# MULTI-DOMAIN NEURAL MACHINE TRANSLATION WITH KNOWLEDGE DISTILLATION FOR LOW RESOURCE LANGUAGES

Velayuthan Menan

238043D

Master of Science (Major Compoment Research)

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2025

# MULTI-DOMAIN NEURAL MACHINE TRANSLATION WITH KNOWLEDGE DISTILLATION FOR LOW RESOURCE LANGUAGES

Velayuthan Menan

238043D

Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science (Major Compoment Research)

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2025

## DECLARATION

I declare that this is my own work and this Dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 27/05/2025

The supervisors should certify the Dissertation with the following declaration.

The above candidate has carried out research for the Master of Science (Major Compoment Research) Dissertation under our supervision. We confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Nisansa de Silva

Signature of the Supervisor: Date: 27/05/2025

Name of Supervisor: Dr. Surangika Ranathunga

Signature of the Supervisor: Date:

## DEDICATION

To my dearest parents, Velayuthan (Appa) and Kanageshwary (Amma),
and to my fiancée, Sambavi,

Thank you for your unconditional love, unwavering support, and endless belief in me.
This journey would not have been possible without you.
This work is dedicated to you.

## ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Nisansa de Silva and Dr. Surangika Ranathunga, for their continuous support and guidance throughout the course of this research. Without your insights and tremendous mentorship, I would not have been able to achieve this level of success. Thank you for always stepping in to help whenever I needed, despite your busy schedules.

I also wish to extend my heartfelt appreciation to the Research Coordinator of the Department of Computer Science & Engineering, Dr. Kutila Gunasekera, for his unwavering support in ensuring this was a smooth and enriching experience. My sincere thanks also go to the academic and non-academic staff of the Department of Computer Science & Engineering for their assistance and for providing the resources necessary to carry out my research.

This work was funded by the Google Award for Inclusion Research (AIR) 2022, received by Dr. Surangika Ranathunga and Dr. Nisansa de Silva. I would like to thank the National Languages Processing (NLP) Centre at the University of Moratuwa for providing the GPUs used to execute the experiments related to this research.

# ABSTRACT

Multi-domain adaptation in Neural Machine Translation (NMT) is crucial for ensuring high-quality translations across diverse domains. Traditional fine-tuning approaches, while effective, become impractical as the number of domains increases, leading to high computational costs, space complexity, and catastrophic forgetting. Knowledge Distillation (KD) offers a scalable alternative by training a compact model using distilled data from a larger teacher model. However, we hypothesize that sequence-level KD primarily distills the decoder while neglecting encoder knowledge transfer, resulting in suboptimal adaptation and generalization, particularly in low-resource language settings where both data and computational resources are constrained.

To address this, we propose an improved sequence-level distillation framework enhanced with encoder alignment using cosine similarity-based loss. Our approach ensures that the student model captures both encoder and decoder knowledge, mitigating the limitations of conventional KD. We evaluate our method on multi-domain German–English translation under simulated low-resource conditions and further extend the evaluation to a bona fide low-resource language, demonstrating the method's robustness across diverse data conditions.

Results demonstrate that our proposed encoder-aligned student model can even outperform its larger teacher models, achieving strong generalization across domains. Additionally, our method enables efficient domain adaptation when fine-tuned on new domains, surpassing existing KD-based approaches. These findings establish encoder alignment as a crucial component for effective knowledge transfer in multi-domain NMT, with significant implications for scalable and resource-efficient domain adaptation.



**Keywords**: Neural Machine Translation, Knowledge Distillation, Multi-Domain Adaptation, Low-Resource Languages, Encoder Alignment, Sequence-Level Distillation

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| Abbreviation | Description |
|---|---|
| EWC | Elastic Weight Consolidation |
| GRU | Gated Recurrent Unit |
| KD | Knowledge Distillation |
| LRL | Low Resource Languages |
| LSTM | Long Short-Term Memory |
| M | Million |
| MT | Machine Translation |
| NLP | Natural Language Processing |
| NMT | Neural Machine Translation |
| RNMT | Recurrent Networks for Neural Machine Translation |
| RNN | Recurrent Neural Networks |
| SMT | Statistical Machine Translation |

# CHAPTER 1

# INTRODUCTION

*"Nothing in life is to be feared; it is only to be understood."*

– Marie Curie

## 1.1 Background and Motivation

Globalization continues to drive the demand for cross-lingual communication; to this end, Machine Translation (MT) has emerged as a key enabling technology, facilitating effective communication and collaboration across linguistic barriers. MT systems have undergone a technological renaissance in the past decade, marked by a shift from Statistical Machine Translation (SMT) [1, 2] to Neural Machine Translation (NMT) [3, 4]. This transition can be attributed to two key innovations in NMT: the introduction of the attention mechanism for encoder-decoder models [5], and the development of a new model architecture, Transformers, with attention mechanisms at its core [6].

Although NMT has significantly outperformed SMT in machine translation tasks [7], it is typically data-hungry and tends to perform poorly under data-poor conditions [8]. This limitation poses a substantial challenge for Low Resource Languages (LRL) . We adopt the threshold proposed by Ranathunga et al. [9], who classify a language as low resource if the available parallel data comprises fewer than 0.5 Million (M) sentence pairs. Given the severity and prevalence of this problem, addressing NMT for LRL has become an extensively studied research direction within MT [3, 10, 11]. This challenge becomes even more pronounced when domain adaptation is required for NMT systems trained on LRL.

Domain-specific NMT systems are in high demand because general-purpose NMT systems often have limited applicability in specialized contexts [12, 13]. This limitation is primarily due to significant lexical and stylistic differences across domains. For instance, an NMT system trained to perform very well in the legal domain might fail to maintain translation quality when translating biomedical documents. Such domain-specific challenges become even more severe in low-resource scenarios, where limited parallel data further constrains the performance of NMT systems [14].

Domain adaptation in NMT has been extensively studied [13], but significant gaps remain, particularly for low-resource languages [3]. A common approach involves fine-tuning general-domain pre-trained models on specific target domains [15]. While effective, this approach becomes cumbersome as the number of domains increases, requiring separate models for each domain. This not only increases the space complexity

of the Machine Translation (MT) system but also incurs higher maintenance and system costs. Moreover, it fails to exploit the shared information inherent across domains, limiting its efficiency [16].

To address these challenges, researchers have increasingly focused on developing a single model capable of handling multiple domains [16–19]. One promising direction is Knowledge Distillation (KD) [20], where domain-specific teacher models (expert models), typically deep neural networks, generate distilled target-side data for each domain. Currey et al. [16] demonstrated that a student model, usually a shallower neural network, can be trained on a mixture of the original and distilled data, effectively capturing domain knowledge in a compact form. This data-centric approach not only reduces the need for maintaining multiple models but also leverages domain similarities, making it both effective and easy to adopt. However, KD-based domain adaptation methods face limitations in low-resource languages, as training effective domain-specific teacher models becomes increasingly challenging due to insufficient data.

Existing KD approaches predominantly rely on sequence-level distillation [21], emphasizing decoder knowledge transfer while often neglecting the encoder component. Consequently, the limited transfer of encoder knowledge may hinder the ability of student models to effectively capture domain-specific nuances, crucial for accurate translation across domains. This motivates the need for improved techniques that address encoder-level knowledge transfer to enhance adaptability and generalization in low-resource, multi-domain scenarios.

## 1.2 Research Problem

Designing multi-domain NMT systems that perform robustly across multiple domains remains challenging, particularly in low-resource settings characterized by both limited training data and constrained computational resources [3, 15]. Existing approaches typically involve fine-tuning separate domain-specific models, which quickly becomes impractical and inefficient as the number of domains grows, due to increased computational overhead and storage demands [15]. Although KD provides a promising direction by leveraging domain similarities to train compact student models [16, 20], its effectiveness under low-resource and compute-poor scenarios is less understood, particularly concerning the adequate transfer of encoder-level knowledge.

Thus, the core research problem addressed in this thesis is: *How can we design a multi-domain NMT system for low-resource languages that scales efficiently to new domains while maintaining minimal or no degradation in performance across existing domains, within the practical constraints of limited data and compute-poor environments?*

This research is motivated by the need to overcome limitations of existing knowl-

edge distillation methods in such settings, where ensuring both adaptability and generalization across multiple domains remains a significant challenge, particularly when encoder knowledge transfer is insufficiently addressed.

## 1.3 Research Objectives

The primary objectives of this research are as follows:

1. **To investigate the limitations of sequence-level KD in NMT, particularly its impact on encoder knowledge transfer in low-resource settings.**
   Traditional sequence-level KD has been widely used for model compression and domain adaptation in NMT. However, prior studies primarily focus on its effectiveness in improving the decoder, with little emphasis on its influence on encoder representations. In low-resource settings, where data scarcity amplifies knowledge transfer challenges, it is crucial to assess whether existing distillation methods efficiently capture and transfer domain-invariant encoder representations. This objective seeks to empirically analyze the extent to which sequence-level KD primarily benefits the decoder, while neglecting encoder alignment, leading to suboptimal generalization across domains.

2. **To design and assess an encoder-alignment strategy to enhance knowledge transfer between teacher and student models for improved multi-domain generalization in NMT.**

   Given the limitations of conventional KD approaches, this research proposes an encoder-aware KD method that explicitly aligns the encoder representations between teacher and student models. The proposed method will be evaluated against traditional sequence-level KD to assess its effectiveness in improving encoder representation learning and its impact on both in-domain and out-of-domain translation performance.

3. **To assess the effectiveness of the proposed encoder-aligned distillation approach in both in-domain and out-of-domain scenarios, demonstrating its applicability in real-world low-resource and compute-constrained environments.**
   The practicality of any domain adaptation method depends not only on theoretical improvements but also on its real-world applicability. This objective focuses on systematically evaluating the proposed approach using a realistic low-resource setup, where both parallel data and computational resources are constrained. The research will measure performance gains in in-domain adaptation as well as generalization to unseen domains, validating whether the proposed

method can serve as a scalable, resource-efficient solution for multi-domain NMT.

## 1.4 Contributions

The main contributions of this thesis are as follows:

1. We empirically investigate the hypothesis that sequence-level KD primarily transfers decoder knowledge, and find that it provides limited support for encoder knowledge transfer, particularly in low-resource settings.
2. We propose an encoder-alignment method that uses cosine similarity between teacher and student encoder representations to encourage the transfer of domain-invariant knowledge and improve generalization.
3. We show that our encoder-aligned distillation method performs consistently better than standard sequence-level KD and other data-centric baselines, across both in-domain and out-of-domain evaluations.
4. We evaluate our method in a practical low-resource scenario with limited parallel data (50K sentences) and constrained computational resources, demonstrating its applicability in compute-poor environments.
5. We extend our evaluation to a bona fide low-resource language, English–Sinhala, and find that encoder alignment continues to be beneficial. Our results also suggest that the optimal strength of alignment is dependent on the dataset.

