RAG Picking Helps: Retrieval Augmented Generation for Machine Translation

Anonymous ACL submission

Abstract

Machine Translation (MT) has considerably improved over the years, especially with the introduction of neural approaches. However, such approaches lack the ability to tackle scenarios like domain adaptation and low-resource settings due to their dependence on only their parametric knowledge. We explore using a retrieval mechanism for MT and provide a detailed analysis of the quantitative and qualitative improvements obtained by its use. We introduce RAGMT, a retrieval augmented generation (RAG)-based multi-task fine-tuning approach for Machine Translation (MT) using non-parametric knowledge sources. We also propose using new auxiliary training objectives that improve the performance of RAG for domain-specific MT. To the best of our knowledge, we are the first to adapt the RAG framework with a multi-task training objective for MT to support end-to-end training. Our experiments demonstrate that retrieval-augmented fine-tuning of MT models under the RAGMT framework results in an average improvement of 12.90 BLEU scores compared to simple fine-tuning approaches on English-German domain-specific translation. We also demonstrate RAGMT's ability to exploit in-domain knowledge bases versus domain-agnostic ones and perform careful ablations over the model components. Qualitatively, RAGMT is easily interpretable, stylistically aligns translation outputs to the domain of interest, and appears to demonstrate "copy-over-translation" behaviour with respect to named entities.

1 Introduction

Neural Machine Translation (NMT) has shown significant improvements in its ability to produce high-quality translations. However, NMT systems often struggle to maintain accuracy and fluency in specialized domains such as medicine, law, and information technology, where domain-specific terminology, sentence structures, tone and context play a crucial role (Chu and Wang, 2018, 2020). General translation models trained on generic datasets lack the ability to capture the nuances and intricacies of these specialized domains, leading to suboptimal translation quality that may fail to convey the intended meaning accurately. This serves as a strong motivation for us to explore the use of non-parametric methods, specifically RAG.

The problem of domain adaptation of translation can be stated as obtaining high-quality translation for a specific domain of interest. When fine-tuning a pretrained NMT model for a particular domain, some of the key challenges include limited availability of in-domain data, catastrophic forgetting, and inadequacy to adapt to domain style and tonality (Saunders, 2021).

Integrating non-parametric memory to parametric neural networks (Khandelwal et al., 2019; Guu et al., 2020; Lewis et al., 2020) has shown great promise when it comes to language models. For the task of MT, various approaches and integration of different types of non-parametric memories have been explored (Bulté and Tezcan, 2019; Moussallem et al., 2019; Zhao et al., 2020; Khandelwal et al., 2020; Zhang et al., 2021; Cai et al., 2021; He et al., 2021; Hoang et al., 2022; Cheng et al., 2023). By incorporating relevant information from these external sources, MT systems can produce more accurate and contextually appropriate translations tailored to the specific domain.

Despite their differences, all the approaches mentioned above either lack the ability to train the model to effectively utilize the retrieved documents and train the memory retriever to retrieve contextually highly relevant documents or fail to make use of weak signals to train the retrieval mechanism. For better domain adaptation, we require to train a model in such a way that it improves the downstream translation task along with its ability of domain-style adaptation and accurate entity translation. Although it has been shown that translation quality improves with retrieval mechanisms, no methodology to train the NMT and retriever component jointly for domain adaptation has been explored.

In this work, we propose a novel fine-tuning approach **RAGMT** to enhance MT systems using an end-to-end multi-task RAG framework for retrieval-augmented machine translation. Our approach builds upon the RAG framework (Lewis et al., 2020), combining document retrieval with a generative model to produce translations enriched with domain-specific knowledge. We utilize a multi-task framework and introduce an explicit document similarity term to the training objective of retrieval-augmented NMT. This results in improved effectiveness of the model for domain-adaptation scenarios.

Our contributions are:

1. **RAGMT** (section 3), a new RAG-based multi-task fine-tuning approach for machine translation with a new end-to-end training objective, along with **Entity masked language modelling (MLM)** as an auxiliary task (Song et al., 2019) and explicit document similarity term that boosts documents that are very similar to the source sentence, in contrast, penalizing documents that are further off.[1]

2. In-depth analysis (section 5) of our proposed approach on domain-specific machine translation using knowledge graphs (KG) as non-parametric sources. Compared with neural and retrieval-based baselines, we achieve an *average improvement of +12.90 BLEU score across domains*. Additionally, we demonstrate that domain-specific knowledge sources provide an *average improvement of +0.625 BLEU score* over domain-agnostic sources.

