RAG Picking Helps: Retrieval Augmented Generation for Machine Translation

Anonymous ACL submission

Abstract

We introduce RAGMT, a retrieval augmented generation (RAG)-based multi-task framework for Machine Translation (MT) using non-parametric knowledge sources. To the best of our knowledge, we are the first to adapt the RAG framework for MT to support end-to-end training and use knowledge graphs as the non-parametric source. We also propose the use of new auxiliary training objectives that improve the performance of RAG for domain-specific MT. Our experiments demonstrate that retrieval-augmented fine-tuning of NMT models under the RAGMT framework results in an average improvement of 2.03 BLEU scores over simple fine-tuning approaches on English to German domain-specific translation. We also demonstrate the efficacy of RAGMT with using in-domain versus domain-agnostic knowledge graphs and careful ablations over the model components. Qualitatively, RAGMT is easily interpretable and appears to demonstrate "copy-over-translation" behaviour over named entities.

1 Introduction

Neural Machine Translation (NMT) systems often struggle to maintain accuracy and fluency in specialized domains such as medicine, law, and information technology (IT), where domain-specific terminology and context play a crucial role (Chu and Wang, 2018). Traditional MT models trained on generic datasets lack the ability to capture the nuances and intricacies of these specialized domains, leading to suboptimal translations that may fail to convey the intended meaning accurately.

To address these challenges, researchers have explored various techniques to enhance MT systems' performance in specialized domains. One promising approach involves integrating retrieval mechanisms into the translation pipeline, enabling MT models to access external knowledge sources such as domain-specific documents or knowledge graphs (Zhao et al., 2020; 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.

The convergence between Natural Language Processing (NLP) and Information Retrieval (IR) convergence has given rise to a powerful paradigm called Retrieval Augmented Generation (RAG). Retrieval Augmented Generation (RAG) (Lewis et al., 2020; Guu et al., 2020) represents a paradigm shift in how we approach language understanding and generation tasks. At its core, RAG combines the strengths of traditional IR methods, which excel at retrieving relevant information from vast corpora, with the expressive power of modern NLP models, capable of generating coherent and contextually relevant text. Developing effective RAG systems requires robust methods for seamlessly integrating retrieval and generation components, as these have traditionally been treated as separate modules in NLP pipelines. RAG's potential extends beyond improving output quality; it also offers enhanced interpretability, robustness to input variations, and adaptability to dynamic contexts, making it particularly valuable for applications where transparency is critical, such as legal or medical domains.

In this work, we propose a novel approach **RAGMT** to enhance MT systems using an end-to-end multi-task RAG framework for the task of external memory-augmented machine translation. Our approach builds upon the

RAG framework (Lewis et al., 2020), which combines document retrieval with a generative model to produce translations enriched with domain-specific knowledge. We propose several key enhancements to the RAG framework to improve its effectiveness in translation for specialized domains.

Our main contributions are as follows:

1. We introduce **RAGMT**, a new RAG-based multi-task framework for machine translation with a new end-to-end training objective. The framework allows the integration of different types of non-parametric knowledge sources.[1]

2. Our new training objective includes a specific document similarity term that boosts documents that are very similar to the source sentence while penalizing documents that are further off.

3. We propose the use of **Entity masked language modelling (MLM)** as an auxiliary task for RAGMT (Song et al., 2019). Entity MLM uses a source sentence with its entities masked as its input. This entity-masked source sentence, along with a set of retrieved documents, are used to reconstruct the source sentence, thereby improving the model's ability on domain-specific translation.

4. We conduct an in-depth analysis of our proposed framework 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 +2.03 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.

5. We conduct a detailed ablation study 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.

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

2 Background and Related Work

Transformer-based approaches for NMT. Transformer models, such as T5 (Raffel et al., 2019), XLM (Lample and Conneau, 2019), MoE (Shazeer et al., 2017), and NLLB (NLLB Team et al., 2022), have become foundational in Neural Machine Translation (NMT) due to their ability to handle complex linguistic structures and long sequences. NLLB, designed for multilingual translation across 200 languages, combines a mixture of experts with dense layers, data filtering, and large-scale pretraining to excel in low-resource scenarios.