## 1.5 Publications

- Surangika Ranathunga, Nisansa de Silva, **Menan Velayuthan**, Aloka Fernando, Charitha Rathnayake
  - "Quality Does Matter: A Detailed Look at the Quality and Utility of Web-Mined Parallel Corpora"
  - EACL 2024 – The 18th Conference of the European Chapter of the Association for Computational Linguistics, Malta, March 17-22, 2024 (h5-index: 56)
  - Published and available online: link
  - Award: Best Low Resource Paper Award

- **Menan Velayuthan**, Dilith Jayakody, Nisansa de Silva, Aloka Fernando, Surangika Ranathunga
  - "Back to the Stats: Rescuing Low Resource Neural Machine Translation with Statistical Methods"
  - WMT 2024 – Ninth Conference on Machine Translation, Miami, Florida, USA, November 15-16, 2024 (h5-index: 45)
  - Published and available online: link
- **Menan Velayuthan**, Surangika Ranathunga, Nisansa de Silva
  - "Encoder-Aware Sequence-Level Knowledge Distillation for Low-Resource Neural Machine Translation"
  - Accepted at LoResMT 2025 – The Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages at NAACL 2025

## 1.6 Other Contributions

- Successfully conducted **"Sequence to Sequence Learning: Hands-on Tutorial"** at MERCon 2023, together with Dr. Uthaya, Dr. Nisansa, and Sritharan Braveenan.
  **[** GitHub **]**
- Successfully conducted the following sessions for Dr. Nisansa's research group:
  - Distilling the Knowledge in a Neural Network **[** YouTube **|** GitHub **]**
  - Graph Neural Networks - Part 1 **[** YouTube **|** GitHub **]**
  - Understanding Low-Rank Adapters **[** YouTube **|** GitHub **]**
- Conducted a TED Talk at IESL on "How To Act Data Rich In A Data Poor Country Borrow from Your Rich neighbor?" discussing the utility of Transfer learning and Finetuning in AI in data poor scenario.

## 1.7 Thesis Structure

This thesis is organized into six chapters, each addressing different aspects of our research on KD for low-resource multi-domain NMT. Below is a brief overview of each chapter:

- **Chapter 1: Introduction**
  This chapter introduces the motivation and background of the research, highlighting the challenges of multi-domain adaptation in NMT, particularly in low-resource and compute-constrained settings. We define the research problem, outline the primary objectives, and summarize the key contributions of this thesis. Additionally, we present our relevant publications.

- **Chapter 2: Literature Review**
  This chapter provides a comprehensive review of prior work in NMT, domain adaptation, and KD. We examine the evolution of NMT architectures, discuss key domain adaptation techniques, and analyze the strengths and limitations of KD in machine translation. The chapter concludes by identifying critical gaps in existing research, motivating our proposed approach.

- **Chapter 3: Methodology**
  This chapter details the proposed encoder-aligned sequence-level distillation method. We first present the hypothesis that standard sequence-level distillation predominantly transfers decoder knowledge, leading to suboptimal encoder representations. To address this, we introduce a novel encoder alignment strategy using cosine similarity. We provide a step-by-step explanation of our approach, including the training process, loss formulation, and justification for key design choices.

- **Chapter 4: Experimental Setup**
  This chapter outlines the experimental framework used to evaluate our proposed method. We describe the hardware and software configurations, the datasets used, and the training and inference setups. We also detail the specific model configurations and hyperparameters employed in our experiments. This chapter ensures transparency and reproducibility of our research.

- **Chapter 5: Results and Discussion**
  This chapter presents and analyzes the experimental results. We begin by discussing the limitations of our initial contrastive encoder alignment approach, which used multiple domain-specific teacher models. Based on these insights, we motivate the transition to our final encoder-aligned distillation method. We then evaluate different values of the encoder alignment factor ($\alpha$) to determine the optimal setting. The chapter is organized into two main stages: (i) multi-domain translation and generalization to unseen domains, and (ii) domain adaptation through fine-tuning on new domains. Finally, we extend our evaluation to a real low-resource language (English–Sinhala) to study the robustness and applicability of our method in more challenging data-scarce conditions.

- **Chapter 6: Conclusion**
  This chapter summarizes the key contributions of our work and discusses the broader implications of our findings. We highlight how our method enhances domain adaptation in low-resource NMT and propose potential directions for future research, including scalability to more domains, alternative alignment techniques, and application to other language pairs.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Overview

NMT has evolved significantly with Transformer-based architectures replacing earlier Recurrent Neural Network-based models due to their superior efficiency and ability to model long-range dependencies. However, domain adaptation in NMT remains a challenge, particularly for low-resource languages, where data scarcity and computational constraints hinder performance. Traditional fine-tuning approaches struggle with scalability and catastrophic forgetting, while multi-domain adaptation techniques attempt to mitigate these issues [22, 23]. KD, particularly sequence-level distillation, has emerged as a promising method for transferring knowledge from large models to smaller, more efficient ones. While KD has shown effectiveness in multi-domain NMT, existing approaches primarily focus on high-resource settings and remain largely data-centric, neglecting encoder knowledge transfer. This chapter explores key NMT architectures, domain adaptation strategies, and KD techniques, identifying gaps that motivate our proposed enhancements to sequence-level distillation for low-resource and compute-constrained scenarios.

## 2.2 Neural Machine Translation (NMT)

### 2.2.1 An Overview of NMT

NMT is a sequence-to-sequence modeling task where the objective is to translate a sentence from a source language into a target language using neural network architectures [7, 24, 25]. Formally, given a source-language sentence $X = (x_1, x_2, \ldots, x_{T_x})$, the goal of NMT is to produce a corresponding target-language sentence $Y = (y_1, y_2, \ldots, y_{T_y})$ by modeling the conditional probability:

$$P(Y|X;\theta) = \prod_{t=1}^{T_y} P(y_t|y_{<t}, X;\theta), \tag{2.1}$$

where $\theta$ represents the trainable parameters of the NMT model.

The earliest successful NMT systems employed RNNs, particularly LSTM or GRU, within an encoder-decoder framework [7, 26]. The encoder processes the source sentence into a hidden representation $h$:

$$h_t = f_{enc}(x_t, h_{t-1}), \tag{2.2}$$

where $f_{enc}$ is the encoder function (typically an RNN cell). The decoder then

generates the target sentence word-by-word, conditioned on previous target words and the encoder representation:

$$s_t = f_{dec}(y_{t-1}, s_{t-1}, c_t), \tag{2.3}$$

where $f_{dec}$ is the decoder function, and $c_t$ is a context vector derived from encoder states, typically using an attention mechanism:

$$c_t = \sum_{i=1}^{T_x} \alpha_{t,i} h_i, \tag{2.4}$$

where attention weights $\alpha_{t,i}$ are computed as:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{T_x} \exp(e_{t,j})}, \tag{2.5}$$

with the alignment scores $e_{t,i}$ often computed using feed-forward neural networks [24].

Despite early successes, RNN-based NMT faced challenges in modeling long-range dependencies and suffered from sequential computation bottlenecks. These issues were addressed by the Transformer architecture, which replaced recurrent computations with multi-head self-attention mechanisms [25]. The self-attention mechanism computes the attention between queries ($Q$), keys ($K$), and values ($V$):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{2.6}$$

where $d_k$ is the dimensionality of the key vectors. Multi-head attention allows the model to jointly attend to information from different representation subspaces.

### 2.2.2 Common Architectures in NMT

**RNN-based NMT Architectures:** Early NMT systems predominantly used Recurrent Neural Networks (RNN) [27] as the backbone for sequence-to-sequence modeling. In these architectures, an encoder RNN, typically a Long Short-Term Memory (LSTM) [28] or a Gated Recurrent Unit (GRU) [29], processes the source sentence word-by-word into a fixed-length vector, and a decoder RNN generates the target sentence sequentially [30]. The introduction of the attention mechanism [24] significantly improved this model by allowing the decoder to dynamically focus on encoder hidden states at each decoding step, alleviating the bottleneck of compressing all information into a single vector [31]. This attention-based encoder–decoder framework, known as Recurrent Networks for Neural Machine Translation (RNMT) , quickly became the de-facto standard for NMT and was adopted in large-scale commercial systems such as Google's GNMT due to its superior translation quality. However, RNN-based approaches faced inherent limitations, particularly regarding parallelization and the mod-

eling of long-range dependencies. By the late 2010s, these models were progressively replaced by convolutional and self-attention-based Transformer architectures [25, 31].

**Transformer-based NMT Architectures:** Transformers revolutionized machine translation by replacing recurrent computations entirely with self-attention mechanisms, enabling parallel computation and effectively capturing long-range dependencies. The Transformer model consists of stacked encoder-decoder layers, each employing multi-head self-attention and position-wise feed-forward networks [25]. This design removes the sequential bottlenecks inherent in RNN-based models, resulting in faster training times and improved contextual representation capabilities [32]. Unlike RNNs, where information propagates sequentially, self-attention allows direct interactions among all tokens in a sequence, significantly reducing the effective distance between words [33]. Additionally, Transformers utilize positional encodings to embed word-order information without relying on recurrence [34]. Recent advancements have optimized the Transformer architecture by dynamically pruning redundant attention heads, thereby improving computational efficiency without compromising translation performance [35]. Due to their scalability and superior effectiveness, Transformers have emerged as the dominant architecture in modern NMT, substantially outperforming previous RNN-based approaches [36].

**Pretrained Models for NMT:** Recent advancements in NMT have been largely driven by pretrained multilingual models, such as NLLB-200 [37], M2M-100 [38], mT5 [39], and mBART [40], which enable translation across a wide range of languages, including low-resource ones. NLLB-200, introduced to improve low-resource language translation, employs a conditional compute model with a sparse Mixture-of-Experts (MoE) architecture and extensive multilingual data mining, yielding superior performance across numerous translation directions [37]. Similarly, M2M-100 was developed to enable truly Many-to-Many translation without relying on English as an intermediary language, a deficiency in earlier approaches that often led to English-centric biases [38]. Meanwhile, mBART, a sequence-to-sequence denoising autoencoder, demonstrated significant improvements in multilingual translation through large-scale monolingual corpora and multilingual fine-tuning, particularly benefiting low-resource language adaptation [40]. mT5, a multilingual extension of the T5 text-to-text transfer transformer, exhibited robust performance across NLP benchmarks, including the ability to generalize to unseen languages, though challenges in typologically diverse low-resource contexts persist [39]. While pretrained models have significantly enhanced translation quality, they still face limitations such as multilingual interference, data scarcity-induced degradation, and representational imbalance, particularly for under-represented languages [41–43]. Recent approaches, including typology-aware transfer learning [44], domain adaptation [45], and lightweight model distillation [46], seek to

further refine low-resource language capabilities, addressing some of these persistent challenges.