3. Ablation study (section 5) on the proposed **RAGMT** training objective, quantifying the contribution of each loss term. Our analysis highlights the impact of the document similarity term with an *average improvement of 1.125 BLEU scores* across domains.

We intend RAGMT to be a generalized framework for retrieval-augmented fine-tuning of NMT models. Hence, we describe it as a modular framework with flexible plug-and-play components.

[1]The codebase for RAGMT and the datasets to replicate our results will be released upon publication.

2 Background and Related Work

Domain-specific Machine Translation Transformer models, such as Raffel et al. (2019), Lample and Conneau (2019), Shazeer et al. (2017), and NLLB Team et al. (2022), have become foundational in NMT due to their ability to handle complex linguistic structures and long sequences. Recent works, including Alves et al. (2024), Wei et al. (2021), Yang et al. (2023) and Zhang et al. (2023) explore the use of LLMs for translation.

Despite these advancements, general models struggle to perform well in domain adaptation scenarios. For domain-specific MT, methods such as Luong and Manning (2015); Khayrallah et al. (2018); Thompson et al. (2019a,b); Lu et al. (2023); Ghazvininejad et al. (2023); Moslem et al. (2023); Anonymous (2024) have been proposed.

Some methodologies, including Khandelwal et al. (2020); Cai et al. (2021); Hoang et al. (2022); Ghazvininejad et al. (2023); Moslem et al. (2023), are particularly focused on integrating non-parametric knowledge in the process of translation.

Retrieval for Text Generation. This class of techniques represent methods for integrating external knowledge for text generation. Retrieval Augmented Generation (RAG) (Khandelwal et al., 2019; Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2021) combines information retrieval with generation, allowing to leverage retrieved documents for better context. Works including Karpukhin et al. (2020); Siriwardhana et al. (2022) advance RAG models in open-domain question answering through domain adaptation. Retrieval augmented text generation has significant advancements, including Lin et al. (2023); Asai et al. (2023); Xu et al. (2023); Wang et al. (2023); Shi et al. (2023). Recent work has also focused on retrieval-augmented fine-tuning, including Lin et al. (2023); Wang et al. (2023); Xu et al. (2023); Liu et al. (2024); Zhang et al. (2024).

3 RAGMT

In this section, we describe the proposed RAGMT approach. First, we formulate machine translation as a retrieve-then-generate process in section 3.1. We then describe the constituent components of RAGMT in section 3.2. RAGMT is described as a framework that allows for end-to-end retrieval-augmented training of NMT models with interchangeable plug-and-play components. In section 3.3, we formulate the auxiliary task used in RAGMT. Lastly, in section 3.4, we describe how the components of RAGMT are optimized end-to-end.

3.1 Problem Formulation

Given an input sentence S in the source language, $S = (s_1, s_2, \dots, s_m)$, the problem of retrieval-augmented machine translation can be formulated as finding the target sentence, $\hat{T} = (t_1, t_2, \dots, t_n)$, by first retrieving a set of helpful documents, $D = \{d_i\}_{i=1}^{K}$, from an external knowledge base, where K is the number of retrieved documents. Then, the generation of the target sentence is conditioned on both the source sentence, S, along with the documents from the retrieved set, D, as given by equation (1).

$$\hat{T} = \underset{T}{\operatorname{argmax}} \sum_{d \in D} P(d|S) P(T|d, S) \tag{1}$$

Knowledge base is a generic term denoting various structures, including KG triples, textual documents, and precomputed embeddings.

3.2 Overview

RAGMT framework (illustrated in Figure 1) consists of four main components: *knowledge base*, *retriever*, *integrator* and *generator*. The knowledge base is a collection of documents that can consist of structured information, such as KGs and wordnets, or unstructured information, such as translation memory.

