Enhancing Neural Machine Translation Through Target Language Data: A kNN-LM Approach for Domain Adaptation

Abudurexiti Reheman [1], Hongyu Liu [1], Junhao Ruan [1], Abudukeyumu Abudula [1], Yingfeng Luo [1], Tong Xiao [1,2], Jingbo Zhu [1,2] *,

[1] School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2] NiuTrans Research, Shenyang, China
rexiti_neu@outlook.com
{xiaotong,zhujingbo}@mail.neu.edu.cn

Abstract

Neural machine translation (NMT) has advanced significantly, yet challenges remain in adapting to new domains . In scenarios where bilingual data is limited, this issue is further exacerbated. To address this, we propose kNN-LM-NMT, a method that leverages semantically similar target language sentences in the kNN framework. Our approach generates a probability distribution over these sentences during decoding, and this distribution is then interpolated with the NMT model's distribution. Additionally, we introduce an n-gram-based approach to focus on similar fragments, enabling the model to avoid the noise introduced by the non-similar parts. To enhance accuracy, we further incorporate cross-lingual retrieval similarity to refine the kNN probability distribution. Extensive experiments on multi-domain datasets demonstrate significant performance improvements in both high-resource and low-resource scenarios. Our approach effectively extracts translation knowledge from limited target domain data, and well benefits from large-scale monolingual data for robust context representation.

1 Introduction

With the introduction of deep learning techniques, especially the revolutionary models like Transformer, neural machine translation (NMT) has made significant progress in recent years (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). Despite these advancements, NMT often suffers from translating in new domains, which is called domain adaptation (Koehn and Knowles, 2017; Isabelle et al., 2017; Lee et al., 2018; Farajian et al., 2017). The primary reason for this issue is the domain mismatch between the training and test datasets. To address this, researchers have developed various strategies to enhance NMT model's domain adaptation capabilities through the incorporation of external domain knowledge or similar examples (Cao and Xiong, 2018; Bulté and Tezcan, 2019; Xu et al., 2020; Meng et al., 2022; Reheman et al., 2023). A particularly promising approach integrates non-parametric methods with NMT, where the non-parametric methods rely on an external memory of additional translation examples (Khandelwal et al., 2021). However, the heavy reliance on high-quality ground-truth bilingual data limits the applicability of these methods in low-resource scenarios.

On the other hand, with the advantages of large data scale, broad domain coverage, and ease of access, monolingual data is widely employed to enhance NMT capabilities (Pang et al., 2024). Back-translation (Sennrich et al., 2016; Caswell et al., 2019; Marie et al., 2020) enriches training data by translating target language data back into the source language using a reverse translation model. Shallow fusion or deep fusion leverages language models that are trained on the target language data to constrain the translation process of NMT (Gülçehre et al., 2015, 2017; Sriram et al., 2018). Although these methods efficiently leverage monolingual data to enhance parametric models, they inevitably increase the training cost.

Recent non-parametric methods, UDA-kNN (Zheng et al., 2021b) and Pseudo-kNN-MT (Reheman et al., 2024), leverage target language data in the kNN-MT framework. The former introduces adapter modules that require NMT model retraining, while the latter pairs similar target sentences with the source input. However, this approach is vulnerable to the noise incorporated by the non-similar parts of the target retrievals and is constrained by the representation capability of the NMT model.

As a non-parametric method, k-nearest neighbor language model (kNN-LM) (Khandelwal et al., 2020) has achieved notable success in language modeling tasks. In the machine translation task,

[*] Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10053–10065
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

since the decoder of a transformer NMT model learns a causal language model implicitly, if the target language instances that are semantically similar to the source input sentence can be provided, a non-parametric LM can be utilized to incorporate the relevant translation information.

Building on this idea, we propose our method, kNN-LM-NMT. Specifically, given some semantically similar target sentences, kNN-LM-NMT generates a probability distribution over these sentences in the kNN-LM framework. This probability distribution is subsequently interpolated with the NMT model's distribution. Additionally, to effectively leverage the similar fragments from the target language sentences, we introduce an n-gram-based approach, which focuses on similar fragments. This enables the model to filter out non-similar parts of the target sentence. Furthermore, to differentiate the contributions of sentences with varying similarity levels, we incorporate cross-lingual retrieval similarity to refine next token probability of the kNN.

We validate our approach on multi-domain datasets for both high-resource and low-resource machine translation tasks. The results demonstrate that our method significantly enhances the translation performance of NMT on both scenarios.

In real-world application, our method can effectively extract useful translation knowledge from a small amount of similar target domain data. At the same time, large-scale target data enables training robust language models for context representation.

2 Preliminaries

For better understanding, we provide an overview of the fundamental concepts underlying kNN-LM and LaBSE.

2.1 kNN-LM

kNN-LM (Khandelwal et al., 2020) is a semi-parametric language modeling approach, which combines the traditional language model with nearest neighbor search techniques, allowing it to leverage large corpora of text data more effectively. Applying kNN-LM involves two main steps: datastore creation and inference using the datastore.

Datastore Creation. Let $\mathcal{D}$ denote a collection of key-value pairs, where key is a high-dimensional representation of a context from a pretrained language model (LM), and value is the corresponding ground-truth next token. Let $\mathcal{Y} = \{Y_1, Y_2, ..., Y_n\}$ be a set of sentences and let $f(\cdot)$ be the mapping function that transfers the context into the vector representation using an LM. For all sentences in $\mathcal{Y}$, the key-value datastore is created as:

$$\mathcal{D} = \{(f(y_{1:t-1}), y_t), \forall y_t \in Y \mid Y \in \mathcal{Y}\}. \quad (1)$$

Here, the datastore size is equal to the total number of tokens in $\mathcal{Y}$.