## 2.3 Domain Adaptation for NMT

### 2.3.1 An Overview of Domain Adaptation

Catastrophic forgetting [22, 47], the degradation of model performance on previous tasks when continually trained on new tasks, remains a significant challenge in domain adaptation for NMT. A straightforward yet effective approach to address this issue was demonstrated by Chu et al. [22] and Liang et al. [48], where models are re-trained from scratch using data from all available domains. While this method can become impractical as the number of domains grows—since it requires all domain-specific data to be accessible simultaneously—it remains feasible in low-resource scenarios, where each domain typically has a relatively small dataset. However, adapting NMT systems to only a single domain is often inadequate for real-world applications requiring multi-domain capabilities. Maintaining separate expert systems for each domain quickly becomes prohibitively expensive and inefficient as the number of domains grows. Consequently, multi-domain NMT systems provide a more scalable and practical alternative [19, 49, 50].

Approaches to domain adaptation in NMT are broadly categorized into two main types: (i) data-centric methods, which primarily leverage modifications or augmentation of training data to enhance domain adaptability [16, 21, 51, 52]; and (ii) model-centric methods, which involve direct modifications to model architectures or adjustments to training processes [53–56]. In this thesis, we adopt a hybrid approach that combines elements of both strategies. Specifically, we build upon successful data-centric distillation methods introduced by Currey et al. [16], while introducing targeted modifications to the training procedure through our encoder-alignment strategy, effectively merging data-driven and model-centric paradigms.

### 2.3.2 Challenges in Domain Adaptation

#### 2.3.2.1 Data Scarcity and Domain Shifts

NMT systems excel when ample in-domain parallel data is available, but performance drops sharply for texts from a new domain with distinct vocabulary or style [13, 14]. In real-world scenarios, high-quality domain-specific parallel corpora are often scarce or nonexistent [15]. As a result, a model trained on one domain (e.g. news text) struggles to translate another domain (e.g. biomedical literature) accurately due to distributional differences (terminology, syntax, formality) between domains [14]. This domain shift problem means that words may take on different translations or meanings

in different contexts, and generic NMT models lack exposure to these domain-specific usages. The limited availability of in-domain bilingual data exacerbates the issue, since data scarcity hampers the model's ability to learn the target domain's language patterns. Researchers therefore leverage out-of-domain data and monolingual corpora to compensate [15]—for example, using large general parallel corpora as a foundation and synthetic in-domain data (via back-translation of monolingual texts [15] to inject domain-specific terminology). Despite such efforts, the fundamental challenge remains: insufficient in-domain examples and significant domain mismatches lead to poor generalization of NMT systems on new domains [13]. Overcoming data scarcity (often by data augmentation or transfer learning) and bridging the domain gap are central goals of domain adaptation techniques.

#### 2.3.2.2 Catastrophic Forgetting When Adapting Models

Adapting an NMT model to a new domain via fine-tuning can introduce the problem of catastrophic forgetting, where gains on the target domain come at the expense of dramatic losses in performance on the original domain [57]. In other words, the model "forgets" previously learned translation knowledge that is not reinforced by the new domain's small dataset. This phenomenon has been widely observed in NMT domain adaptation [58]. Fine-tuning on limited in-domain data tends to over-specialize the model to that domain's distribution, erasing some of the general-domain capability acquired during initial training. The result is a domain-adapted model that translates in-domain sentences well but performs poorly on out-of-domain sentences it used to handle correctly [57]. Such forgetting is especially problematic if the system is expected to retain broad coverage (for example, an online translator that must handle multiple topics). Mitigating catastrophic forgetting is an active research focus: approaches include regularization strategies (e.g. $L_2$ regularization or dropout during fine-tuning) to constrain parameter updates [15], and Elastic Weight Consolidation (EWC) [59, 60] or KD [20] methods to preserve the original model's knowledge [15, 57]. Another strategy is to mix a portion of general-domain data into the fine-tuning process (mixed fine-tuning) so that the model retains proficiency on the base domain while adapting [15]. Overall, preventing the model from forgetting earlier competencies when adapting to new domains remains a key challenge, closely tied to the trade-off between specialization and generalization in NMT.

#### 2.3.2.3 Effective Transfer Learning Techniques

Identifying effective transfer learning techniques for domain adaptation is challenging, as NMT models can be adapted in numerous ways and each comes with trade-offs. The simplest and most common approach is fine-tuning: taking a model trained on a large out-of-domain corpus and continuing training on the smaller in-domain cor-

pus [22]. Fine-tuning often yields quick gains on the target domain, but as noted, it can overfit to the new data or cause forgetting of the original domain. To address these issues, various enhanced techniques have been proposed. One line of work focuses on data-centric methods: for example, data selection or instance weighting where out-of-domain sentence pairs that are most relevant to the target domain are weighted or chosen to aid adaptation [15]. This ensures the model learns more from sentences similar to the in-domain style. Another approach is generating synthetic parallel data using in-domain monolingual text (through back-translation), effectively augmenting the scarce parallel data [61, 62]. On the model side, researchers explore multi-domain training and architectural tweaks. Mixed fine-tuning, which intermixes in-domain and out-of-domain data during training, has shown success in improving in-domain performance without sacrificing out-of-domain quality [22, 23]. There are also techniques like adding domain tags or domain-specific tokens to inputs so that a single model can handle multiple domains by conditioning on the domain label [15, 63, 64]. Adversarial training with a domain discriminator is another method: the NMT encoder is trained to produce representations that a domain classifier cannot easily distinguish (forcing the model to learn domain-invariant features) [65]. Furthermore, ensemble or multi-model methods can be used – e.g., combining a general model and an in-domain model, or using specialist models for each domain and switching between them. Each of these transfer learning strategies comes with challenges: instance weighting and adversarial methods add complexity to training, and multi-domain models may struggle to match a dedicated model's performance on any single domain. A survey of domain adaptation techniques indicates that no single method dominates; often, the best results come from a combination of techniques, such as mixing data augmentation with fine-tuning and regularization [13]. The challenge for practitioners is to choose an adaptation approach that effectively transfers knowledge from the general domain to the target domain while minimizing overfitting and maintaining as much of the original model's capability as possible.

### 2.3.3 Multi-domain Adaptation Strategies

Multi-domain NMT adaptation has been addressed with a range of strategies in recent years. One simple yet effective approach is to attach domain tags or special tokens to the input, allowing a single translation model to condition on the domain and thereby produce domain-appropriate outputs [66]. Other methods allocate domain-specific model components or parameters for each domain: for example, training with domain-specific adapter layers or separate encoder/decoder modules that capture domain-specific patterns while sharing common translation knowledge [67]. Some works also employ multi-task learning signals such as auxiliary domain classifiers or adversarial objectives to influence the encoder's representations, encouraging the

model to either distinguish between domains or learn domain-invariant features as needed [68]. In addition, data-centric techniques like intelligent data selection and curriculum learning have been explored: here the training data from different domains are weighted or scheduled (from generic to specific) so that the model gradually focuses on in-domain sentences, which has been shown to improve multi-domain performance without sacrificing low-resource domains [69]. More recently, meta-learning has been applied to NMT domain adaptation, where the idea is to "learn how to learn" new domains – the model is trained on simulated domain-shift tasks so that it can quickly adapt to an unseen target domain with only a small amount of in-domain data [70].

Each approach comes with distinct challenges and advantages when handling multiple domains. A fundamental issue is how to retain both general and domain-specific translation quality across domains without catastrophic forgetting or interference (adapting to one domain hurting performance on others) [69]. Domain tagging offers simplicity and a unified model; indeed, a single tag-conditioned NMT model can compare favourably to dedicated per-domain models in translation quality [71]. However, tag-based systems are inflexible in adding new domains and may not fully capture niche domain terminology or style without additional adaptation – each new domain typically requires retraining the entire model [72]. Approaches that introduce domain-specific parameters (e.g. adapters or specialized layers) help mitigate such trade-offs by preserving distinctive domain information within a shared model, often yielding higher accuracy on each domain. For instance, a word-level domain mixing model with per-domain modules outperformed prior multi-domain NMT systems on several benchmarks [67], and explicitly maximizing the mutual information between translations and domain labels was shown to further enhance domain specialization, pushing a multi-domain model to state-of-the-art results on all domains [68]. Meanwhile, meta-learning methods excel when in-domain data is scarce or domains are numerous: one study reported that a meta-trained NMT model significantly outperformed standard fine-tuning by about 2.5 BLEU with only a few hundred in-domain sentences [70]. Finally, dynamic curriculum learning and data weighting strategies can balance specialization and generalization, improving in-domain translation while maintaining strong performance on other domains. Empirical findings suggest that no single technique dominates universally; each method offers comparative benefits depending on the data availability, number of domains, and whether rapid extensibility to new domains is required.

## 2.4 Knowledge Distillation in NMT

### 2.4.1 An Overview of Knowledge Distillation

KD, first introduced by Hinton [20], is a training paradigm in which knowledge from a larger, more powerful *teacher* model is transferred to a smaller, compact *student* model. The main objective of KD is to compress knowledge into more efficient architectures, making deployment feasible under resource-constrained conditions. KD has demonstrated considerable success across various fields, notably computer vision and natural language processing [20, 21, 73, 74].

The fundamental principle behind KD is that the student model learns to mimic the predictive distribution of the teacher rather than relying exclusively on traditional ground-truth labels. Formally, given the teacher model's output distribution $p^T$ and the student model's distribution $p^S$, the distillation loss is typically defined via the Kullback-Leibler (KL) divergence as follows:

$$\mathcal{L}_{KD} = \sum_{i=1}^{|V|} p^T(y_i|x;\tau) \cdot \log \frac{p^T(y_i|x;\tau)}{p^S(y_i|x;\tau)}, \tag{2.7}$$

where $x$ is the input, $y_i$ is a token from the vocabulary $V$, and $\tau$ is a temperature parameter controlling the smoothness of probability distributions. Higher temperatures produce softer probability distributions, facilitating better transfer of knowledge about inter-class relationships [20].

Typically, KD employs a joint optimization objective combining the distillation loss with a standard supervised loss, such as cross-entropy:

$$\mathcal{L}_{total} = (1-\alpha)\mathcal{L}_{CE} + \alpha\mathcal{L}_{KD}, \tag{2.8}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss computed with respect to ground-truth labels, and $\alpha$ balances the two loss components.

KD approaches can be broadly classified into three categories based on the form of knowledge transferred: (i) response-based distillation, which leverages teacher predictions [20]; (ii) feature-based distillation, aligning intermediate representations between models [75]; and (iii) relation-based distillation, transferring relational information among model layers or inputs [76, 77].

### 2.4.2 Knowledge Distillation for NMT

Extending KD to sequence-to-sequence tasks, such as NMT, poses specific challenges due to the structured nature of sequence generation. To address this, Kim and Rush [21] introduced *sequence-level distillation*, an adaptation of KD tailored explicitly for sequence generation tasks [78]. Sequence-level distillation trains the student model

to directly mimic the teacher-generated sequences, thereby capturing richer contextual and structural knowledge essential for tasks such as translation.

Building upon this concept, Currey et al. [16] demonstrated that sequence-level distillation could effectively support multi-domain adaptation in NMT. Their results indicated that a single compact student model trained through KD could achieve superior performance compared to multiple domain-specific teacher models. Furthermore, Liang et al. [17] investigated sequence-level distillation within continual learning frameworks [79], progressively expanding a translation model's coverage across domains. They found that sequence-level distillation effectively mitigated catastrophic forgetting, enabling the model to retain knowledge from previously learned domains while adapting efficiently to new domains.