Similar to Bromley et al. (1993), we use a dual encoder structure for the retriever. It consists of the document encoder, Encoder_D and the source encoder, Encoder_S. The relevance score between a source sentence S, and a candidate document, d, is defined as the dot product of their encodings:

$$f(S, d) = \text{Encoder}_D(d)^T \text{Encoder}_S(S)$$

The encoding of the documents is generated using the document encoder and is stored in a vector index. We use FAISS (Johnson et al., 2019) for this purpose. When a source sentence is provided to the retriever, it encodes the sentence and passes it to FAISS to retrieve the most relevant documents from the knowledge base. The retriever component $P_\eta(d|S)$, parametrized by η gives the relevance of the document d, given the source sentence, S as:

$$P_\eta(d|S) \propto \exp(\text{Encoder}_D(d)^T \text{Encoder}_S(S))$$

Given the source sentence, S, and the retrieved set of documents, D, the generator finally performs the downstream translation task and the auxiliary task of entity MLM (described in section 3.3). The conditional probability for a translation candidate T, given by the generator for the translation task, is defined as:

$$P(T|S, D) = \prod_{i=i}^{n} P(t_i|S, D, t_{<i})$$

$$= \prod_{i=i}^{n} \sum_{d \in D} P_\eta(d|S) P_\theta(t_i|S, d, t_{<i})$$

where, $P_\theta(t_i|S, d, t_{<i})$, parametrized by θ, gives the probability of generation of the current token, t_i, based on the source sentence, S, one of the documents from the retrieved set of documents, d and the previous generation context, $t_{<i}$.

As there is no optimal strategy for encoding the retrieved documents and the source sentence to prepare the input for the generator, we introduce a plug-and-play integrator component. Different types of knowledge bases can require different integration strategies. KGs, for example, have structural information, which needs to be encoded in the input to the generator (Shen et al., 2020; Sun et al., 2020; Wen et al., 2024). The simplest integration strategy is to prepend the source sentence with documents from the retrieved set.

RAGMT enables training of the two parametric components, the retriever and the generator, with the use of different knowledge bases and integration strategies. Similar to Lewis et al. (2020), RAGMT does end-to-end propagation of the gradients, allowing for joint training of the two components.

The translation output is obtained using the generator with the loss function, L_G, as given in equation 2.

$$L_G = -\sum_{i=1}^{n} \log P(t_i|S, D, t_{<i}) \tag{2}$$

3.3 Auxiliary Task: Entity Masked Language Modelling

We introduce an auxiliary task derived from entity-masked language modelling (E-MLM) (Song et al.,

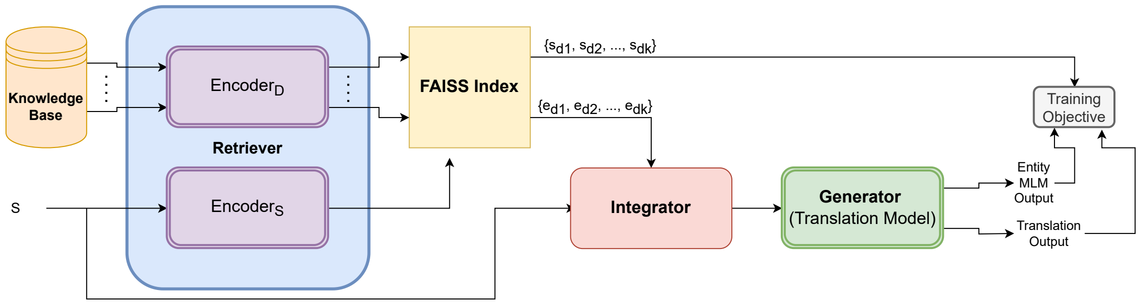

Figure 1: **RAGMT Architecture**: The KB consists of documents to be retrieved, which are indexed using FAISS over the embeddings computed using Encoder$_D$. For a source sentence, S, The retriever first encodes S using Encoder$_S$, then retrieves documents using the FAISS index. The retrieved documents, along with the source sentence, are then inputs for the Integrator, which outputs the formatted input to be used by the Generator.

2019; Siriwardhana et al., 2022) training to enhance the model's capacity to integrate external knowledge. This auxiliary task supplements the primary training objective by providing additional context about named entities in the input text. By training the model to predict masked entities within the input text, we aim to improve its understanding of domain-specific terminology and entities, enhancing translation accuracy and domain adaptation capabilities.

For a particular training pair, (S, T), where S is the source sentence and T is the target sentence, let the retrieved results from the retriever be $D = R(S) = \text{top} - \text{k}(P_\eta(.|x)) = \{d_1, \ldots, d_K\}$ where η parameterizes the retriever model, R. Let S_M be the source sentence with named entities masked. The task of entity MLM is to predict the masked entities in the source sentence, given the set D and S_M, as stated in Equation (3).

$$\hat{S} = \underset{S}{\operatorname{argmax}} P(S|S_M, D) \tag{3}$$

This auxiliary task is a form of multi-task learning, where multiple learning tasks are performed simultaneously, and each task aids the learning of the other task.