Inference. During the inference phase, the representation of the previously generated tokens, $y_{1:t-1}$, at each time-step is taken as a query, denoted as $q = f(y_{1:t-1})$, to retrieve k-nearest neighbors $\mathcal{N}$ from the datastore $\mathcal{D}$. This retrieval process employs vector distance measuring metrics, such as L^2 distance or cosine similarity, to ascertain the proximity of the context representation to the stored instances. Subsequently, the kNN distribution, $p_{k\mathrm{NN}}$, over the vocabulary is then obtained by normalizing the negative distances with temperature, and aggregating the probability of same tokens afterwards. The kNN distribution is formulated as:

$$p_{k\mathrm{NN}}(y_t|\hat{y}_{1:t-1}) \propto \sum_{(k_j,v_j)\in\mathcal{N}} \mathbb{1}_{y_j=v_j} \exp(-d(q,k_j)/T), \quad (2)$$

where $d(\cdot,\cdot)$ represents the distance function that calculates the distances between the query vector and the keys from the datastore and T is the temperature.

In the end, the final generation probability is calculated by linearly interpolating the kNN distribution with the LM distribution, as:

$$p(y_t \mid \hat{y}_{1:t-1}) = \lambda p_{k\mathrm{NN}}(y_t \mid \hat{y}_{1:t-1}) + (1-\lambda)p_{\mathrm{LM}}(y_t \mid \hat{y}_{1:t-1}), \quad (3)$$

where λ is the interpolation hyperparameter.

2.2 LaBSE

LaBSE (Language-agnostic BERT Sentence Embedding) (Feng et al., 2022) is a multilingual sentence embedding model designed to generate language-agnostic embeddings that capture semantic meaning across diverse languages.

In cross-lingual retrieval tasks, LaBSE encodes sentences from different languages into a unified representation space. Each sentence is first processed through the model, resulting in a dense vector representation. These representations can

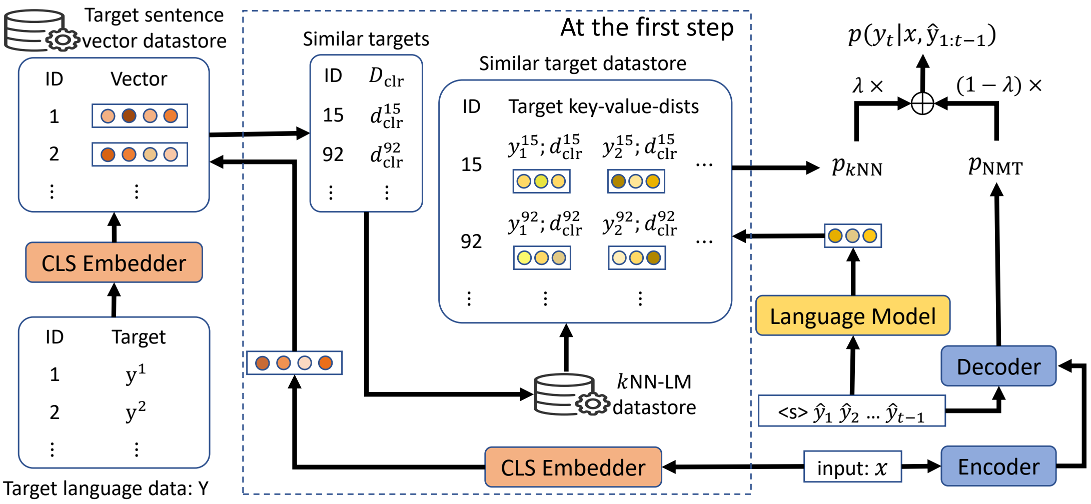

Figure 1: The illustration of the proposed kNN-LM-NMT method. CLS Embedder refers to the cross-lingual sentence embedding model. The Encoder and Decoder at the right-bottom refer to the encoder and decoder of a standard transformer NMT model.

be compared using similarity functions. By computing the similarity of sentences in different languages, LaBSE effectively retrieves semantically similar sentences across languages. In this paper, we use it to retrieve similar target sentences for the source input sentence.

3 Methodology

Given a limited set of similar sentences in the target language, our objective is to extract valuable knowledge from them to enhance the NMT by integrating this knowledge into the translation process. As a robust non-parametric language modeling method, kNN-LM effectively incorporates the probability of the next token into the target generation process through approximate similarity search. Moreover, since these similar sentences exhibit a certain degree of similarity to the input sentence, it is possible to leverage LMs to integrate the relevant translation knowledge from these sentences without referencing the source sentences.

3.1 Retrieving similar target Sentences

Given the source input sentence X and the target language dataset $\mathcal{Y} = \{Y_1, Y_2, ..., Y_n\}$, we utilize the cross-lingual embedding model, LaBSE, denoted as $e(.)$, to derive their dense vector representations as follows:

$$h_X = e(X), \quad (4)$$

$$\mathcal{H}_\mathcal{Y} = \{e(Y) \mid Y \in \mathcal{Y})\}. \quad (5)$$

Subsequently, we calculate the Euclidean Distances (also referred to as L^2 distances) between the source sentence embedding h_X and each target sentence embedding in $\mathcal{H}_\mathcal{Y}$. Finally, we rank the target sentences based on their distances and select the top-k nearest ones as the final retrieved targets.

3.2 Integrating kNN-LM into NMT

Similar to other kNN-based methods, we first construct a kNN-LM datastore $\mathcal{D}$ using the entire target language dataset. This process is identical to the datastore construction described in §2.1. Both this datastore and the target sentence vector datastore (used for retrieving similar target language sentences in §3.1) are built offline. During the decoding process, the kNN probability distribution for the next token is generated from $\mathcal{D}$ and then interpolated with the probability distribution from the NMT model. The overall process of this method is illustrated in Figure 1.