However, prior research has not extensively explored the application of sequence-level KD within resource-constrained settings, such as scenarios involving low-resource languages or limited computational resources. Addressing this gap, our work specifically targets the combined challenges of low-resource language data availability and computational constraints, proposing KD-based strategies designed explicitly for these demanding scenarios.

### 2.4.3 Sequence-Level Distillation

Sequence-level distillation extends the traditional KD framework to sequence-to-sequence tasks like NMT. Unlike standard KD, which transfers knowledge via class probabilities at the token level (response-based distillation), sequence-level distillation operates at the sequence level by training the student model directly on teacher-generated outputs. This approach, introduced by Kim and Rush [21], was designed to address challenges unique to structured prediction tasks.

Formally, given an input sentence $X = (x_1, x_2, \ldots, x_{T_x})$, a teacher model $T$ generates a translated sentence:

$$\hat{Y}^T = \arg\max_{Y'} P(Y'|X; \theta_T), \tag{2.9}$$

where $Y'$ is a candidate sequence, $\theta_T$ are the teacher model's parameters, and $\hat{Y}^T$ is the teacher-generated translation. The student model $S$, parameterized by $\theta_S$, is then trained directly on the pairs $(X, \hat{Y}^T)$ using standard supervised learning objectives, typically cross-entropy loss:

$$\mathcal{L}_{SEQ-KD} = -\sum_{t=1}^{T_y} \log P(y_t^T | y_{<t}^T, X; \theta_S), \tag{2.10}$$

where $y_t^T$ is the $t$-th token of the teacher-generated sequence $\hat{Y}^T$. This contrasts with token-level KD, which utilizes the soft target probabilities over the entire vocab-

ulary at each token.

The key distinction between sequence-level distillation and conventional KD is that the former captures sequence-level structural and contextual knowledge explicitly by utilizing the teacher-generated sequences as hard targets. This explicit sequence modeling approach better captures translation decisions, word ordering, and phrase-level coherence, which are essential for sequence-generation tasks like NMT.

However, this method is predominantly data-centric: it does not explicitly encourage the alignment of internal representations or intermediate features between teacher and student models. Instead, it enriches the dataset itself by augmenting original parallel data with teacher-generated translations, effectively creating a new, distilled training corpus. Thus, sequence-level distillation fundamentally differs from standard feature-based or relation-based distillation approaches, as it operates exclusively by altering the training data rather than the model architecture or training algorithm itself.

Due to its simplicity and effectiveness, sequence-level distillation has become a popular choice for multi-domain adaptation and continual learning in NMT, primarily because it easily leverages the teacher's domain expertise without extensive modification to the underlying student model architecture[16, 80, 81].

## 2.5 Embedding Aligment

### 2.5.1 Alignment Methods In Natural Language Processing (NLP)

Alignment methods in NLP span a range of distance or divergence measures, each with pros and cons. *Euclidean (L2) distance* is a straightforward choice (e.g. minimizing mean squared error between embedding vectors), but it is sensitive to vector magnitude and often underperforms without normalization [82]. *Cross-entropy alignment* treats alignment as a classification problem – for instance, selecting the correct target word or sentence among candidates – and has been used in contrastive frameworks to align representations from parallel data [83]. This can yield strong fine-grained alignment (as in matching translation pairs), but requires careful negative sampling and can be brittle if the alignment signal is noisy. *KL-divergence* is another option for aligning probability distributions (e.g. between an encoder's output distribution and a reference), commonly seen in teacher-student model distillation; however, it is asymmetric and can be ineffective on its own (directly mimicking a teacher's output via KL often yields only limited gains) [84]. More recently, *Wasserstein distance* (optimal transport) has attracted attention as it accounts for the geometry of the representation space. Wasserstein-based alignment can align distributions with minimal overlap and has shown high accuracy (even achieving state-of-the-art bilingual lexicon induction in one OT-based approach) [85]. The downside is computational cost: solving optimal transport for large vocabularies or high-dimensional features is expensive (scaling su-

perlinearly with data size) [85], making methods like Sinkhorn or quantization-based approximations necessary in practice. In summary, each alignment criterion balances different trade-offs – Euclidean vs. cosine focus on vector space geometry, cross-entropy and KL on probabilistic matching, Wasserstein on global distribution shape – and their effectiveness can vary by task and data characteristics.

### 2.5.2 Cosine-based Alignment Methods

Cosine-based alignment has emerged as a superior choice for aligning neural encoder representations, especially in bilingual and multilingual settings. Cosine similarity ignores vector length and emphasizes direction, which better captures semantic equivalence in high-dimensional spaces where embedding norms may vary [82]. Empirical studies in the last few years repeatedly show that maximizing cosine (or cosine-derived measures) between aligned language representations leads to stronger alignment and transfer performance than alternative metrics. For example, aligning multilingual BERT's contextual embeddings by enhancing cosine-similarity of translation pairs yields large gains in cross-lingual tasks – Cao et al. [86] report that after applying a cosine-based alignment procedure, zero-shot translation accuracy on XNLI for some languages matched a fully supervised baseline [86]. In NMT, fine-tuning encoder outputs to increase cosine similarity between true source–target word pairs has achieved state-of-the-art word alignment quality, outperforming traditional statistical aligners by substantial margins [87]. These results echo a general finding that cosine-oriented objectives produce better-aligned embeddings: replacing raw Euclidean loss with a cosine-based criterion (or its improvements like RCSLS) significantly boosts bilingual lexicon induction and cross-lingual embedding mapping accuracy [82]. In essence, cosine alignment provides a robust, scale-invariant gauge of similarity that aligns encoder spaces in a way that more faithfully reflects underlying semantic correspondences, explaining why it is now prevalent in encoder alignment for NMT and other NLP applications.

Cosine-based encoder alignment has emerged as a widely adopted method due to its clear advantages over alternative alignment techniques. Unlike Euclidean distance, which is sensitive to vector magnitude, cosine similarity effectively captures semantic relationships by focusing on directional alignment in high-dimensional spaces. Compared to cross-entropy and KL-divergence, which rely on probabilistic distributions and require careful sampling, cosine-based alignment offers a *scale-invariant*, *computationally efficient*, and *empirically validated* approach for aligning neural representations. Furthermore, recent studies demonstrate that maximizing cosine similarity between source and target embeddings in multilingual and NMT settings achieves *state-of-the-art alignment quality* and improved transferability, outperforming traditional statistical and probabilistic methods. These advantages have led to its preva-

lence in modern encoder alignment strategies, particularly within low-resource and multilingual NLP tasks.

## 2.6 Summary and Research Gaps

This chapter has reviewed key aspects of NMT, domain adaptation techniques, and KD, particularly sequence-level distillation. NMT has evolved from recurrent-based architectures to the Transformer model, which has become the dominant architecture due to its efficiency and superior translation quality. Despite these advances, domain adaptation remains a critical challenge, especially for low-resource languages, where data scarcity and domain shifts hinder performance. Fine-tuning and multi-domain adaptation techniques offer promising solutions but introduce issues such as catastrophic forgetting and scalability concerns when extending to multiple domains.

KD has emerged as an effective approach to mitigate these challenges by transferring knowledge from a larger teacher model to a smaller student model. Sequence-level distillation, in particular, has proven useful for NMT by enabling student models to learn directly from teacher-generated translations. However, prior studies have primarily focused on high-resource settings, and little work has explored its effectiveness in low-resource and compute-constrained environments. A key limitation of existing KD approaches is that they predominantly distill knowledge into the decoder while neglecting the encoder, leading to suboptimal generalization across domains. Additionally, conventional KD remains primarily data-centric, focusing on modifying training data rather than enhancing the transfer of internal representations between teacher and student models.

To address these gaps, this thesis investigates an enhanced sequence-level KD approach tailored for low-resource settings. Specifically, we propose incorporating **cosine-based encoder alignment** to improve knowledge transfer in NMT. Unlike conventional KD, our method explicitly aligns teacher and student encoder representations, ensuring better retention of domain-invariant features. By evaluating our approach in a **compute-poor, low-resource scenario**, we aim to demonstrate that our method not only improves domain adaptation but also enhances generalization to unseen domains while maintaining efficiency. This work contributes to bridging the gap in KD research for NMT, particularly in resource-constrained settings where both data and computational resources are limited.

# CHAPTER 3

# METHODOLOGY

## 3.1 Motivation and Hypothesis

NMT models trained using sequence-level KD [21] achieve superior performance compared to direct training on ground-truth data. However, we hypothesize that *sequence-level distillation primarily distills the decoder, leading to suboptimal knowledge transfer in encoder-decoder architectures*. This occurs because sequence-level KD focuses on aligning teacher and student predictions at the target-sequence level, overlooking the internal representations learned by the encoder.

To address the limitations of encoder knowledge transfer, we initially explored a contrastive encoder alignment method using multiple domain-specific teacher models as both positive and negative examples. However, this approach introduced significant complexity and computational overhead due to managing multiple teacher encoders and computing contrastive losses. Based on these observations, we adopted a more streamlined solution: a **cosine-based encoder alignment** mechanism. This approach directly aligns the student encoder with a single teacher encoder by minimizing a cosine embedding loss between their final-layer representations. By simplifying the alignment process, our method improves generalization across domains while maintaining efficiency in low-resource settings.

## 3.2 Initial Proposed Approach

Our initial approach explored a contrastive encoder alignment strategy using multiple domain-specific teacher models. The motivation behind this design was to leverage both high-performing and poorly performing teacher encoders as positive and negative examples, respectively, in a contrastive loss framework. We hypothesized that this approach would enhance the transfer of robust encoder representations to the student model by pushing it closer to strong domain-specific encoders while distancing it from less effective ones.

### 3.2.1 Overview of the Initial Approach

The method involved the following steps:

- **Step 1: Train Domain-Specific Teacher Models** We first trained separate large teacher models, each dedicated to a specific domain. These models were trained on domain-specific datasets to capture domain-specialized representations.