Equation (4) below shows the loss function for entity MLM loss, where $M = \{m_1, m_2, \ldots, m_k\}$ is the set of positions in the entity masked source sentence, corresponding to named entities.

$$L_{\text{MLM}} = -\sum_{m \in M} \log P(\hat{s}_m = s_m | S_M, D) \tag{4}$$

where $\hat{s}_m$ denotes the source token predicted by the generator.

With its entity reconstruction objective, the entity MLM loss further aligns the model's outputs with the retrieved documents. This auxiliary loss complements the primary loss (L_G) by encouraging the model to produce fluent, accurate translations closely aligned with the content and context of the retrieved documents.

3.4 Training

Along with the generator loss, L_G, RAGMT explicitly models the similarity between the source sentence, S, and the retrieved document set, $D = \{d_i\}_{i=1}^{K}$ with a document similarity-based loss, L_D, as given in equation 5.

$$L_D = \left(-\sum_{i=1}^{K} \log(s_{d_i}) \right) \tag{5}$$

where, s_{d_i} is given by $f(S, d_i)$.

The model parameters of RAGMT, η and θ are optimized using the final training objective:

$$L = L_G \cdot L_D + L_{\text{MLM}}$$

where L_G is the generator model's loss and L_{MLM} is the entity MLM loss, and L_D is the document similarity term. The first component of the RAGMT training objective is a product between L_G and L_D, as both the retrieval and generation processes are mutually dependent. By multiplying the two terms, we enforce that both retrieval and generation work well together. If the retrieval component retrieves irrelevant documents, the translation generated suffers; conversely, if the generator doesn't utilize the retrieved documents effectively, the overall translation still suffers.[2]

[2]We conducted experiments with the more traditional formulation of the training objective, $L = L_G + L_D + L_{\text{MLM}}$, but found that our formulation performs significantly better where the external knowledge base actively contributes to the generation.

Domain	# Training Samples	# KG Triples
Law	222927	454148
Medical	17982	37176
Koran	467310	753082
IT	248099	471002

Table 1: Dataset statistics: English-German Domain Specific Parallel Corpus. The table shows the number of training data points in the dataset, along with the number of knowledge graph triples extracted as described in section 4.1.

Model Name	IT	Koran	Law	Medical
Baseline FT	38.35	16.26	45.48	39.99
Hoang et al. (2022)	33.84	**27.53**	52.17	46.95
Khandelwal et al. (2020)	48.63	19.22	61.11	**54.54**
Cai et al. (2021)	35.33	16.26	53.97	50.32
Ghazvininejad et al. (2023)	33.58	20.34	45.92	50.38
RAGMT	**49.12**	26.99	**61.23**	54.36

Table 2: Comparison of BLEU score of different setups on domain adaptation. Each setup is described in section 4.

4 Experiments

4.1 Dataset

We utilized the English and German parallel corpus from Aharoni and Goldberg (2020), a re-split version of the multi-domain data set from Koehn and Knowles (2017). The dataset comprises Law, Medical, Koran, IT and Subtitles domains. We leave out the Subtitles domains from all our experiments since the data lacks consistency in terms of the constituent topics. Hence, a cohesive knowledge base could not be constructed from the data. Each domain consists of 2000 validation and 2000 test points.

For our main experiment, comparing the domain adaptation of MT, we use the complete training sets of each domain and translate in German-to-English direction.

For all the other experiments, we use a randomly sampled subset of 15000 data points from the training set of each domain. This was done primarily for two reasons: 1) We wanted to restrict the amount of available fine-tuning data to reflect real-world settings where domain-specific fine-tuning data is limited. 2) Our available compute was insufficient to run experiments using the entire training datasets. In this constrained setting, we have carefully compared it against existing baseline systems, as detailed below.

Knowledge Base We conduct all our experiments with a knowledge base made up of knowledge graphs. For this purpose, we extract in-domain knowledge graphs for each domain mentioned above, using a pre-trained multilingual model, REBEL (Resource Extraction from BERT Embeddings for Linked data) (Cabot and Navigli, 2021). The dataset statistics have been depicted in table 1.