Knowledge-intensive tasks in NMT. Domain-specific machine translation, especially in specialized fields like medicine or law, demands precise handling of terminology and context, which general-purpose MT systems often fail to achieve. To improve accuracy, strategies such as integrating domain-specific terminologies, fine-tuning with domain-specific parallel corpora, and using domain-specific knowledge graphs have been developed. These approaches not only enhance the translation of specialized terms but also ensure semantic consistency within the domain. Furthermore, they are particularly effective in low-resource scenarios, leveraging multilingual transfer learning and external linguistic resources to improve translation accuracy (Sennrich et al., 2015).

Knowledge structures play a critical role in enhancing NMT by providing additional context and semantic enrichment. For instance, knowledge graphs (KGs) capture complex relationships and contextual information, offering a nuanced understanding of data that improves translation, particularly in domain-specific contexts. Despite challenges related to scalability and consistency, KGs significantly contribute to the effectiveness of NMT systems. Wordnets, as described by (Fellbaum, 2000),

serve as comprehensive lexical databases that organize concepts hierarchically, facilitating efficient storage and retrieval of linguistic information. The IndoWordnet, discussed by (Bhattacharyya, 2010), extends this concept to Indian languages, supporting cross-lingual information retrieval and machine translation.

Knowledge infusion techniques represent significant advancements in NMT by integrating external knowledge to improve both translation quality and contextual relevance. Retrieval Augmented Generation (RAG) (Guu et al., 2020; Lewis et al., 2020) combines information retrieval with generation, allowing NMT models to leverage retrieved documents for better context. REALM (Guu et al., 2020) enhances this by introducing a masked language pre-training step, integrating external knowledge during both pre-training and fine-tuning. RETRO further improves translation by using a KG-based approach to generate relevant textual explanations, enhancing interpretability and coherence. Additionally, synthetic data generation, as explored by (Lewis et al., 2019), augments training datasets with diverse examples, improving model performance, particularly in domain-specific tasks. (Siriwardhana et al., 2022) also advances RAG models in open-domain question answering through domain adaptation, using Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) to retrieve relevant passages for accurate answer generation.

Knowledge infusion with machine translation has been further advanced with methods like k-nearest-neighbour machine translation (kNN-MT) (Khandelwal et al., 2020), which enhances NMT by integrating nearest neighbour retrieval without additional training. This method improves translation across various domains. Similarly, (Cai et al., 2021) introduces a monolingual translation memory (TM) approach, particularly effective in low-resource or domain adaptation scenarios, where the system retrieves relevant sentences to enhance translation accuracy. (Zhang et al., 2021) proposes the PDC framework, integrating bilingual dictionaries into NMT to improve translation accuracy, especially for rare words, by combining components like the pointer, disambiguator, and copier. This framework shows significant improvements over traditional NMT models, particularly in handling rare vocabulary. Approaches such as those by (Bulté and Tezcan, 2019), (He et al., 2021), and (Hoang et al., 2022) further enhance translation by using fuzzy matching to retrieve similar documents, with improvements in how source and target information are encoded and interacted with during the translation process.

3 Methodology

In this section, we describe the proposed framework, highlighting the constituent components: adaptive changes to the RAG architecture for the task of MT, introduced auxiliary task, and the training objective.

3.1 Problem Formulation

Given an input sentence S in the source language, in tokenized form, $S = (s_1, s_2, \ldots, s_m)$, the problem of retrieval-augmented machine translation can be formulated as finding the target sentence, $\hat{T}$ in tokenized form, $\hat{T} = (t_1, t_2, \ldots, t_n)$, as given by equation (1),

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

where D is the set of retrieved documents from the knowledge base. Knowledge base is a generic term denoting various structures, including KG triples, textual documents, and even precomputed embeddings.

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

The machine translation output is obtained using the Generator, with the loss function, L_G, as given in Equation (2).