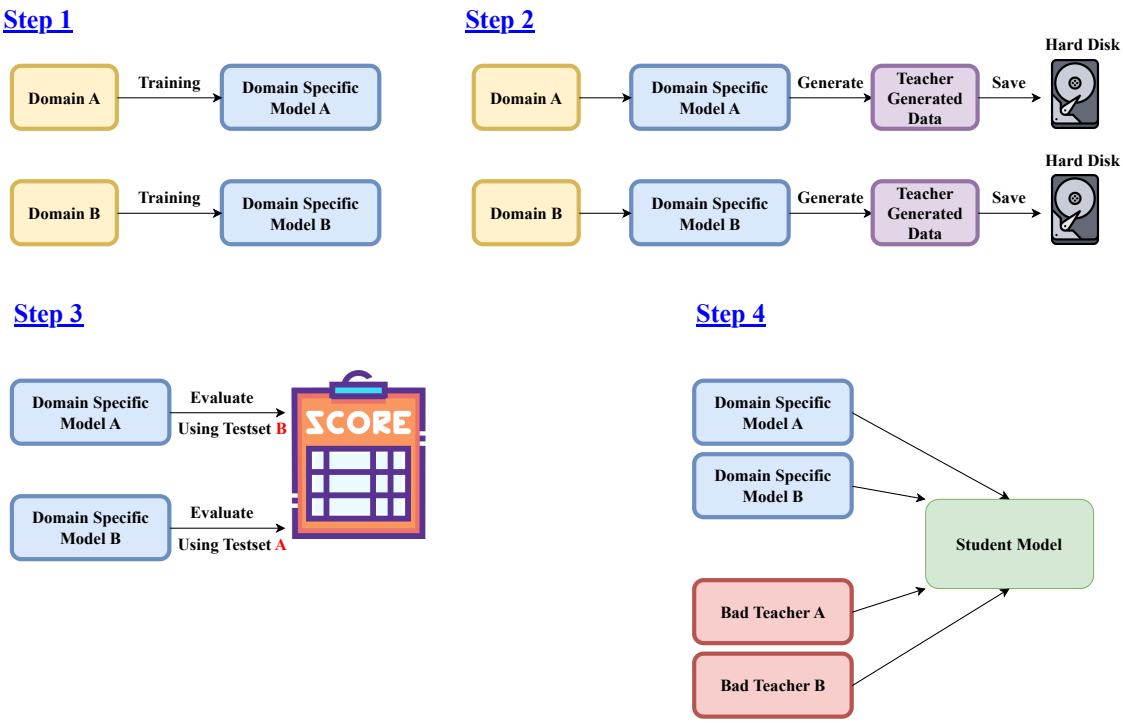


Fig. 3.1: Illustration of the steps involved in the initial proposed approach.

- **Step 2: Generate Distilled Data** Each domain-specific teacher was then used to decode the source-side sentences from its respective domain, generating synthetic target-side translations. These teacher-generated translations formed the distilled dataset for each domain, capturing the teacher's learned distribution.
- **Step 3: Identify "Bad Teacher" Models** We evaluated each domain-specific teacher on its corresponding domain's test set. Subsequently, we identified other domain-specific teachers that performed poorly on the given domain. These underperforming teachers were labeled as "Bad Teachers" and were later used as negative examples in the encoder alignment loss function.
- **Step 4: Train the Student Model with Contrastive Encoder Alignment** A smaller student model was trained on the distilled datasets. During training, a contrastive loss was applied between the student's encoder embeddings and the encoder embeddings from both the domain-specific teacher (as the positive example) and the "Bad Teacher" (as the negative example).

This methodology is visually summarized in Figure 3.1.

### 3.2.2 Limitations of the Initial Approach

As detailed in Section 5.1, our experimental results revealed a key limitation of this strategy. The approach, while conceptually promising, introduced added complexity and memory overhead due to the simultaneous involvement of multiple teacher models and their outputs. Specifically, maintaining multiple domain-specific teachers and computing contrastive loss across all encoders increased GPU memory consumption and made the training pipeline cumbersome.

### 3.2.3 Transition to the Final Methodology

Through iterative experimentation, we observed that consolidating the teacher models into a single multi-domain teacher yielded a more streamlined and memory-efficient framework. By distilling knowledge from a unified teacher model trained on all domains, we reduced GPU memory usage and simplified the training process. Additionally, using a single teacher allowed for consistent and cohesive encoder alignment, eliminating the need for contrastive components based on "Bad Teachers."

Given these advantages, we ultimately abandoned the initial approach in favor of the final methodology, which employs a single multi-domain teacher with cosine-based encoder alignment, as described in Section 3.3.

## 3.3 Overview of Our Final Approach

Our methodology consists of four main steps, illustrated in Figure 3.2 and outlined in Algorithm 1. The training process involves first learning a strong teacher model and then distilling its knowledge into a compact student model, enhanced with encoder alignment.

- **Step 1: Train a Teacher Model** We first train a high-capacity Transformer model (teacher) on the available bilingual dataset until convergence.
- **Step 2: Generate Distilled Data** The teacher model is used to decode the source-side sentences, generating synthetic target translations. These teacher-generated translations serve as distilled data, capturing the teacher's learned distribution.
- **Step 3: Construct the Distilled Mixed Dataset (DMD)** The original training dataset and the teacher-generated dataset are combined to form the Distilled Mixed Dataset (DMD).
- **Step 4: Train the Student Model with Encoder Alignment** A smaller student model is trained on the DMD. Unlike standard KD, we introduce cosine embed-

ding loss between the mean-pooled encoder representations of the teacher and student to encourage encoder knowledge transfer.

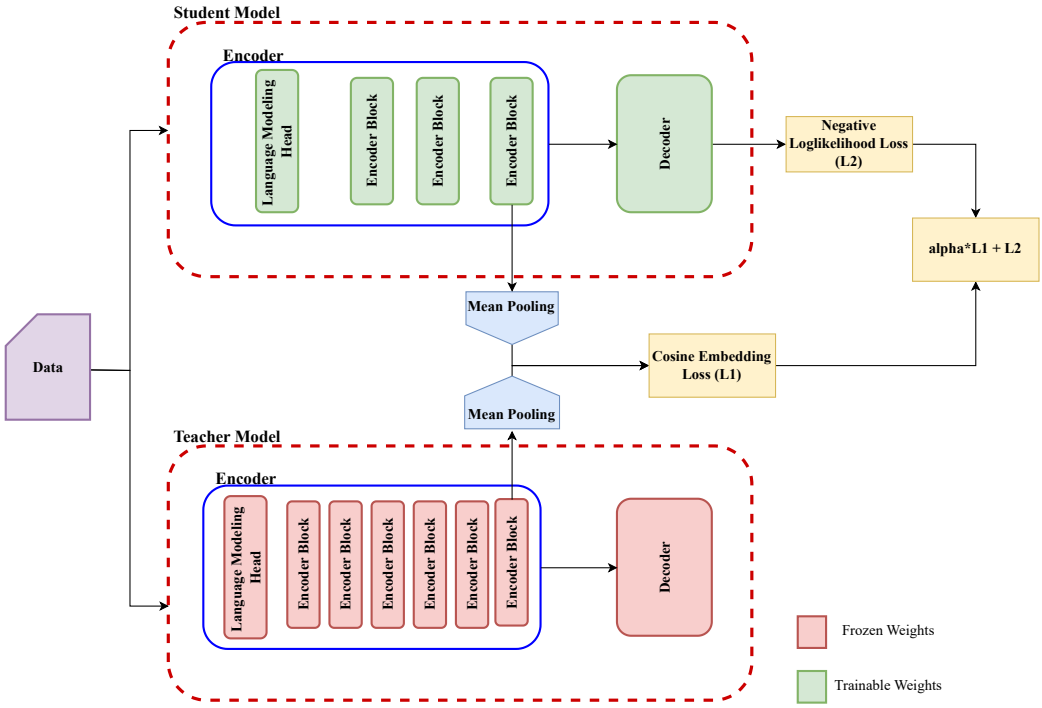


Fig. 3.2: The figure illustrates our proposed final method. The teacher model's weights are frozen, while the student model's weights are learnable. The final loss is computed as the sum of two components: (1) the cosine embedding loss between the mean-pooled final encoder layer outputs of the teacher and student models, and (2) the negative log-likelihood loss of the student model. Note that the input data for both the teacher and student models is the *distill mixed dataset*.

## 3.4 Limitations of Sequence-Level Distillation

Sequence-level distillation has been widely adopted in NMT due to its ability to improve translation fluency and reduce exposure bias [21]. However, a significant limitation of **vanilla sequence-level distillation** is that it primarily affects the decoder while **neglecting the encoder**.

Since sequence-level distillation trains the student model to **mimic teacher-generated sequences**, the decoder benefits the most from learning structured translation patterns. However, the encoder's **latent representations are not explicitly constrained**, leading to a mismatch between the way the student encoder processes source inputs and the way the teacher does. This discrepancy weakens the student model's ability to generalize across domains.

By focusing only on output sequences, standard sequence-level distillation does not encourage better **encoder representation transfer**, making it ineffective for domain adaptation tasks where **shared encoder knowledge** across multiple domains is critical. This motivates our approach to explicitly align encoder representations.

**Algorithm 1:** Sequence-Level Knowledge Distillation with Encoder Alignment

**Input :** Parallel training data $D = \{(X_i, Y_i)\}_{i=1}^{N}$

**Input :** Transformer Teacher Model $T$, Transformer Student Model $S$

**Input :** Hyperparameter $\alpha$ (encoder alignment weight)

**Output:** Trained Student Model $S$

**Step 1: Train the Teacher Model**

Train $T$ on $D$ using the standard NMT loss:

$$\mathcal{L}_{\text{NMT}} = -\sum_{t=1}^{T_y} \log P(y_t | y_{<t}, X; \theta_T)$$

**Step 2: Generate Distilled Data**

**foreach** *sentence* $X_i$ *in* $D$ **do**

  Generate teacher translation $\hat{Y}_i^T$ using $T$

**Step 3: Construct Distilled Mixed Dataset (DMD)**

Create $D_{\text{DMD}} = D \cup D_T$, where:

$$D_T = \{(X_i, \hat{Y}_i^T)\}_{i=1}^{N}$$

**Step 4: Train the Student Model with Encoder Alignment**

**foreach** *batch* $(X, Y, \hat{Y}^T)$ *in* $D_{DMD}$ **do**

  Compute student output $P(y_t | y_{<t}, X; \theta_S)$;

  Compute mean-pooled encoder representations $h_T, h_S$;

  Compute cosine embedding loss:

$$\mathcal{L}_{\text{cosine}} = 1 - \cos(h_T, h_S)$$

  Compute negative log-likelihood loss:

$$\mathcal{L}_{\text{NLL}} = -\sum_{t=1}^{T_y} \log P(y_t^T | y_{<t}^T, X; \theta_S)$$

  Compute total loss:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{cosine}} + \mathcal{L}_{\text{NLL}}$$

  Update student model parameters $\theta_S$ using backpropagation;

**return** *Trained Student Model* $S$

## 3.5 Cosine-Based Encoder Alignment

To address the limitations of sequence-level distillation, we introduce a **cosine embedding loss** to align the encoder representations of the teacher and student models. Given the final-layer encoder outputs $h_T$ and $h_S$ from the teacher and student, we apply mean pooling to obtain a fixed-length vector representation of each sequence. The cosine embedding loss is then computed as:

$$\mathcal{L}_{\text{cosine}} = 1 - \cos(h_T, h_S), \tag{3.1}$$

where $\cos(h_T, h_S)$ is the cosine similarity between the teacher and student encoder representations.

By aligning the encoders in **representation space**, we ensure that the student model learns meaningful latent representations that better match those of the teacher, improving domain adaptation and generalization across unseen data.