At the first step of decoding, we extract a sentence-specific datastore $\mathcal{D}_{\text{sim}}$ for the source sentence. Specifically, for an input sentence X, we use the method described in §3.1 to obtain a set of similar target sentences, along with their cross-lingual distances. Then, we extract the corresponding keys and values from the kNN datastore based on the similar target sentence IDs and construct triples in the form of $<key, value, dist>$. Here, "$dist$" is the sentence-level cross-lingual retrieval distance of a target sentence to which the token belongs. These triples are then stored to construct $\mathcal{D}_{\text{sim}}$.

At each time-step of the decoding process, we input the previously generated target tokens $y_{1:t-1}$ into the LM, the same model which is used to

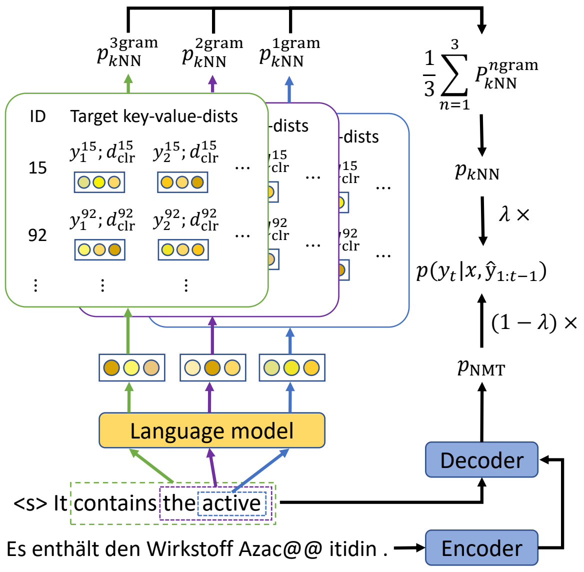

Figure 2: An example of incorporating n-gram based kNN-LM with NMT.

build the kNN datastore, to obtain query vector $q = f(y_{1:t-1})$. Next, we use q to retrieve the k-nearest neighbors from $\mathcal{D}_{\text{sim}}$ and construct the kNN probability distribution. Unlike the standard kNN probability generation method, we incorporate cross-lingual retrieval distances to refine the kNN probability. This method is described in detail in §3.4. Finally, the kNN distribution is interpolated with the NMT distribution as:

$$p(y_t \mid x, \hat{y}_{1:t-1}) = \lambda p_{k\text{NN}}(y_t \mid \hat{y}_{1:t-1}) + (1-\lambda) p_{\text{NMT}}(y_t \mid x, \hat{y}_{1:t-1}). \quad (6)$$

3.3 Integrating n-gram-based kNN-LM with NMT

Since we utilize similar target language sentences, the non-similar parts may introduce noise. Additionally, due to the auto-regressive nature of causal LMs, the representation at a time-step inherently contains all the information of its prefix. To mitigate the negative impact of noise in the representation, we propose a local representation method based on n-gram sequences. An example of this method is illustrated in Figure 2.

First, we construct n-gram datastores using the target language dataset. Specifically, each sentence in $\mathcal{Y}$, we input every n-gram segment into the LM to obtain its hidden representation as the key, with the corresponding next token as the value. Hence, the n-gram kNN datastore is constructed as:

$$\mathcal{D}^{ngram} = \{(f(y_{t-n:t-1}), y_t), \forall y_t \in Y \mid Y \in \mathcal{Y}\}. \quad (7)$$

where $f(\cdot)$ is the text-to-vector mapping function using an LM, Y is the target sentence, and $y_{t-n:t-1}$ is the n-gram segment at time-step t. Notably, we take the hidden representation h_t as the key for $y_{t-n:t-1}$, reflecting the nature of casual LMs.

At the first time-step of decoding, we retrieve the similar target sentences and extract a sentence-specific n-gram datastore $\mathcal{D}_{\text{sim}}^{ngram}$, using the same method described in §3.2. At each time-step of decoding, we extract n-gram segment $y_{t-n:t-1}$ from the previously generated target sequence $y_{1:t-1}$ and input $y_{t-n:t-1}$ into the LM to obtain its representation, which is then taken as query to construct the n-gram kNN distribution from $\mathcal{D}_{\text{sim}}^{ngram}$. If the length of $y_{1:t-1}$ is insufficient to extract n-gram segment, the corresponding n-gram kNN distribution is not constructed for that time-step.

Furthermore, we construct multiple n-gram kNN distributions, and take the average of these distributions to obtain the final kNN distribution, which is then interpolated with NMT distribution as:

$$p(y_t \mid x, \hat{y}_{1:t-1}) = \lambda \frac{1}{n} \sum_{1}^{n} p_{k\text{NN}}^{ngram} + (1-\lambda) p_{\text{NMT}}(y_t \mid x, \hat{y}_{1:t-1}). \quad (8)$$

3.4 Integrating Cross-lingual Similarity

The generation of kNN probabilities relies on the retrieved similar target sentences, independent of the source sentence information. Nevertheless, the cross-lingual retrieval distance reflects the degree of translational correspondence between them, providing essential information for kNN probability. Therefore, we propose two distinct approaches to integrate cross-lingual retrieval distances into the kNN probability generation framework.

