naslědk naslědk, napšawo a nalewo!**	36.56
$y^{+Semantic+Formal}$	**Wobejźujomy se** *dopředka,* **naslědk, napšawo a nalewo!**	80.91

Table 7: The translation examples are from experiments done on the German to Lower-Sorbian (de→dsb) dataset.

x	1 . The Committee shall consist of two representatives from each Member State .	BLEU
y	(1) Der Ausschuß besteht aus je zwei Vertretern jedes Mitgliedstaats .	
smt	**(1) Die Agentur hat einen Verwaltungsrat , der sich aus je einem Vertreter der Mitgliedstaaten und zwei Vertretern der Kommission zusammensetzt .**	
fml_1	(1) Der Ausschuß *besteht* aus	
fml_2	Ausschuss *besteht* aus	
y^{Base}	(1) Der Ausschuß setzt sich aus zwei Vertretern je Mitgliedstaat zusammen .	33.43
$y^{+Semantic}$	**(1) Der Ausschuß setzt sich aus je zwei Vertretern jedes Mitgliedstaats zusammen .**	58.28
$y^{+Semantic+Formal}$	**(1) Der Ausschuß** *besteht* **aus je zwei Vertretern jedes Mitgliedstaats .**	100.00

Table 8: The translation examples are from experiments done on the English to German (en→de) dataset.

x	So beschrieb der Maler Jan Bück sein ambivalentes Verhältnis zur industriellen Wende in den Lausitzen.	BLEU
y	Tak wopisowaše moler Jan Buk swój ambiwalentny poćah k industrielnemu přewrótej we Łužicomaj.	
smt	**»Grilowane kołbaski zaso wulkotnje słodźa!«, praji Lina zahorjena.**	
fml_1	we *Łužicomaj.*	
y^{Base}	Tak wopisowaše moler Jan Buk jeho ambiwalentny poměr k industrialnym přewróće we Łužicach.	32.52
$y^{+Semantic}$	**Tak wopisowaše moler Jan Buk swoju ambiwalentny poměr k industrijowemu přewrótej we Łužicach.**	35.42
$y^{+Semantic+Formal}$	**Tak wopisowaše moler Jan Buk swój ambiwalentny poměr k industrielnemu přewrótej we** *Łužicomaj.*	76.12

Table 9: The translation examples are from experiments done on the German to Upper-Sorbian (de→hsb) dataset.

x	Die Erzieherin beobachtet die gegenseitige Hilfe der Kinder, wenn eines von ihnen nicht das Sorbische verstand.	BLEU
y	Kubłarka wobkedźbuje wzajomnu pomoc dźěći, hdyž njeje jedne z nich serbšćinu rozumiło.	
smt	**Kubłarka reaguje na situacije, w kotrychž trjeba so zaživjace dźěćo přidatnu podpěru (n.př. při nawjazanju kontakta k druhim dźěćom).**	
fml_1	hdyž *njeje* jedne z	
y^{Base}	Kubłarka wobkedźbuje mjezsobnu pomoc dźěći, hdyž njeje jedna z nich serbski njerozum.	28.65
$y^{+Semantic}$	**Kubłarka wobkedźbuje mjezsobnu pomoc dźěći, hdyž njebě jedne z nich serbšćinu rozumiło.**	46.60
$y^{+Semantic+Formal}$	**Kubłarka wobkedźbuje mjezsobnu pomoc dźěći, hdyž** *njeje* **jedne z nich serbšćinu rozumiło.**	76.92

Table 10: The translation examples are from experiments done on the German to Upper-Sorbian (de→hsb) dataset.

x	Es fehlen noch Dachboden, Keller, Garage, Hof, Garten.	BLEU
y	Feluju hyšći najšpa, piwnica, garaža, dwór, zagroda.	
smt	**Buźćo wjasołe w naźeji, sćerpne w tešnosći, hobstawne w módlenju.**	
fml_1	*Feluju*	
y^{Base}	Feluju hyšći najšpy, piwnica, garaž, gumno.	8.09
$y^{+Semantic}$	**Póbrachujo hyšći najšpy, piwnica, garaža, dwór, zagroda.**	43.47
$y^{+Semantic+Formal}$	*Feluju* **hyšći najšpy, piwnica, garaža, dwór, zagroda.**	48.89

Table 11: The translation examples are from experiments done on the German to Lower-Sorbian (de→dsb) dataset.