3.2 Overview of RAGMT

The RAGMT framework (illustrated in Figure 1) comprises four main components:

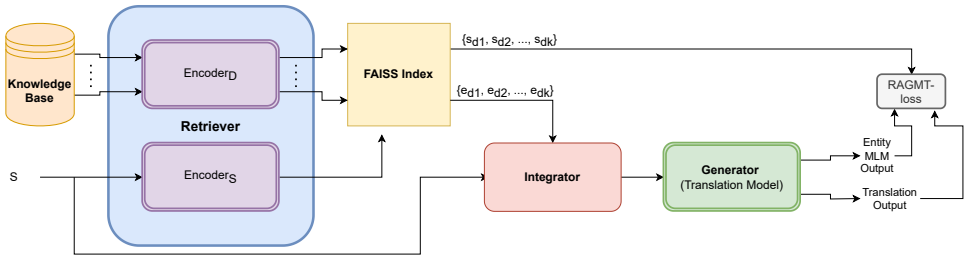

Figure 1: **RAGMT Architecture**: The KB consists of documents to be retrieved, which are indexed using FAISS over the embeddings computed using Encoder$_D$. For a source sentence, S, The retriever first encodes S using Encoder$_S$, then retrieves documents using the FAISS index. The retrieved documents,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

Knowledge Base, **Retriever**, **Integrator**, and **Generator**. The Knowledge Base can consist of structured information, such as KGs, Wordnets, etc., or unstructured documents. The Retriever consists of the document encoder, Encoder_D and the source encoder, Encoder_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 documents, along with the source sentence, are taken as input by the **Integrator** component, which allows various operations to be performed using the two, preparing inputs for the generator, such as simple concatenation over text, operations with the encodings, etc. The **Generator** finally performs the downstream task of machine translation, along with the auxiliary task of entity MLM.

3.3 Adapting RAG for NMT

We adopt the RAG framework for NMT by configuring the RAG architecture to support end-to-end training of all three parametric components: the **Document (Context) encoder**, the **Source encoder**, and **the Generator**. Integrating these components into a unified architecture enables the model to seamlessly incorporate context retrieval and translation generation, enhancing its ability to leverage external knowledge for translation tasks.

3.4 Auxiliary Task: Entity Masked Masked Language Modelling

We introduce an auxiliary task derived from entity Masked Masked Language Modelling (MLM) (Song et al., 2019) 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 $D = R(S) = \text{topk}(P_\eta(.|x)) = \{d_1, \ldots, d_k\}$ 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).

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

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.

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

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

The entity MLM loss, with its entity reconstruction objective, 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.

3.5 Final Training Objective

The final training objective of RAGMT can be written as follows:

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

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

where L_G is the generator model's loss and L_{MLM} is the entity MLM loss. L_D is a document similarity-based loss computed using the similarity between the source sentence (s) and every document d_i in the set of retrieved documents D. These similarity denoted by s_{d_i} is simply computed as a dot product between embeddings for s and d_i, via Encoder_S and Encoder_D respectively.

4 Experiments

4.1 Dataset

We utilized the English and German parallel corpus introduced by (Koehn and Knowles, 2017) at the First Workshop on Neural Machine Translation (WNMT) in 2017 and re-split by (Aharoni and Goldberg, 2020). This dataset provides a diverse range of text samples across different domains, allowing us to evaluate the adaptability of our models across various domains. The dataset consists of data from 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.

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 4.3.

Hence, a cohesive KG could not be extracted from the data. As specified by (Aharoni and Goldberg, 2020), we use 2000 validation and 2000 test points for each domain. We perform all the experiments with 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 not sufficient to run experiments using the entire training datasets. In this constrained setting, we have carefully compared against existing baseline systems as detailed below.

4.2 Experimental Setup

We conduct a series of experiments to evaluate the performance of the different translation models for domain adaptation. Each experiment involved fine-tuning and testing the models on domain-specific datasets, intending to assess their ability to adapt to different domains and leverage domain-specific knowledge for translation. We conduct the following experiments:

1. *Domain-adaptation of NMT models, using in-domain KG.* We build a KG for each domain as described in section 4.3, using the complete training subset of the data. Table 1 shows the training data size and corresponding KG size in terms of number of triples. [2]

[2]The subset of dataset used for training, and extracted in-domain KGs will be released upon publication.

2. *Domain-agnostic KG vs. In-domain KG.* We compare the impact of using an in-domain KG (built from the data of the same distribution as the training data) over a domain-agnostic KG. We use ConceptNet (Speer et al., 2016) as the domain-agnostic KG.