Updating kNN Distance (UD). For an input sentence X, we have the similar target sentence datastore $\mathcal{D}_{\text{sim}}$ (we denote this in both kNN and n-gram kNN occasions). At time-step t, we feed the previously generated target prefix $\hat{y}_{1:t-1}$ or the n-gram segment $\hat{y}_{t-n:t-1}$ into the LM to obtain its representation h_q. Using h_q as the query ($q = h_q$), we retrieve k-nearest neighbors from $\mathcal{D}_{\text{sim}}$, obtaining kNN distances d^{knn} and cross-lingual retrieval distances d^{clr} of the top-k neighbors. The kNN probability is constructed based on the multiplication of d^{clr} and $d^{k\text{nn}}$ as follows:

$$p_{k\text{NN}}(y_t \mid x, \hat{y}_{1:t-1}) \propto \sum_{(k_j, v_j, d_j^{\text{clr}}) \in \mathcal{D}_{\text{sim}}} \mathbb{1}_{y_j=v_j} \exp(-d_j^{\text{clr}} d^{k\text{nn}}(q, k_j)/T). \tag{9}$$

Updating kNN Probability (UP). Unlike UD, this approach transforms the cross-lingual distances into similarity scores and updates kNN probability accordingly. Specifically, after retrieving k-nearest neighbors from the datastore, we compute their similarity scores s^{clr} by applying an exponential function to the negative cross-lingual distance:

$$s_j^{\text{clr}} = \exp(-d_j^{\text{clr}}), \quad d_j^{\text{clr}} \in \mathcal{D}_{\text{sim}}. \tag{10}$$

Following the construction of the kNN probability, we multiply each neighbor's probability by its corresponding similarity score and then aggregate probabilities of identical tokens. The updated kNN probability is thus formulated as:

$$p_{k\text{NN}}(y_t \mid x, \hat{y}_{1:t-1}) \propto \sum_{(k_j, v_j) \in \mathcal{D}_{\text{sim}}} \mathbb{1}_{y_t=v_j} \exp(-d^{k\text{nn}}(q, k_j)/T) s_j^{\text{clr}}. \tag{11}$$

4 Experiments

We evaluate the effectiveness of our approach on a publicly available multi-domain dataset in high-resource and low-resource translation settings. We also explore the usability of LLM in the kNN component of our method.

Models, Datasets and Evaluation Metrics. We employ Facebook's champion model on WMT19 German-English (De-En) news translation task[1] (Ng et al., 2019). For the kNN component, we use their WMT19 English LM [2], trained on the monolingual Newscrawl dataset of that year. We test our method on the test sets of IT, Koran, Law and Medical domains of the Multi-domain dataset which is originally introduced in Koehn and Knowles (2017) and resplit by Aharoni and Goldberg (2020). We take the target training data of each domain as the monolingual target language dataset. The data statistics are given in Table 1. The translation quality is evaluated using SacreBLEU (Post, 2018) and COMET-22[3] (Rei et al., 2022).

Split	WMT19	IT	Koran	Law	Medical
Train	33M	223k	17k	467k	248k
Valid	6002	2000	2000	2000	2000
Test	2000	2000	2000	2000	2000

Table 1: Statistics of datasets.

Experimental Settings. We utilize the cross-lingual embedding model LaBSE (Feng et al., 2022) to convert the sentences from both source and target language datasets into sentence embeddings. We employ dense vector search library FAISS (Johnson et al., 2021) for both the cross-lingual retrieval and the kNN search from the datastore, using its built-in L^2 distance function for vector distance measuring. In the cross-lingual retrieval, we retrieve the top-32 similar sentences from the target dataset for each source input sentence. In the kNN search, we retrieve $k = 8$ neighbors from the kNN datastore. For the key of the kNN datastore, following the recommendations from Khandelwal et al. (2020); Xu et al. (2023), we extract the input for the "FFN" in the last layer of the LM decoder. Regarding the kNN temperature, we search it from $\{0.5, 1, 5, 10, 20, 50\}$. For interpolation, we search λ from $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. For n-gram kNN method, we set the n-gram from "1-gram" to "5-gram". For decoding, we set the *beam size* to 5 and *length penalty* to 1.0.

Baselines. We compare our method with existing ones that leverage target language data in NMT, especially with non-parametric kNN methods.

- NMT: The general-domain NMT model.
- Shallow Fusion (Gülçehre et al., 2017): Interpolates the probability distribution of a trained LM with NMT distribution during the inference. For fair comparison, we also finetune the LM on the training data of multi-domain dataset.
- BT (Sennrich et al., 2016): Standard back-translation approach that augments the training data by translating the target data back to the source language.
- Pseudo-kNN-MT (Reheman et al., 2024): A kNN-based method that creates pseudo-bilingual data by pairing similar target sentences with source sentences for kNN-MT.
- UDA-kNN (Zheng et al., 2021b): Incorporates target language data into kNN-MT by training a specialized adapter in the transformer NMT model.

[1] https://dl.fbaipublicfiles.com/fairseq/models/wmt19.de-en.ffn8192.tar.gz

[2] https://dl.fbaipublicfiles.com/fairseq/models/lm/wmt19.en.tar.gz

[3] https://github.com/Unbabel/COMET

Methods	**IT**		**Koran**		**Law**		**Medical**		**Average**	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
NMT	38.43	82.46	17.07	72.57	45.99	85.38	41.97	83.16	35.86	80.89
Shallow Fusion	36.44	81.05	17.32	72.34	46.15	84.87	40.98	81.73	35.22	79.99
Shallow Fusion(ftLM)	37.57	81.92	17.62	72.71	47.87	85.47	42.15	82.37	36.30	80.62
UDA-kNN	40.67	82.71	18.98	**73.40**	51.17	85.90	45.95	83.79	39.19	81.45
Pseudo-kNN-MT	40.63	82.51	18.46	72.24	53.03	84.68	49.36	82.43	40.37	80.46
BT-kNN	41.58	82.96	20.35	73.00	54.43	85.96	49.47	83.93	41.46	81.46
kNN-UD	41.61	83.06	20.51	72.93	54.34	85.72	49.73	83.82	41.55	81.38
kNN-UP	**42.47**	**83.10**	20.47	72.91	56.27	85.53	**52.57**	83.57	**42.95**	81.27
n-gram-kNN-UD	42.18	83.06	**20.63**	73.28	55.61	86.07	50.07	83.91	42.12	81.58
n-gram-kNN-UP	42.45	83.01	20.39	73.26	**56.91**	**86.19**	50.83	**83.97**	42.65	**81.61**

Table 2: SacreBLEU and COMET scores of WMT19 De-En model on the multi-domain test sets. The best results are highlighted in bold, and the second-best results are underlined for clarity.