## 3.6 Choice of Mean Pooling for Encoder Representations

For computing cosine similarity, we use **mean pooling** over encoder outputs to obtain a fixed-dimensional representation. We justify this choice as follows:

- **Better Sequence Representation:** Mean pooling provides a **global summary** of encoder activations, capturing important contextual features without being dominated by a single token.
- **Empirical Validation:** Studies have shown that **mean pooling outperforms max pooling** and **attention pooling** for sequence-to-sequence models when used for representation alignment [88].
- **Computational Efficiency:** Unlike attention-based pooling, mean pooling is lightweight and efficient, making it suitable for low-resource settings.

## 3.7 Loss Function and Optimization

The total training loss combines both **cosine embedding loss** and **negative log-likelihood loss**:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{cosine}} + \mathcal{L}_{\text{NLL}}, \tag{3.2}$$

where:

- $\mathcal{L}_{\text{cosine}}$ encourages **encoder alignment** between teacher and student.

- $\mathcal{L}_{\text{NLL}}$ is the standard sequence-level distillation loss, ensuring the student model learns proper target translations.
- $\alpha$ is a **hyperparameter** controlling the trade-off between the two loss terms.

## 3.8 Constructing the Distilled Mixed Dataset (DMD)

To effectively combine the benefits of both ground-truth data and distilled translations, we construct a *Distilled Mixed Dataset (DMD)* which serves as the training corpus for the student model. The goal of this dataset is to preserve the fidelity of human-annotated parallel data while incorporating the fluency and consistency provided by teacher-generated outputs.

The DMD is constructed through the following steps:

1. **Teacher Model Training:** A high-capacity teacher model is first trained on the original parallel corpus comprising multiple domains. This model learns rich translation patterns and domain-specific mappings from the data.
2. **Sequence-Level Distillation:** Once trained, the teacher model is used to decode the source sentences from the original training set. For each source sentence, the model generates a corresponding synthetic target sentence using beam search decoding. These teacher-generated translations form the *distilled dataset*.
3. **Dataset Merging:** The distilled target sentences are then paired with their original source sentences to form a synthetic parallel dataset. This is merged with the original ground-truth dataset (comprising human-labeled target sentences) to form the DMD. Each source sentence thus appears twice in the DMD: once with the original target and once with the distilled target.

The resulting Distilled Mixed Dataset (DMD) combines the original parallel sentence pairs with synthetic translations generated by the teacher model for the same source sentences. This dataset is used to train student models in the S-DMD configuration, both with and without encoder alignment. The DMD enables a controlled training setup that integrates original and model-generated supervision to evaluate the impact of encoder alignment under consistent input conditions.

## 3.9 Summary

The methodology initially explored a contrastive encoder alignment strategy using multiple domain-specific teacher models. Each teacher model was trained on a specific domain, and both strong and weak performing teachers were identified. These teacher encoders were used as positive and negative examples, respectively, in a contrastive

loss to align the student encoder representations. However, this approach increased system complexity and memory overhead due to maintaining multiple teacher models and contrastive computations.

To address these limitations, the final approach adopts a simplified and more scalable design using a single multi-domain teacher model. Knowledge transfer is enhanced by introducing a cosine embedding loss between the student and teacher encoders, computed on mean-pooled encoder representations.

The training pipeline is organized into four stages: training the multi-domain teacher model, generating synthetic translations through sequence-level distillation, constructing the Distilled Mixed Dataset (DMD) by combining original and synthetic parallel data, and finally, training the student model. The DMD serves as a consistent input for evaluating both the baseline and encoder-aligned student models. This final approach reduces computational overhead while enabling effective encoder knowledge transfer and improved generalization across domains in low-resource settings.

# CHAPTER 4

# EXPERIMENTAL SETUP

## 4.1 Experimental Framework

This chapter details the experimental setup used to evaluate our proposed approach. We outline the hardware and software configurations, the datasets used, and the training and inference procedures. Our experiments were designed to address the research question: *Can encoder alignment improve sequence-level distillation in a low-resource, compute-constrained setting?*

We simulate a low-resource setting because LRLs often lack diverse domains, and the available domain-specific data is typically insufficient to rigorously test our hypothesis. To create a realistic low-resource scenario, we trained models with limited parallel data and evaluated their ability to generalize to new domains. The experiments were conducted in two main stages:

- **Stage 1**: Evaluating our encoder alignment method on a multi-domain setting.
- **Stage 2**: Assessing the impact of encoder alignment when fine-tuning on new domains.

## 4.2 Hardware and Software Configurations

### 4.2.1 Hardware Specifications

All experiments were conducted on a single machine with the following hardware configuration:

- **CPU**: Intel i9-9900K
- **RAM**: 64GB
- **GPU**: Nvidia Quadro RTX 6000 (24GB VRAM)

This setup represents a compute-constrained environment, as opposed to large-scale distributed training on multi-GPU clusters. The choice of limited resources aligns with our focus on low-resource settings, where computational constraints are common.

### 4.2.2 Software Specifications

We used the HuggingFace Transformers library [89] for model implementation. Evaluation was conducted using chrF score from the `evaluate`[1] library of HuggingFace.

[1] https://github.com/huggingface/evaluate

## 4.3 Datasets

### 4.3.1 Dataset Selection and Preprocessing

We selected the German-to-English translation task and used six domain-specific datasets from the OPUS repository [90]. The details of these datasets are provided in Table 4.1.

| Dataset Name (Citation) | Abbreviation | Description |
|---|---|---|
| Europarl [91] | parl | Parliamentary proceedings |
| JRC-Acquis [90] | law | Legal domain corpus |
| EMEA [90] | med | Biomedical texts from the European Medicines Agency |
| News Commentary [90] | news | Collection of news articles |
| OpenSubtitles [92] | opensub | Movie and TV subtitles dataset |
| TED 2020 [93] | ted | Transcripts from TED talks |

**TABLE** 4.1: Domain-specific datasets used in our experiments.

To emulate a low-resource setting, we randomly sampled and deduplicated sentences based on length constraints. The dataset splits are as follows:

- Training set: 50K parallel sentences.
- Development and test sets: 1K sentences each.
- Sentences were selected with word counts between 4 and 120.

### 4.3.2 Experimental Stages

**Stage 1**: We trained models on a **multi-domain setting** using a combination of *parl*, *law*, *news*, and *med*. Additionally, models were evaluated on the **Flores200** (flores) development set [94] to assess **out-of-domain generalization**.

**Stage 2**: To assess domain adaptation, we fine-tuned the models from Stage 1 on **opensub** and **ted** datasets using standard fine-tuning techniques.

## 4.4 Training and Inference Details

### 4.4.1 Model Selection

**Teacher Model**: We selected **T5-small** [95] (77M parameters) as the teacher model. **Student Model**: The student model was derived by **removing three encoder and decoder layers** from T5-small. Both models were randomly initialized to ensure that no pre-trained knowledge influenced the results.

### 4.4.2 Training Setup

All models were trained under the following conditions:

- **Max Epochs**: 100 (early stopping with patience = 4)
- **Learning Rate**: $2 \times 10^{-4}$
- **Batch Size**: 64 (gradient accumulation = 2, effective batch size = 128)
- **Max Sequence Length**: 120 tokens

### 4.4.3 Inference and Generation

During inference, the same settings were used as in training:

- **Beam Search**: Beam size = 5
- **Max Sequence Length**: 120 tokens

Beam search ensures **high-quality translations** during both inference and KD-based data generation.

## 4.5 Model Configurations

To evaluate the impact of our proposed methodology, we trained five different models. Table 4.2 provides an overview of these configurations.

| **Model Name** | **Description** |
|---|---|
| L-ADO | Large model trained on the All-Domain Original dataset. |
| S-ADO | Small model trained on the All-Domain Original dataset. |
| S-ADD | Small model trained on the All-Domain Distilled dataset. |
| S-DMD-NoAlign | Small model trained on the Distilled Mixed Dataset (DMD) without encoder alignment. |
| S-DMD-Align | Small model trained on the DMD with encoder alignment (proposed method). |

**TABLE** 4.2: Model configurations used in our experiments.

The term *All-Domain* refers to the combined dataset of *parl*, *law*, *news*, and *med*. The L-ADO model serves as the teacher model for all student models trained via KD.

## 4.6 Summary

This chapter detailed our experimental setup, including the hardware, software, datasets, training, and inference methodologies. We designed our experiments to simulate low-resource and compute-constrained environments, testing our encoder alignment method under realistic conditions. The next chapter presents the results and analysis of our experiments.

# CHAPTER 5

# RESULTS AND DISCUSSION

## 5.1 Discussion on the Initial Approach

The initial approach was designed around training domain-specific teacher models and employing contrastive loss using "bad teachers" to improve encoder alignment. The idea was that for each domain, we would select a poorly performing teacher as a negative example to push the student encoder's representation away from it.

Table 5.1 presents the performance of each domain-specific expert model across different domains. For instance, the *Ted Expert* performed the worst on both the *law* and *medicine* test domains, while the *Medicine Expert* performed poorly on *news* and *ted* domains.

**TABLE** 5.1: ChrF scores for teacher models on different test domains. The lowest-performing model for each domain is bolded.

| Test Domain | Teacher Model | ChrF Score |
|---|---|---|
| Law | Law Teacher | 20.74 |
| | Medicine Teacher | 2.23 |
| | News Teacher | 1.26 |
| | **Ted Teacher** | **0.67** |
| Medicine | Law Teacher | 1.90 |
| | Medicine Teacher | 19.40 |
| | News Teacher | 0.81 |
| | **Ted Teacher** | **0.55** |
| News | Law Teacher | 1.06 |
| | **Medicine Teacher** | **0.53** |
| | News Teacher | 3.52 |
| | Ted Teacher | 1.33 |
| Ted | Law Teacher | 0.36 |
| | **Medicine Teacher** | **0.32** |
| | News Teacher | 2.13 |
| | Ted Teacher | 4.67 |

As shown in Table 5.2, these bad teachers were selected for each domain and used to calculate the contrastive loss during student training.

However, when we evaluated the final models trained under this methodology, the performance did not meet expectations. Table 5.3 shows the multi-domain perfor-

**TABLE** 5.2: IDENTIFIED BAD TEACHERS FOR EACH DOMAIN.

| Domains | Bad Teacher |
|---|---|
| Law | Ted Teacher |
| Medicine | Ted Teacher |
| News | Medicine Teacher |
| Ted | Medicine Teacher |

mance of our approach against other baseline methods. Notably, our approach underperformed across all domains when compared to simple baselines such as S-ADO and S-ADD.

| Model | med | parl | law | news | Flores |
|---|---|---|---|---|---|
| L-ADO | 63.27 | 56.33 | 63.73 | 53.80 | 50.89 |
| S-ADO | 62.31 | 55.66 | 62.39 | 53.28 | 50.23 |
| S-ADD | 62.36 | 56.08 | 62.92 | 53.87 | 50.90 |
| S-DMD-NoAlign | 61.38 | 55.49 | 61.89 | 52.91 | 49.88 |
| Our Approach | 58.93 | 54.55 | 58.34 | 50.88 | 47.01 |

**TABLE** 5.3: ChrF scores of models trained with various configurations, evaluated on in-domain and out-of-domain test sets.