4.2 Implementation Details

RAGMT is set up with a retriever based on Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) and generator based on NLLB200 (NLLB Team et al., 2022).[3,4]

The RAGMT setup that we use for our experiments consists of the following components:

1. *Knowledge Base*: We use knowledge graphs as the knowledge base. Each document consists of a KG triple of the form, *<h,r,t>*, where h is the head, t is the tail, and r is the relationship of the triple, respectively.

2. *Retriever*: The retriever consists of a dual-encoder setup, as described in section 3.2. We use Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) for the document and source encoders.

3. *Generator*: We use the pre-trained 600M parameter checkpoint of NLLB-200 (NLLB Team et al., 2022) as the generator.

4. *Integrator*: For all our experiments, we use the simple integration strategy of prepending the source sentence, S, with the retrieved document, d, so the input for the generator after we obtain the retrieved set of documents, becomes *<d, [SEP], S>*.

4.3 Baselines

RAGMT is compared with the following baselines: Baseline FT, Khandelwal et al. (2020), Hoang et al. (2022), Cai et al. (2021) and Ghazvininejad et al. (2023). We consider an NMT model fine-tuned without any retrieval mechanism as our baseline (*Baseline FT*). For this purpose, we fine-tune the 600M parameter checkpoint of NLLB-200 NLLB Team et al. (2022).[5,6]

[3]Available at: `https://huggingface.co/facebook/nllb-200-distilled-600M`

[4]We discuss the training details in appendix A.

[5]Details about the compared models are presented in appendix B.

[6]We couldn't compare our work with (Anonymous, 2024), as the code was not publicly available, and the experimental settings presented in their work differ from ours.

	BLEU		chrF++		TER		BERTScore		COMET	
Domain	Baseline FT	RAGMT	Baseline FT	RAGMT	Baseline FT	RAGMT	Baseline FT	RAGMT	Baseline FT	RAGMT
German → English										
IT	38.35	**49.12**	55.00	**72.50**	52.23	**33.42**	0.89	**0.89**	0.68	**0.80**
Koran	16.26	**26.99**	30.50	**44.00**	75.42	**58.66**	0.60	**0.75**	0.45	**0.58**
Law	45.48	**61.23**	60.00	**78.00**	40.65	**28.58**	0.85	**0.91**	0.75	**0.85**
Medical	39.99	**54.36**	54.50	**74.50**	50.76	**31.75**	0.78	**0.88**	0.70	**0.83**
English → German										
IT	35.36	**45.32**	61.28	**65.12**	50.43	**44.56**	0.89	**0.90**	0.48	**0.60**
Koran	18.43	**24.34**	**43.67**	40.23	**68.12**	70.15	**0.83**	0.78	0.25	**0.30**
Law	42.34	**57.23**	66.45	**68.75**	45.87	**42.10**	0.91	**0.93**	0.58	**0.65**
Medical	37.54	**50.34**	63.12	**64.80**	48.35	**46.25**	0.90	**0.91**	0.53	**0.55**

Table 3: Comparison of metrics for the baseline and RAGMT setups in En→De and De→En directions.

Model Name	BLEU	chrf++	TER	BERTScore	COMET
Law Domain					
(1) Baseline FT	35.45	60.07	47.51	**0.84**	0.51
(2) ConceptNet	36.23	61.73	45.18	0.81	0.52
(3) In-domain KG	**37.42**	**62.12**	**43.79**	0.82	**0.54**
Medical Domain					
(1) Baseline FT	36.60	57.26	42.33	0.78	0.55
(2) ConceptNet	38.82	58.61	42.15	0.79	0.56
(3) In-domain KG	**39.12**	**59.12**	**41.55**	**0.83**	**0.58**
Koran Domain					
(1) Baseline FT	20.85	43.02	63.21	0.74	0.21
(2) ConceptNet	**22.56**	**45.94**	62.22	**0.79**	0.28
(3) In-domain KG	22.34	44.37	**61.78**	0.76	**0.30**
IT Domain					
(1) Baseline MT	27.77	48.69	54.64	0.79	0.38
(2) ConceptNet	28.71	48.92	53.37	0.78	0.38
(3) In-domain KG	**29.94**	**49.12**	**52.30**	**0.81**	**0.41**

Table 4: Comparison of domain-agnostic vs domain-specific knowledge graph with RAGMT across various domains.