- BT-kNN: Generates synthetic bilingual data by translating the target dataset back to the source language, which is then used for kNN-MT datastore construction.

4.1 Main Experiment

We conduct this experiment on the multi-domain test set using the WMT19 De-En model. Regarding the reverse NMT model for BT-kNN, we use Facebook's WMT19 En-De model [4], which was trained using the same dataset as the forward model.

Results. The experimental results presented in Table 2 reveal observations below: first, our proposed method demonstrates superior performance compared to other baseline models, dominantly achieving the best and second best results on both BLEU and COMET metrics across most domains, except the COMET score for the Koran domain. Notably, our method even outperforms the competitive BT-kNN baseline, which utilizes bilingual data generated through back-translation. Secondly, the results indicate that while shallow fusion offers limited enhancement to NMT performance after learning the target domain knowledge by fine-tuning, other non-parametric methods exhibit substantial improvements in translation quality. This underscores the effectiveness of non-parametric approaches in rapidly adapting to new domain. Third, our approach consistently benefits from the n-gram enhancement, which suggests that the n-gram method effectively mitigates the noise introduced by non-similar parts of the target sentence to some extent. Lastly, comparing the performance of the cross-lingual retrieval distance incorporation methods, UP performs better than UD. This is mainly because the nature of the "softmax" function that generates kNN probability using the distance. When the kNN distance is either too large or too small, the UD approach, which multiplies the kNN distance by the cross-lingual retrieval distance, has a limited impact on the resulting probability distribution. In contrast, the UP approach directly updates the kNN probability distribution, enabling more effective utilization of the cross-lingual similarity.

4.2 Low Resource Machine Translation

Target language monolingual data has been demonstrated to be beneficial in low-resource machine translation scenarios. To evaluate the performance of our method under such conditions, we conduct experiments on the multi-domain dataset under low-resource settings. We employ the German-to-English low-resource machine translation model provided by Pseudo-kNN-MT (Reheman et al., 2024). This model was trained on 500k bilingual data obtained by uniformly sampling from the cleaned WMT21 De-En news translation task dataset. For the training data of the target LM, we mix the target data of this 500k dataset with the target training data from each domain of the multi-domain dataset. Subsequently, we train a 12-layer decoder-only GPT model (Radford et al., 2019) with an embedding dimension of 768. To ensure consistency in the target vocabulary, we utilize the target language vocabulary of the NMT model for tokenization. In addition to the baselines above, we also compare our method with the standard BT in this experiment.

Results. The experimental results in Table 3 demonstrate that our method achieves the best performance across three domains. In terms of average

[4] https://dl.fbaipublicfiles.com/fairseq/models/wmt19.en-de.ffn8192.tar.gz

Methods	IT	Koran	Law	Medical	Average
NMT	28.69	10.68	28.42	30.17	24.49
BT	31.72	12.73	**41.32**	38.08	30.96
UDA-kNN	31.74	14.44	32.52	35.68	28.60
Pseudo-kNN	30.61	13.97	36.75	38.06	29.85
BT-kNN	32.35	13.28	36.65	37.41	29.92
kNN-UD	30.80	14.72	38.41	38.89	30.71
kNN-UP	31.76	14.73	41.07	40.95	32.13
n-gram-UD	31.68	**15.02**	40.11	40.13	31.74
n-gram-UP	**32.94**	14.66	41.10	**41.52**	**32.56**

Table 3: SacreBLEU scores of WMT21 De-En sample-500k model on the multi-domain test sets. COMET scores are given in Appendix A.

Methods	IT	Koran	Law	Medical	Average
NMT	35.90	16.95	45.04	39.77	34.42
kNN-UD	38.03	19.58	51.62	45.97	38.80
kNN-UP	38.92	19.21	52.94	46.61	39.42
n-gram-UD	40.22	**20.88**	53.23	47.88	40.55
n-gram-UP	**41.31**	20.68	**54.91**	**49.13**	**41.51**

Table 4: SacreBLEU scores of WMT19 De-En Llama3.1-dict model on the multi-domain test sets. COMET scores are given in Appendix A

BLEU scores, our four proposed methods consistently outperform all other baselines except the BT. While BT achieves optimal performance in the Law domain, our method performs comparably in this domain. From the perspective of data scale, BT, as a parametric method, benefits more from a larger monolingual data. Therefore, it achieves the best performance in the law domain, which contains 467k sentences. Conversely, our method, as a non-parametric approach, focuses more on the similarity between the target and source sentence. Consequently, even in the Koran domain with only 17k sentences, our method significantly outperforms BT.