To further fine-tune this approach, we experimented with different pooling mechanisms for the encoder representations. Table 5.4 shows that **average pooling** consistently outperformed *attention* pooling strategy, suggesting that average pooling provides a more stable representation for encoder alignment.

| Pooling Method | ChrF |
|---|---|
| **Average Pooling** | **34.97** |
| Attention Pooling | 32.96 |

**TABLE** 5.4: Comparison of pooling methods based on average ChrF score for model trained on single individual domain.

Despite these adjustments, our approach did not surpass even the simplest baseline. We believe since the domain-specific teachers were trained independently, their encoder representations lacked a shared representation space. Consequently, the learned vector spaces for each teacher could overlap. This undermined the contrastive signal, as the bad teacher's embeddings could be located near embeddings from other well-performing teachers in certain domains.

In essence, the contrastive loss struggled to effectively push the student encoder away from the bad teacher embeddings, as the distance between positive and negative examples was ambiguous. This observation is akin to the saying, *"too many cooks spoil the broth"*.

Ultimately, this insight led us to abandon this multi-teacher contrastive approach in favor of a single all-domain teacher model, which provides a unified representation space and simplifies the distillation process, both in terms of performance and computational overhead.

## 5.2 Discussion of our Final Approach

In this chapter, we present and analyze the experimental results of our proposed methodology. Our primary objective is to evaluate whether sequence-level distillation primarily distills the decoder, leading to suboptimal knowledge transfer, and whether our proposed encoder alignment method mitigates this limitation. We examine the performance of models across two stages:

- **Stage 1**: Multi-domain performance using the original training domains (*med, parl, law, news*) and generalization to out-of-domain data (*flores*).
- **Stage 2**: Fine-tuning on unseen domains (*opensub* and *ted*) to assess domain adaptation capability.

By analyzing the impact of encoder alignment and KD, we demonstrate how our approach improves multi-domain translation while being efficient in **low-resource and compute-constrained settings**.

### 5.2.1 Hyperparameter Selection: Impact of $\alpha$

To optimize the contribution of encoder alignment in our loss function, we experimented with different values of the attenuation factor ($\alpha$). Table 5.5 presents the results.

| $\alpha$ | **med** | **parl** | **law** | **news** | **flores** |
|---|---|---|---|---|---|
| 1.0 | 63.44 | 56.84 | 63.99 | 54.55 | 52.64 |
| 1.5 | 63.56 | 56.91 | 63.88 | 54.68 | 52.50 |
| 2.0 | 63.43 | 56.92 | 64.08 | 54.88 | 52.90 |

**TABLE** 5.5: ChrF scores of our model trained with different $\alpha$ values, evaluated on in-domain test sets (med, parl, law, news) and the out-of-domain flores development-test set.

The results indicate a general improvement in performance as $\alpha$ increases, with $\alpha = 2.0$ achieving the best overall scores across both in-domain and out-of-domain evaluations. Notably, performance in the law domain and the flores test set exhibited slight variations, suggesting that beyond a certain point, increasing $\alpha$ does not yield further gains. Given the computational cost of each training run (12+ hours, trained

until convergence with maximum epochs set to 100), we selected $\alpha = 2.0$ for all subsequent experiments.

### 5.2.2 Stage 1: Multi-Domain Performance and Generalization

Table 5.6 presents the performance of different models trained with various configurations. Our proposed method, S-DMD-Align, consistently achieves the highest scores across all domains, surpassing both small student models and even the large teacher model (L-ADO).

| Model | med | parl | law | news | Flores |
|---|---|---|---|---|---|
| L-ADO | 63.27 | 56.33 | 63.73 | 53.80 | 50.89 |
| S-ADO | 62.31 | 55.66 | 62.39 | 53.28 | 50.23 |
| S-ADD | 62.36 | 56.08 | 62.92 | 53.87 | 50.90 |
| S-DMD-NoAlign | 61.38 | 55.49 | 61.89 | 52.91 | 49.88 |
| S-DMD-Align | **63.43** | **56.92** | **64.08** | **54.88** | **52.90** |

**TABLE** 5.6: ChrF scores of models trained with various configurations, evaluated on in-domain test sets (med, parl, law, news) and the out-of-domain flores development-test set.

These results strongly support our hypothesis that sequence-level distillation primarily distills the decoder, leading to suboptimal knowledge transfer. The S-ADD model, which follows the traditional sequence-level distillation approach, fails to outperform its teacher model (L-ADO), confirming that vanilla distillation does not effectively transfer knowledge.

By explicitly aligning the encoder representations, S-DMD-Align significantly outperforms all baselines, even surpassing the large model (L-ADO) while using half the encoder and decoder layers. This validates our hypothesis that enhanced encoder alignment improves generalization across multiple domains while maintaining computational efficiency.

Furthermore, out-of-domain performance on flores shows that our method generalizes better than traditional sequence-level distillation approaches, as S-DMD-Align outperforms S-ADD by +2.67 ChrF points. This indicates that encoder alignment helps retain generalization capabilities even for unseen test domains.

### 5.2.3 Stage 2: Domain Adaptation Performance

To assess domain adaptation capabilities, we fine-tuned models from Stage 1 on unseen domains (*opensub* and *ted*). Table 5.7 presents the results.

Consistent with Stage 1 results, S-DMD-Align continues to outperform all baselines, including the large model (L-ADO). This confirms that our encoder alignment

| Model | opensub | ted |
|---|---|---|
| L-ADO | 39.94 | 51.32 |
| S-ADO | 39.21 | 51.03 |
| S-ADD | 39.59 | 50.51 |
| S-DMD-NoAlign | 39.13 | 50.41 |
| S-DMD-Align | **40.43** | **51.94** |

**TABLE** 5.7: ChrF scores for Stage 1 models fine-tuned on single domains (Open Subtitles and Ted2020) to evaluate domain adaptation.

strategy enhances domain adaptation, allowing smaller models to retain strong translation capabilities even when fine-tuned on new domains.

The performance gap between S-DMD-Align and S-DMD-NoAlign highlights the importance of encoder alignment in knowledge transfer. S-ADD, which follows traditional sequence-level distillation, performs worse than S-ADO, reinforcing that vanilla distillation alone is insufficient for domain transfer.

## 5.3 Evaluation on Bona Fide Low-Resource Language

To evaluate the robustness of our method in a genuinely low-resource setting, we conduct an ablation study on the English–Sinhala language pair. This language direction is characterized by limited parallel data and narrow domain coverage, making it a representative testbed for evaluating NMT models under severe data constraints. In this study, we focus on examining the impact of the alignment weighting parameter $\alpha$ within our proposed encoder-aligned distillation framework.

| $\alpha$ | ccalign | opensub | gov | Flores |
|---|---|---|---|---|
| 1.0 | 38.91 | 28.71 | 44.25 | **28.11** |
| 2.0 | 39.06 | 28.88 | 44.35 | 28.04 |
| 3.0 | 38.79 | 28.21 | 43.80 | 27.43 |
| 4.0 | **39.54** | **28.91** | **44.66** | 27.54 |
| 5.0 | 37.96 | 28.43 | 43.65 | 27.73 |
| 6.0 | 36.27 | 27.89 | 41.85 | 25.41 |
| 7.0 | 38.59 | 28.86 | 43.91 | 27.71 |

**TABLE** 5.8: ChrF scores of our model trained on the English–Sinhala language pair with different $\alpha$ values using the distilled dataset, evaluated on three in-domain test sets and the out-of-domain Flores200 development-test set.

Table 5.8 presents the ChrF scores of our model trained with different $\alpha$ values on the distilled dataset. Evaluations were performed on three in-domain datasets—*ccalign*, *opensub*, and *gov*—as well as one out-of-domain test set, *Flores*. This setup enables

efficient experimentation in a constrained environment and facilitates hyperparameter tuning for encoder alignment under low-resource conditions.

We observe that in-domain performance peaks consistently at $\alpha = 4.0$, achieving the highest scores across all three datasets. This suggests that in low-resource settings, assigning greater importance to encoder alignment substantially improves domain-specific translation quality. On the other hand, the highest out-of-domain performance on *Flores* occurs at $\alpha = 1.0$, indicating that equal weighting between the NMT loss and the alignment loss better supports generalization beyond the training domains. Importantly, performance does not follow a monotonic trend with increasing $\alpha$; rather, it varies across datasets, highlighting the dataset-dependent nature of the alignment loss contribution.

| **Model** | $\alpha$ | **ccalign** | **opensub** | **gov** | **Flores** |
|---|---|---|---|---|---|
| L-ADO | – | 41.95 | 28.88 | **48.44** | 29.81 |
| S-ADO | – | 39.23 | 28.58 | 45.69 | 28.34 |
| S-ADD | – | 38.41 | 28.67 | 43.46 | 27.15 |
| S-DMD-NoAlign | – | 42.34 | 30.11 | 47.62 | 30.47 |
| S-DMD-Align | 1.0 | 42.78 | 30.36 | 47.25 | 30.54 |
| S-DMD-Align | 4.0 | **43.11** | **30.42** | 48.20 | **31.03** |

**TABLE** 5.9: ChrF scores of models trained with various configurations for the English–Sinhala translation direction, evaluated on three in-domain test sets and the out-of-domain Flores200 development-test set.

Building upon the hyperparameter tuning results, we select $\alpha = 1.0$ and $\alpha = 4.0$ for further evaluation on the full Distilled Mixed Dataset (DMD). As shown in Table 5.9, our encoder-aligned model with $\alpha = 4.0$ outperforms all baselines across the board, including the strong large-teacher model (L-ADO) in three out of four domains. The only exception is the *gov* domain, where L-ADO slightly outperforms our best model by 0.24 ChrF points. Notably, the S-ADD model, which employs vanilla sequence-level distillation, performs worse than even the small model trained on the original dataset (S-ADO), reinforcing the inadequacy of sequence-level KD in low-resource scenarios.

Although $\alpha = 1.0$ yielded the best *Flores* performance during tuning, it is consistently outperformed by $\alpha = 4.0$ across all domains when trained on the combined dataset. This indicates that alignment-based models benefit from a higher alignment loss weight during full training. Overall, our findings highlight that in bona fide low-resource directions, greater emphasis on encoder alignment plays a pivotal role in enhancing both in-domain accuracy and out-of-domain generalization.

## 5.4 Summary of Findings

Our experiments validate the primary hypothesis that sequence-level distillation primarily benefits the decoder while neglecting encoder knowledge transfer. By introducing encoder alignment, we demonstrate consistent improvements in both performance and generalization across domains. Our findings are summarized as follows:

- Small models trained with our proposed method outperform even their large teacher counterparts across both in-domain and out-of-domain scenarios.
- Encoder alignment significantly enhances multi-domain NMT performance, with models trained on a distilled mixed dataset (DMD) showing marked improvements over standard sequence-level distillation.
- Our method generalizes better to unseen domains, particularly in low-resource and compute-constrained settings.
- Distilled models with encoder alignment retain superior domain adaptation capabilities, even when fine-tuned on new domains.
- In a bona fide low-resource setting (English–Sinhala), we find that encoder alignment yields substantial gains, with $\alpha = 4$ performing best across in-domain and out-of-domain evaluations.
- The impact of the encoder alignment hyperparameter ($\alpha$) is dataset-dependent, with higher values offering stronger improvements in resource-poor and domain-limited conditions.