	Baseline FT	RAGMT
Training Latency	1x	3x
Inference Latency	1x	1.71x

Table 5: Comparison of latency between baseline fine-tuning approach and RAGMT fine-tuning

Approach	Law	Medical	Koran	IT
Baseline FT	35.45	36.60	20.85	27.77
RAGMT - w/o L_{MLM}	37.17	39.02	21.98	29.68
RAGMT - w/o L_D	36.52	38.94	20.64	28.21
RAGMT	**37.42**	**39.12**	**22.34**	**29.94**

Table 6: Ablation on the RAGMT training objective. The BLEU scores obtained across all the domains, using different settings described in section 5.

Example	**Retrieved Documents**	**Translation Outputs**
(Source) Your doctor will prescribe Truvada with other antiretroviral medicines. **(Reference Transalation)** Ihr Arzt wird Ihnen Truvada in Kombination mit anderen antiretroviralen Arzneimitteln verschreiben.	(1) Truvada instance of antiretroviral combination therapy (2) Truvada instance of antiretroviral therapies	Ihr Arzt wird Truvada zusammen mit anderen antiretroviralen Arzneimitteln verschreiben.
(Source) Convert current frame to an inline frame **(Reference Translation)** Aktuellen Rahmen in einen im Text mitfließenden Rahmen umwandeln	(1) convert files facet of file format (2) inline frames type of frames	Aktuellen Rahmen in einen Inline-Rahmen umwandeln

Table 7: Example translation using RAGMT. The retrieved documents are contextually relevant to the source as well as target sentence, with the retrieved entities being used in both the source as well as the target sentence.

4.4 Evaluation

For evaluating the performance of all the setups in our experiments, we utilize a comprehensive set of metrics, including BLEU (Post, 2018), chrF++ (Popović, 2015), TER (Snover et al., 2006), BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020).[7]

5 Results and Analysis

Domain adaptation of MT. We first test if fine-tuning the proposed framework using a domain-specific dataset, with a retrieval mechanism applied over an in-domain KG, would improve performance over the baseline approaches. Table 2 shows the BLEU scores of all the compared approaches on the domain adaptation experiment for German to English translation. Compared to the Baseline MT, RAGMT improves performance by an average of 12.90 BLEU scores, with the largest improve-

[7]We defer the discussion on metrics other than BLEU score to appendix.

ment on the Law domain data with 15.75 BLEU score improvement. This signifies that fine-tuning an NMT model using the RAGMT framework for MT on a domain-specific dataset with access to an in-domain knowledge base, such as KGs, helps improve the performance of the MT model. Comparing the proposed RAGMT framework with the Khandelwal et al. (2020), we observe an average improvement of 2.05 BLEU scores, with the largest improvement of 7.77 BLEU scores on the Koran domain.

To further test the generalizability of RAGMT, we conduct experiments in both translation directions. The results across various metrics have been shown in table 3 for the baseline fine-tuning approach and RAGMT.

Domain-specific KG vs Domain-agnostic KG. We explore the effects of using an in-domain knowledge base as opposed to a domain-agnostic knowledge base. We use the domain-specific KGs extracted for each domain (as described in section 4.1) for fine-tuning RAGMT for the respective domain. We use ConceptNet (Speer et al., 2016) as our domain-agnostic KG. Table 4 shows the difference in the performance of the RAGMT framework with in-domain KG instead of using a domain-agnostic KG. Using domain-specific KG, we observe an average improvement of 0.62 BLEU scores over the use of ConceptNet, with improvements in three of the four domains. We analyze the performance degradation in the Koran domain later in this section.

Latency RAGMT employees a retrieve-then-translate mechanism with end-to-end gradient propagation during training. Compared to the baseline fine-tuning approach, RAGMT introduces additional parameters due to the addition of the retriever component. We compare the latency incurred by RAGMT fine-tuning over the baseline fine-tuning approach. For both the training and inference, we retrieve top 5 documents from the knowledge base. We observe that RAGMT is nearly 3 orders of magnitude slower than the baseline during training, while nearly 1.7 orders of magnitude slower during inference.

Ablations on the RAGMT training objective. We analyze the contribution of each of the constituent components of the training objective as described in section 3.4. We compare the performance of RAGMT framework under the following settings: (1) **RAGMT - w/o $L_{\mathbf{MLM}}$**, the RAGMT training objective without the loss from the Entity MLM component; (2) **RAGMT - w/o L_D**, the RAGMT objective without the explicit Document Similarity component; and (3) **RAGMT**, the training objective as described in section 3.4.