4.3 Large Language Models in kNN Component

Large language models (LLMs) demonstrate remarkable success across various NLP tasks (Brown et al., 2020; Alves et al., 2024; Luo et al., 2025; Muennighoff et al., 2025; Zheng et al., 2025). Trained on massive corpora, LLMs exhibit strong representational capabilities. Given the critical role of context representation in our method for searching nearest neighbors, we investigate the application of LLMs within the kNN component. We employ Llama 3.1-8B [5]. It is essential to keep vocabulary consistency between the target language of NMT model and the LLM to integrate their probability distributions. Accordingly, we train an NMT model on the WMT19 De-En dataset using the Llama dictionary. The model architecture of this NMT model is identical to that in the main experiment. To determine the optimal *key*, we conduct preliminary experiments using both the "FFN-input" and "Hidden-state" from the last layer of the Llama model (see Appendix B for detailed results). The "last-FFN-input" yields better performance and is thus selected as the *key*.

As presented in Table 4, the experimental results reveal that while the baseline NMT model exhibits a performance decline compared to the one from main experiment - attributable to the Llama vocabulary implementation - our method substantially enhances translation performance over the NMT baseline. Although the results closely match those of the main experiments, they do not fully align, mainly due to the relative weakness of the NMT model used, whose generated distributions are less accurate for interpolation.

5 Analysis

5.1 Impact of Cross-lingual Similarity Integration

To evaluate the contribution of cross-lingual retrieval similarity to translation performance, we conduct an ablation study by removing the integration mechanism of cross-lingual retrieval distance, denoted as d^{clr}. As shown in Table 5, removing d^{clr} leads to performance degradation in both methods. Specifically, the average SacreBLEU scores of kNN-UD and kNN-UP drop significantly from 41.55 and 42.95 to 38.81, highlighting the strong reliance on d^{clr}. In contrast, n-gram-UD remains nearly unchanged, while n-gram-UP experiences a slight decline.

We attribute these results to the inherent noise-handling capabilities of the methods. The n-gram method constructs probabilities from similar n-gram segments, making it easier to exclude noise introduced by non-similar parts. In contrast, the kNN method relies on the entire prefix, making it more sensitive to noise. The d^{clr} helps determine the level of noise in the sentence for kNN, explaining its stronger reliance on this mechanism.

[5] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B

Methods	IT	Koran	Law	Medical	Avg
NMT	38.43	17.07	45.99	41.97	35.86
kNN-UD	41.61	20.51	54.34	49.73	41.55
kNN-UP	42.47	20.47	56.27	52.57	42.95
No d^{clr}	39.93	18.99	50.06	46.25	38.81
n-gram-UD	42.18	20.63	55.61	50.07	42.12
n-gram-UP	42.47	20.47	56.27	52.57	42.95
No d^{clr}	42.18	20.57	55.64	50.40	42.20

Table 5: Effect of removing cross-lingual retrieval distance on SacreBLEU scores. "No d^{clr}" indicates the removal of cross-lingual retrieval similarity.

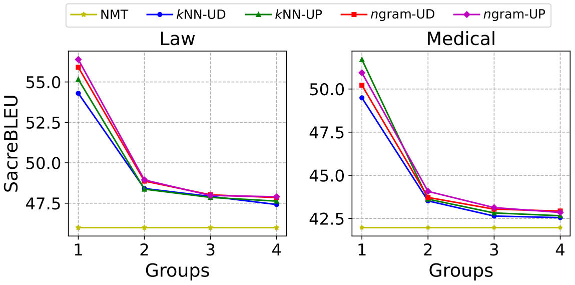

Figure 3: SacreBLEU scores of WMT19 De-En model on Law and Medical domains across similarity-based groups.

5.2 Impact of Target Language Similarity

Our method benefits from the similar content in the target sentence. To investigate the target similarity impact, we sort retrieved target sentences by similarity and divide them into four groups, with group 1 having the highest similarity and groups 2-4 showing progressively lower. We conduct experiments using the WMT19 De-En model on the Law and Medical domains.

The results in Figure 3 show performance declines sharply with decreasing similarity, followed by a more gradual decline. This highlights the critical importance of target sentence similarity, as higher similarity provides more informative cues for translation.

5.3 Impact of Cross-lingual Model

The retrieval of semantically similar sentences from the target data with high recall rates is relevant to the representational capabilities of cross-lingual models. To this end, we evaluate our method using several cross-lingual models, including E5 (Wang et al., 2022b), LASER2 (Heffernan et al., 2022), and MuSR (Gao et al., 2023), under the main experiment settings. Since the multi-domain dataset is bilingual, we simulate a scenario with near-perfect target similarity recall: we first retrieve similar source sentences using edit distance, as:

$$d(x_i, x_j) = ED(x_i, x_j)/\max(|x_i|, |x_j|), \quad (12)$$

where $ED(\cdot,\cdot)$ is Edit Distance function and $|x|$ is the length of x, then extract the corresponding target language sentences as the final target retrievals.

Models	Methods	IT	Koran	Law	Medical	Avg
-	NMT	38.43	17.07	45.99	41.97	35.86
E5	kNN-UD	41.87	20.47	54.86	49.49	41.67
	kNN-UP	42.13	20.60	56.09	51.89	42.68
	n-gram-UD	42.12	20.64	56.10	49.69	42.14
	n-gram-UP	42.11	<u>20.72</u>	56.84	50.03	42.43
Laser2	kNN-UD	41.27	19.88	54.16	46.93	40.56
	kNN-UP	41.60	20.27	56.00	49.53	41.85
	n-gram-UD	41.41	20.58	55.41	48.84	41.56
	n-gram-UP	41.43	20.52	56.18	50.01	42.04
Musr	kNN-UD	41.37	20.57	54.44	49.59	41.49
	kNN-UP	42.17	20.66	56.73	51.67	42.81
	n-gram-UD	42.20	**20.78**	55.87	49.96	42.20
	n-gram-UP	42.27	20.61	57.03	51.08	42.75
SrcED	kNN-UD	41.11	20.32	54.61	48.34	41.10
	kNN-UP	<u>43.24</u>	19.86	**58.11**	**52.42**	**43.41**
	n-gram-UD	42.43	20.38	55.88	49.61	42.08
	n-gram-UP	**43.33**	19.91	<u>57.92</u>	<u>51.91</u>	<u>43.27</u>

Table 6: SacreBLEU score comparison of applying different cross-lingual retrieval models. The best and second best results are highlighted in bold and underline, respectively.