3. *Ablations on **RAGMT**.* To analyze the impact of the individual loss terms of the **RAGMT** training objective, we perform an ablation study involving L_{MLM} (the entity MLM loss) and L_D (the document similarity-based loss) that are newly added along with the generation loss.

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), and BERTScore (Zhang et al., 2019). BLEU (Bilingual Evaluation Understudy) is widely used to measure n-gram precision and is often considered a standard for machine translation evaluation. chrF++ provides an alternative by focusing on character n-grams, offering better sensitivity to small translation differences, especially in morphologically rich languages. TER (Translation Edit Rate) measures the number of edits needed to change a hypothesis translation into one of the references, emphasizing error correction. Finally, BERTScore leverages the semantic representations from BERT to evaluate translations based on contextual embeddings, capturing nuanced meaning and context beyond surface-level similarity. These metrics offer a well-rounded evaluation framework to assess translation quality from multiple perspectives.

The main systems we compare in our experiments are:

1. **Baseline MT.** We consider NMT model, without any external memory augmentation applied, as our baseline MT model. For this purpose (NLLB Team et al., 2022) is used.

2. **RAT-SI.** We set up RAT-SI as described in (Hoang et al., 2022). The setup uses fuzzy-matching to retrieve relevant documents and uses the training samples as the knowledge base, as opposed to our use of KGs. We always use (NLLB Team et al., 2022) as the generator to maintain a common starting point for each approach.

3. **RAGMT.** Our proposed framework that uses a KG extracted from the training set of each domain. The setup consists of document and source encoders, for which we use Dense Passage Retrieval detailed in (Karpukhin et al., 2020), and a generator, for which we use NLLB (NLLB Team et al., 2022).

4.3 KG Extraction

To facilitate domain-specific knowledge integration, we extracted datastores from the training data for each domain using REBEL (Resource Extraction from BERT Embeddings for Linked data) (Cabot and Navigli, 2021). REBEL enables the extraction of domain-specific knowledge using pre-trained BERT embeddings, allowing us to enrich our translation models with relevant domain information.

4.4 Implementation Details

All the systems described in section 4.2 use the pre-trained 600M parameter checkpoint of NLLB-200 (NLLB Team et al., 2022) as the generator. For the RAGMT setup, Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) is used as the encoders in the Retrieval module. 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*.

For the RAGMT setup, we use FAISS (Johnson et al., 2019) to index the knowledge base encoding, for faster training and evaluation time retrieval. For all our experiments, we extract the top 5 documents from any of the knowledge bases, for a consistent comparison.

5 Results and Analysis

Domain adaptation of MT. We first test if fine-tuning the proposed framework using a

Model Name	BLEU	chrf++	TER	BERTScore
Law Domain				
(1) Baseline MT	35.45	60.07	47.51	**0.84**
(2) RAT-SI	35.82	58.42	45.83	0.83
(3) RAGMT	**37.42**	**62.12**	**43.79**	0.82
Medical Domain				
(1) Baseline MT	36.6	57.26	42.33	0.78
(2) RAT-SI	37.61	**60.28**	43.14	**0.84**
(3) RAGMT	**39.12**	59.12	**41.55**	0.83
Koran Domain				
(1) Baseline MT	20.85	43.02	63.21	0.74
(2) RAT-SI	21.01	43.38	63.35	0.74
(3) RAGMT	**22.34**	**44.37**	**61.78**	**0.76**
IT Domain				
(1) Baseline MT	27.77	48.69	54.64	0.79
(2) RAT-SI	28.45	**52.61**	53.84	0.8
(3) RAGMT	**29.94**	49.12	**52.3**	**0.81**

Table 2: Domain adaptation performance of different experimental setups. Details about each setup are described in section 4.2.