Table 6 presents the BLEU score comparison across domains for each ablation. Across all domains, the variations of the RAGMT training objective result in higher BLEU scores than the baseline. The obtained results signify that the *Document Similarity* component substantially contributes to the training objective with an average difference of 1.12 BLEU score due to its removal. The loss from the Entity MLM component results in an average difference of 0.24 BLEU scores across domains. Overall, we observe consistent improvement in performance across domains with the addition of each of the two components, showing the efficacy of the proposed RAGMT training objective and justifying the inclusion of each component.

Quantitative and Qualitative Analysis. We quantitatively analyze the benefits of using a non-parametric knowledge base for MT using the RAGMT framework by looking at the entity overlap in the translation outputs. More precisely, for each entity present in the translation output, we categorize the entity into four categories: (1) Present only in the source sentence; (2) Present only in the knowledge base; (3) Present in both; (4) Present in neither. While using an in-domain datastore, on average, the entities are present in both the source sentence and knowledge base 38.5% times, as opposed to the domain-agnostic knowledge base, where entities are present 35.25% times. Compared with the domain-agnostic KG, we see a lower proportion of entities being exclusively present only in the KG for all domains except Koran. Unlike the other three domains, Koran has 19% translated entities exclusively present in the domain-agnostic KG setup and only 11% translated entities exclusively present in the domain-specific KG. This potentially explains why the domain-agnostic KG yields higher BLEU scores for the Koran domain compared to the domain-specific KG. Table 7 shows a few examples of translations performed using the RAGMT framework. For the second example (taken from the IT domain), we can observe that the reference translation does not consist of the phrase *inline frame*, but it is present in the translation output.

To further investigate the domain adaptation capabilities of RAGMT, we conducted a clustering-based analysis. Specifically, a domain fine-tuned BERT (Devlin et al., 2019) encoder was employed to extract dense representations of target-side sentences from the training sets across different domains. These representations were used to identify cluster centers for each domain. We then evaluated how closely the translations generated by the RAGMT system aligned with their respective domain-specific cluster centers. On average, the translations achieved an alignment accuracy of 89%, with the law domain exhibiting the highest accuracy at 95%, and the Koran domain demonstrating the lowest accuracy at 79%.

6 Conclusion and Future Work

We address the shortcomings of NMT models due to their reliance on just their parametric knowledge. We present RAGMT, a RAG-based multi-task MT fine-tuning approach to enhance machine translation using non-parametric knowledge bases. Compared to existing baselines, we show the efficacy of our approach to the problem of domain-specific MT using knowledge graphs as the knowledge base. Our approach improves the performance of the baseline MT model using both domain-agnostic and domain-specific knowledge graphs across all domains. For future work, we aim to use the proposed framework for other nuanced MT tasks, such as low-resource language adaptation, accurate entity translation, and usage of other non-parametric knowledge sources.

7 Limitations

- We study the efficacy of RAGMT for the setting of domain-adaptation of MT. The same framework can be adapted for low-resource MT settings, however, the efficacy and analysis of RAGMT for such a setting is yet to be studied.

- There is an inherent trade-off with increasing the number of retrieved documents using RAG versus improving BLEU scores. The former can improve the quality of the generated translations but leads to increased computational overhead. This balance needs to be considered depending on the downstream task.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*.

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

Anonymous. 2024. DragFT: Adapting large language models with dictionary and retrieval augmented fine-tuning for domain-specific machine translation. In *Submitted to ACL Rolling Review - June 2024*. Under review.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *ArXiv*, abs/2310.11511.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. 2021. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann.

Bram Bulté and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Conference on Empirical Methods in Natural Language Processing*.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual translation memory. In *Annual Meeting of the Association for Computational Linguistics*.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023. Lift yourself up: Retrieval-augmented text generation with self-memory. In *Advances in Neural Information Processing Systems*, volume 36, pages 43780–43799. Curran Associates, Inc.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2020. A survey of domain adaptation for machine translation. *J. Inf. Process.*, 28:413–426.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *ArXiv*, abs/2302.07856.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *ArXiv*, abs/2002.08909.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Annual Meeting of the Association for Computational Linguistics*.

Cuong Hoang, Devendra Singh Sachan, Prashant Mathur, Brian Thompson, and Marcello Federico. 2022. Improving retrieval augmented neural machine translation by controlling source and fuzzy-match interactions. *ArXiv*, abs/2210.05047.