The distributions of cross-lingual retrieval distances vary significantly across these models. For fair comparison, we scale their distance distributions to align with the scale of LaBSE in the main experiment. The experimental results are shown in Table 6.

The results show that, compared with the results obtained using LaBSE for cross-lingual retrieval in the main experiment, the differences are marginal except Laser2, indicating that their cross-lingual recall rates have negligible impact on this task. The source edit distance based retrieval approach, by enabling more precise target retrievals, yields the most significant improvement in our method, especially for the UP approach. Additionally, kNN-UP achieves the highest average SacreBLEU score, better than its n-gram based counterpart. We believe this is because when target sentence similarity is higher, the performance degradation caused by the noise in the non-similar parts of the sentences is reduced.

Methods	IT	Koran	Law	Medical
NMT	177.44	182.04	191.93	179.63
# Target Sents	223k	17k	467k	248k
kNN Dstore Size	3.16M	0.52M	19.07M	6.90M
UDA-kNN	29.14	101.61	12.88*	17.06
BT-kNN	29.75	102.83	13.42*	17.47
kNN(UD/UP)	30.04	31.92	27.00	28.01
n-gram(UD/UP)	29.13	28.87	30.33	29.10

Table 7: SacreBLEU score comparison of applying different cross-lingual retrieval models. "$*$" represents using half of the datastore due to the GPU memory limitation.

5.4 Efficiency

We evaluate the translation speed of our method against other kNN-based methods under the main experiment settings. We use a batch size of 1 and a FlatL2 FAISS index. Results with *tokens/second* are given in Table 7.

Experimental results demonstrate that the performance of UDA-kNN and BT-kNN is highly sensitive to the kNN datastore size, with decoding speed degrading significantly with larger datastores. In our method, the target sentence vector datastore and the kNN-LM datastore are constructed offline. Therefore, apart from the decoding time of the NMT model, the time overhead of our method mainly occurs as below. (1) T_{clr}: cross-lingual retrieval at the first step of the decoding, perform once per sentence; (2) T_{lm}: computing the query vector via a forward-pass of the LM on the current target sequence or the n-grams; (3)$T_{k\mathrm{nn}}$: searching for kNNs from the similar target datastore $\mathcal{D}_{\mathrm{sim}}$. Although our method incurs additional T_{clr} and T_{lm}, the orders of magnitude smaller $\mathcal{D}_{\mathrm{sim}}$ (e.g., a datastore of 32 target similar sentences with an average length of 30 tokens, totaling 960 vectors) enables a smaller $T_{k\mathrm{nn}}$, avoiding scalability issues.

6 Related Work

Utilizing Monolingual Data in NMT. In NMT research, the scarcity of bilingual data often lets researchers turn to the utilization of monolingual data. To our knowledge, Gülçehre et al. (2015) first investigated the usage of monolingual data by training an LM on the target language data and integrating it in the inference phase. Sriram et al. (2018) proposed to keep the LM fixed, but train the NMT model from scratch by taking LM as a component. Back-translation (Sennrich et al., 2016; He et al., 2016; Fadaee et al., 2017; Edunov et al., 2018) enhances the training data by translating the target language data back into the source language. Among the recent works, Cai et al. (2021) proposed jointly training the NMT model with a retriever that retrieves similar target sentences from corpora and enhancing NMT with the retrievals. Reheman et al. (2024) proposed to leverage similar target sentences to construct pseudo bilingual sentences and perform kNN-MT. Their difference from ours is that we take the cross-lingual similarity and n-gram segments into account and combine a non-parametric LM with the NMT.

kNN Based Methods in NMT. kNN-LM (Khandelwal et al., 2020) and kNN-MT (Khandelwal et al., 2021) represent the first attempts of kNN methods on language modeling and NMT task, respectively. For efficiency, a series of studies has been proposed to optimize kNN-MT in MT community. Martins et al. (2022b); Wang et al. (2022a) proposed to reduce the datastore scale or key dimension. By utilizing word alignment, Meng et al. (2022) narrows down the search space in advance to accelerate the kNN search. Instead of retrieving one token in each step, Martins et al. (2022a) proposed to retrieve a chunk of tokens at a time. From the perspective of denoising, Zheng et al. (2021a) and Jiang et al. (2022) proposed to dynamically control the information that comes from nearest neighbors.

7 Conclusion

We propose an innovative approach to enhance the domain adaptability of NMT by leveraging target language data. We incorporate translation information from semantically similar target sentences using the kNN-LM framework. To fully utilize the similar target segments, we also propose an n-gram-based method. Additionally, we further improve the model's robustness to noise by incorporating cross-lingual retrieval-based similarity. The experimental results demonstrate significant performance improvements over baselines across high-resource and low-resource translation settings. In future work, we will explore methods that are more robust to neighbors with varying distances.

8 Limitations

Despite its powerful performance, our method still has limitations as follows. Since our approach applies a cross-lingual retrieval model to search similar sentences from the target language data, its performance directly affects the recall rate of similar sentences and, consequently, the overall translation. Additionally, ensuring the consistency between the vocabulary of the LM and the target vocabulary of the NMT model is essential for interpolating the probability distributions.