Model Name	BLEU	chrf++	TER	BERTScore
Law Domain				
(1) Baseline MT	35.45	60.07	47.51	**0.84**
(2) ConceptNet	36.23	61.73	45.18	0.81
(3) Domain-specific KG	**37.42**	**62.12**	**43.79**	0.82
Medical Domain				
(1) Baseline MT	36.6	57.26	42.33	0.78
(2) ConceptNet	38.82	58.61	42.15	0.79
(3) Domain-specific KG	**39.12**	**59.12**	**41.55**	**0.83**
Koran Domain				
(1) Baseline MT	20.85	43.02	63.21	0.74
(2) ConceptNet	**22.56**	**45.94**	62.22	**0.79**
(3) Domain-specific KG	22.34	44.37	**61.78**	0.76
IT Domain				
(1) Baseline MT	27.77	48.69	54.64	0.79
(2) ConceptNet	28.71	48.92	53.37	0.78
(3) Domain-specific KG	**29.94**	**49.12**	**52.3**	**0.81**

Table 3: Comparison of domain-agnostic vs domain-specific knowledge graph with RAGMT across various domains.

domain-specific dataset, with a retrieval mechanism applied over an in-domain KG, would improve performance over the baseline approaches. Table 2 shows the performance of all the compared approaches on the domain adaptation experiment. Compared to the Baseline MT, RAGMT improves performance by an average of 2.03 BLEU scores, with the largest improvement on the Medical domain data with 2.52 BLEU score improvement. This signifies that fine-tuning a generator 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 RAT-SI approach, we observe an average improvement of 1.48 BLEU scores, with the largest improvement of 1.6 BLEU scores on the Law domain dataset. This improvement can be attributed to the adaptive retrieval mechanism employed by the RAGMT framework along with the **document similarity term** incorporated in the RAGMT training objective, compared to the fuzzy-matching based retrieval used by RAT-SI.

Domain-specific KG vs Domain-agnostic KG. The domain-specific KG has been extracted from the training subset of the domain-specific data. At the same time, we use ConceptNet (Speer et al., 2016) as our domain-agnostic KG. Table 3 shows the difference in performance of the RAGMT framework 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.

Ablations on the RAGMT training objective. We analyze the contribution of each of the constituent components of the training objective as described in section 3.5. We compare the performance of RAGMT framework under the following settings: (1) **RAGMT -** $L_{\textbf{RAG-Seq}}$, i.e., using the RAG-Sequence loss proposed by (Lewis et al., 2020); (2) **RAGMT - w/o** $L_{\textbf{MLM}}$, the RAGMT training objective without the loss from the Entity MLM component; (3) **RAGMT - w/o** L_D, the RAGMT objective without the explicit Document Similarity component; and (4) **RAGMT**, the training objective as described in section 3.5.

Table 4 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 RAG Sequence Loss. The obtained results signify that the *Document Similarity* component substantially contributes to the training objective with an average difference of

Domain	RAGMT - $L_{\text{RAG-Seq}}$	RAGMT - w/o L_{MLM}	RAGMT - w/o L_D	RAGMT
Law	34.67	37.17	36.52	**37.42**
Medical	34.96	39.02	38.94	**39.12**
Koran	21.54	21.98	20.64	**22.34**
IT	26.11	29.68	28.21	**29.94**

Table 4: Ablation on the RAGMT training objective. The BLEU scores obtained across all the domains, using different settings described in section 5.

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 5 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.

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 5: 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.

6 Conclusion

We present a Retrieval Augmented Generation (RAG) based multi-task MT framework to enhance machine translation using non-parametric knowledge bases. We show the efficacy of our new framework, compared to existing baselines, on the problem of domain-specific MT using knowledge graphs as the non-parametric knowledge base. Our approach improves the performance of the baseline MT model using both domain-agnostic as well as domain-specific knowledge graphs across all domains. For future work, we aim to focus on using 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, such as WordNet, tabular data, etc.

7 Limitations

- Due to resource limitations, we conduct experiments with a limited subset of the training data. Although the experiments demonstrate the efficacy of the proposed framework, fine-tuning using the complete training dataset would potentially offer more improvements over the baseline.
- 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 is a balance that 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*.

Pushpak Bhattacharyya. 2010. Indowordnet. In *International Conference on Language Resources and Evaluation*.

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.

Christiane Fellbaum. 2000. Wordnet: an electronic lexical database. *Language*, 76:706.

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.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *ArXiv*, abs/1511.06709.

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.

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*.

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

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.