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

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest neighbor machine translation. *ArXiv*, abs/2010.00710.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *ArXiv*, abs/1911.00172.

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. 2018. Regularized training objective for continued training for domain adaptation in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 36–44, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *NMT@ACL*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *ArXiv*, abs/1901.07291.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke S. Zettlemoyer, and Scott Yih. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *ArXiv*, abs/2310.01352.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Building gpt-4 level conversational qa models. *ArXiv*, abs/2401.10225.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *ArXiv*, abs/2305.06575.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. *ArXiv*, abs/2310.14451.

Diego Moussallem, Mihael Arcan, Axel-Cyrille Ngonga Ngomo, and Paul Buitelaar. 2019. Augmenting neural machine translation with knowledge graphs. *ArXiv*, abs/1902.08816.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp

Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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, Belgium, Brussels. Association for Computational Linguistics.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *ArXiv*, abs/2009.09025.

Danielle Saunders. 2021. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *J. Artif. Intell. Res.*, 75:351–424.

Noam M. Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ArXiv*, abs/1701.06538.

Tao Shen, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, and Weizhu Chen. 2020. Exploiting structured knowledge in text via graph-guided representation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8980–8994, Online. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models. In *North American Chapter of the Association for Computational Linguistics*.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Kumar Rana, and Suranga Nanayakkara. 2022. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *ArXiv*, abs/1905.02450.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. CoLAKE: Contextualized language and knowledge embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3660–3670, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019a. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Thompson, Rebecca Knowles, Xuan Zhang, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019b. HABLex: Human annotated bilingual lexicons for experiments in machine translation. 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 1382–1387, Hong Kong, China. Association for Computational Linguistics.

Boxin Wang, Wei Ping, Lawrence C. McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Instructretro: Instruction tuning post retrieval-augmented pretraining. *ArXiv*, abs/2310.07713.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. MindMap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10370–10388, Bangkok, Thailand. Association for Computational Linguistics.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence C. McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *ArXiv*, abs/2310.03025.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *ArXiv*, abs/2305.18098.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *ArXiv*, abs/2306.10968.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei A. Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *ArXiv*, abs/2403.10131.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Tong Zhang, Long Zhang, Wei Ye, Bo Li, Jinan Sun, Xiaoyu Zhu, Wenxin Zhao, and Shikun Zhang. 2021. Point, disambiguate and copy: Incorporating bilingual dictionaries for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Yang Zhao, Lu Xiang, Junnan Zhu, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4495–4505, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Training Details

All models have their maximum input and output length set to 1024. We use the Adam optimizer (Kingma and Ba, 2014) and train each setup for a maximum of 50K steps. All the models are trained with *fp16 precision*. We extract the top 5 documents from the knowledge base for all our experiments. As described in section 3.2, we use FAISS (Johnson et al., 2019) to index the knowledge base encoding for faster retrieval during training and inference time.

B Compared Setups

We compare RAGMT with the following setups:

- Baseline FT: We consider an NMT model fine-tuned without any retrieval mechanism as our baseline. For this purpose, we fine-tune the 600M parameter checkpoint of NLLB-200 NLLB Team et al. (2022).
- Khandelwal et al. (2020): This approach uses a k-nearest-neighbour-based retrieval over a translation memory based knowledge base with no additional training of NMT models.
- Hoang et al. (2022): This approach uses a fuzzy-matching-based retrieval mechanism over a source-target translation-based knowledge base and performs a zero-shot adaptation of NMT models.
- Cai et al. (2021): This approach uses monolingual translation memory, retrieves them by source side similarity and adopts a dual encoder (source and target) architecture.
- Ghazvininejad et al. (2023): This approach uses LLMs for translation via dictionary-based prompting.

C Additional Result Analysis

The performance of RAGMT generalizes well across various metrics we use for evaluation. As shown in table 3, RAGMT shows consistent and significant improvement in terms of chrF++, TER, BERTScore and COMET scores, except for the Koran domain, where the baseline FT approach shows better TER and BERTScore values. A similar trend can be observed in table 4, where the in-domain KG performs well compared to domain-agnostic KG for all domains except the Koran domain. We study this behaviour in section 5 by qualitatively analyzing the nature of the Koran domain.

The significant improvement shown by the RAGMT system in our experiments indicate its ability for other nuanced MT tasks, such as low-resource adaptation and accurate entity translation, although a detailed study needs to be conducted to conclusively make the claim.