9 Acknowledgements

This work was supported in part by the National Science Foundation of China (Nos. 62276056 and U24A20334), the Fundamental Research Funds for the Central Universities, the Yunnan Fundamental Research Projects (No. 202401BC070021), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No.B16009).

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7747–7763. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M. Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733.

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*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Bram Bulté and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1800–1809. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7307–7318. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 3042–3047. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 1: Research Papers*, pages 53–63. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 567–573. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, Nicola Bertoldi, and Marcello Federico. 2017. Neural vs. phrase-based machine translation in a multi-domain scenario. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 280–284. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Learning multilingual sentence representations with cross-lingual consistency regularization. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track, Singapore, December 6-10, 2023, pages 243–262. Association for Computational Linguistics.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Comput. Speech Lang.*, 45:137–148.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2101–2112. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George F. Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2486–2496. Association for Computational Linguistics.

Hui Jiang, Ziyao Lu, Fandong Meng, Chulun Zhou, Jie Zhou, Degen Huang, and Jinsong Su. 2022. Towards robust k-nearest-neighbor machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5468–5477. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39. Association for Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Yingfeng Luo, Tong Zheng, Yongyu Mu, Bei Li, Qinghong Zhang, Yongqi Gao, Ziqiang Xu, Peinan Feng, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2025. Beyond decoder-only: Large language models can be good encoders for machine translation. *CoRR*, abs/2503.06594.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5990–5997. Association for Computational Linguistics.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2022a. Chunk-based nearest neighbor machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4228–4245. Association for Computational Linguistics.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2022b. Efficient machine translation domain adaptation. *CoRR*, abs/2204.12608.

Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. 2022. Fast nearest neighbor machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 555–565, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 314–319. Association for Computational Linguistics.

Jianhui Pang, Baosong Yang, Derek Fai Wong, Yu Wan, Dayiheng Liu, Lidia S. Chao, and Jun Xie. 2024. Rethinking the exploitation of monolingual data for low-resource neural machine translation. *Comput. Linguistics*, 50(1):25–47.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abudurexiti Reheman, Yingfeng Luo, Junhao Ruan, Chunliang Zhang, Anxiang Ma, Tong Xiao, and JingBo Zhu. 2024. Exploiting target language data for neural machine translation beyond back translation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 12216–12228. Association for Computational Linguistics.

Abudurexiti Reheman, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. 2023. Prompting neural machine translation with translation memories. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13519–13527. AAAI Press.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022. COMET-22: unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 578–585. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2018. Cold fusion: Training seq2seq models together with language models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Dexin Wang, Kai Fan, Boxing Chen, and Deyi Xiong. 2022a. Efficient cluster-based k-nearest-neighbor machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2175–2187. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533.

Frank F. Xu, Uri Alon, and Graham Neubig. 2023. Why do nearest neighbor language models work? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 38325–38341. PMLR.

Jitao Xu, Josep Maria 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, ACL 2020, Online, July 5-10, 2020*, pages 1580–1590. Association for Computational Linguistics.

Tong Zheng, Yan Wen, Huiwen Bao, Junfeng Guo, and Heng Huang. 2025. Asymmetric conflict and synergy in post-training for llm-based multilingual machine translation. *CoRR*, abs/2502.11223.

Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021a. Adaptive nearest neighbor machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 368–374. Association for Computational Linguistics.

Xin Zheng, Zhirui Zhang, Shujian Huang, Boxing Chen, Jun Xie, Weihua Luo, and Jiajun Chen. 2021b. Nonparametric unsupervised domain adaptation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4234–4241. Association for Computational Linguistics.

A COMET Scores

We give some experimental results on COMET metric here. Table 8 and Table 9 correspond to Table 3 and 4 from the main part, respectively.

Methods	IT	Koran	Law	Medical	Average
NMT	77.40	59.88	75.67	77.36	72.58
BT	79.65	63.12	81.19	81.07	76.26
UDA-kNN	78.66	63.24	78.07	79.50	74.87
Pseudo-kNN	78.21	62.82	76.75	78.86	74.16
Bt-kNN	78.50	62.69	77.60	79.41	74.55
kNN-UD	78.23	62.18	77.21	78.96	74.15
kNN-UP	78.29	62.14	77.47	79.28	74.30
n-gram-UD	78.38	62.01	77.43	79.15	74.24
n-gram-UP	78.53	62.18	76.97	79.41	74.27

Table 8: COMET scores of WMT21 De-En sample 500k model on the multi-domain test sets.

Methods	IT	Koran	Law	Medical	Average
NMT	82.62	72.46	85.61	83.41	81.03
kNN-UD	82.49	72.33	85.40	83.25	80.87
kNN-UP	82.30	72.17	86.19	83.05	80.93
n-gram-UD	83.33	73.06	86.24	84.17	81.70
n-gram-UP	83.45	73.17	86.51	84.28	81.85

Table 9: Comet scores of WMT19 De-En Llama3.1-dict model on the multi-domain test sets.

B Other results

To determine either the "last FFN input" or the "last hidden state" of the Llama3.1-8B model gets better representation, we conduct this experiment. The results reveal that the "last FFN input" demonstrate superior performance. Consequently, we select it in the formal experiments.

Key	Methods	IT	Koran	Law	Medical
-	NMT	35.90	16.95	45.04	39.77
FFN	kNN-UD	38.03	19.58	51.62	45.97
	kNN-UP	38.92	19.21	52.94	46.61
Hid	kNN-UD	38.31	18.64	46.24	41.87
	kNN-UP	38.56	18.72	46.26	41.82

Table 10: SacreBLEU scores of WMT19 De-En Llama3.1-dict model on the multidomain test sets. "FFN" denotes the last ffn input as the key, "Hid" denotes last hidden state as the key.