These results establish encoder-aware KD as an effective, scalable, and resource-efficient solution for multi-domain NMT, particularly well-suited to the challenges of low-resource languages.

## 5.5 Alignment Between Objectives and Contributions

To ensure clarity and coherence in how the research objectives of this thesis were addressed, this section presents a structured mapping between each objective and the corresponding methodological strategies and outcomes. The table below outlines how the goals defined in Chapter 1 were translated into concrete experiments, design decisions, and evaluations throughout the course of the study. This mapping highlights the alignment between the intended aims and the actual contributions made, offering a consolidated view of how each component of the research supports the overarching goal of improving multi-domain neural machine translation in low-resource settings.

| Research Objective | Approach Taken | Deliverable / Evidence |
|---|---|---|
| To investigate the limitations of sequence-level KD in NMT, particularly its impact on encoder knowledge transfer in low-resource settings | Performed empirical comparisons between student models trained with and without KD to analyze the extent of encoder knowledge transfer; identified that standard KD focuses predominantly on distilling decoder behavior, with minimal transfer of encoder-level representations | Discussion in Section 5.2.2; Quantitative results in Table 5.6 showing underperformance of S-ADD; Summarized in Section 5.4 |
| To design and assess an encoder-alignment strategy to enhance knowledge transfer between teacher and student models for improved multi-domain generalization in NMT | Introduced a cosine-based encoder alignment loss between mean-pooled final encoder outputs of teacher and student models; integrated the alignment mechanism into the distillation pipeline for improved encoder representation learning | Methodology described in Chapter 3; Visual representation in Figure 3.2; Implementation details in Algorithm 1 |
| To assess the effectiveness of the proposed encoder-aligned distillation approach in both in-domain and out-of-domain scenarios, demonstrating its applicability in real-world low-resource and compute-constrained environments | Trained distilled T5-small models (3 encoder and decoder layers) using 50K parallel sentence pairs on a single-GPU setup; evaluated performance across multiple in-domain and out-of-domain test sets (using Flores for generalization), and conducted a separate evaluation on the English–Sinhala language pair to assess robustness in a bonafide low-resource setting | Results across multi-domain and generalization scenarios in Tables 5.6 and 5.7; Low-resource validation on English–Sinhala in Table 5.9; Experimental setup documented in Chapter 4 |

**TABLE** 5.10: Alignment between research objectives, methodological approaches, and corresponding deliverables.

# CHAPTER 6

# CONCLUSION

In this thesis, we addressed the challenge of building scalable and efficient multi-domain NMT systems for low-resource languages under compute-constrained settings. We began by identifying a key limitation of conventional sequence-level Knowledge Distillation (KD): while effective at transferring decoder knowledge, it fails to adequately distill encoder representations, leading to limited domain generalization.

To overcome this, we proposed an encoder-aligned distillation framework that explicitly aligns the final-layer encoder representations of teacher and student models using a cosine embedding loss. Our method was iteratively developed, starting with a contrastive alignment strategy using multiple domain-specific teachers, which was later abandoned due to its added complexity and inefficiency. We instead adopted a streamlined approach using a single multi-domain teacher model, improving both performance and training efficiency.

Comprehensive experiments on German–English translation under simulated low-resource conditions demonstrated that our method consistently outperforms standard KD and baseline models across in-domain, out-of-domain, and domain adaptation tasks. Further evaluation on the English–Sinhala language pair, a bona fide low-resource language, confirmed that the proposed method remains effective in truly low-resource settings. We also observed that the optimal encoder alignment weight is dataset-dependent, reinforcing the importance of context-sensitive tuning in low-resource scenarios.

Overall, our findings underscore the significance of encoder-level knowledge transfer in KD and establish encoder-aligned distillation as a robust and scalable solution for multi-domain NMT in low-resource environments.

## Limitations

Despite the promising results demonstrated by our encoder-aligned distillation framework, there are a few important limitations to acknowledge. First, due to computational constraints, our experiments were conducted using T5-small models. While our method outperformed baseline approaches in this setting, its effectiveness on larger encoder-decoder architectures remains unverified. Further investigation is required to assess whether the proposed alignment strategy scales favorably with model size.

Second, the scope of this work is limited to encoder-decoder architectures. Many state-of-the-art large language models adopt decoder-only designs, for which our method has not been evaluated. As such, its applicability beyond the encoder-decoder paradigm remains an open question.

## Future Work

Building on the current findings, future work could focus on extending the proposed method to larger pretrained architectures, evaluating whether encoder alignment continues to provide benefits at scale. Additionally, exploring adaptation strategies suitable for decoder-only models may help generalize encoder-aware distillation to a broader range of architectures.

# REFERENCES


[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, 2007, pp. 177–180.

[2] A. Lopez, "Statistical machine translation," *ACM Computing Surveys (CSUR)*, vol. 40, no. 3, pp. 1–49, 2008.

[3] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–37, 2023.

[4] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–38, 2020.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *arXiv preprint arXiv:1409.3215*, 2014.

[8] R. Sennrich and B. Zhang, "Revisiting low-resource neural machine translation: A case study," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 211–221. [Online]. Available: https://aclanthology.org/P19-1021/

[9] S. Ranathunga, E.-S. A. Lee, M. P. Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," 2021.

[10] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl, and A. Birch, "Survey of low-resource machine translation," *Computational Linguistics*, vol. 48, no. 3, pp. 673–732, 2022.

[11] R. Östling and J. Tiedemann, "Neural machine translation for low-resource languages," *arXiv preprint arXiv:1708.05729*, 2017.

[12] S. Manchanda and G. Grunin, "Domain informed neural machine translation: Developing translation services for healthcare enterprise," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada, Eds. Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 255–261. [Online]. Available: https://aclanthology.org/2020.eamt-1.27/

[13] D. Saunders, "Domain adaptation and multi-domain adaptation for neural machine translation: A survey," *Journal of Artificial Intelligence Research*, vol. 75, pp. 351–424, 2022.

[14] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, T. Luong, A. Birch, G. Neubig, and A. Finch, Eds. Vancouver: Association for Computational Linguistics, Aug. 2017, pp. 28–39. [Online]. Available: https://aclanthology.org/W17-3204/

[15] C. Chu and R. Wang, "A survey of domain adaptation for neural machine translation," in *Proceedings of the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1304–1319. [Online]. Available: https://aclanthology.org/C18-1111/

[16] A. Currey, P. Mathur, and G. Dinu, "Distilling multiple domains for neural machine translation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4500–4511. [Online]. Available: https://aclanthology.org/2020.emnlp-main.364/

[17] Y. Liang, F. Meng, J. Wang, J. Xu, Y. Chen, and J. Zhou, "Continual learning with semi-supervised contrastive distillation for incremental neural machine translation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10 914–10 928. [Online]. Available: https://aclanthology.org/2024.acl-long.588/

[18] X. Pan, M. Wang, L. Wu, and L. Li, "Contrastive learning for many-to-many multilingual neural 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 (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 244–258. [Online]. Available: https://aclanthology.org/2021.acl-long.21/

[19] M. Pham, J. Crego, F. Yvon, and J. Senellart, "Generic and specialized word embeddings for multi-domain machine translation," in *Proceedings of the 16th International Conference on Spoken Language Translation*, J. Niehues, R. Cattoni, S. Stüker, M. Negri, M. Turchi, T.-L. Ha, E. Salesky, R. Sanabria, L. Barrault, L. Specia, and M. Federico, Eds. Hong Kong: Association for Computational Linguistics, Nov. 2-3 2019. [Online]. Available: https://aclanthology.org/2019.iwslt-1.26/

[20] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[21] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1317–1327. [Online]. Available: https://aclanthology.org/D16-1139/

[22] C. Chu, R. Dabre, and S. Kurohashi, "An empirical comparison of domain adaptation methods for neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 385–391. [Online]. Available: https://aclanthology.org/P17-2061/

[23] R. Dabre, A. Fujita, and C. Chu, "Exploiting multilingualism through multistage fine-tuning for low-resource neural 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1410–1416. [Online]. Available: https://aclanthology.org/D19-1146/

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016. [Online]. Available: https://arxiv.org/abs/1409.0473

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[26] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179/

[27] B. Ghojogh and A. Ghodsi, "Recurrent neural networks and long short-term memory networks: Tutorial and survey," 2023. [Online]. Available: https://arxiv.org/abs/2304.11461

[28] R. C. Staudemeyer and E. R. Morris, "Understanding lstm – a tutorial into long short-term memory recurrent neural networks," 2019. [Online]. Available: https://arxiv.org/abs/1909.09586

[29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014. [Online]. Available: https://arxiv.org/abs/1412.3555

[30] S. A. Mohamed, A. A. Elsayed, Y. Hassan, and M. A. Abdou, "Neural machine translation: past, present, and future," *Neural Computing and Applications*, vol. 33, pp. 15 919–15 931, 2021.

[31] M. X. Chen, O. Firat, A. Bapna, M. Johnson, W. Macherey, G. Foster, L. Jones, N. Parmar, M. Schuster, Z. Chen, Y. Wu, and M. Hughes, "The best of both worlds: Combining recent advances in neural machine translation," 2018. [Online]. Available: https://arxiv.org/abs/1804.09849

[32] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling neural machine translation," 2018. [Online]. Available: https://arxiv.org/abs/1806.00187

[33] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Łukasz Kaiser, "Universal transformers," 2019. [Online]. Available: https://arxiv.org/abs/1807.03819

[34] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," 2020. [Online]. Available: https://arxiv.org/abs/1909.11942

[35] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" 2019. [Online]. Available: https://arxiv.org/abs/1905.10650

[36] J. Kasai, H. Peng, Y. Zhang, D. Yogatama, G. Ilharco, N. Pappas, Y. Mao, W. Chen, and N. A. Smith, "Finetuning pretrained transformers into RNNs," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10 630–10 643. [Online]. Available: https://aclanthology.org/2021.emnlp-main.830/

[37] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, ..., and J. Wang, "No Language Left Behind: Scaling Human-Centered Machine Translation," 2022. [Online]. Available: https://arxiv.org/abs/2207.04672

[38] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," 2020. [Online]. Available: https://arxiv.org/abs/2010.11125

[39] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41/

[40] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.47/

[41] H. M. Alemayehu, H. M. Zahera, and A.-C. Ngonga Ngomo, "Error analysis of multilingual language models in machine translation: A case study of English-Amharic translation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 19 758–19 768. [Online]. Available: https://aclanthology.org/2024.emnlp-main.1102/

[42] E.-S. A. Lee, S. Thillainathan, S. Nayak, S. Ranathunga, D. I. Adelani, R. Su, and A. D. McCarthy, "Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation?" in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 58–67. [Online]. Available: https://aclanthology.org/2022.findings-acl.6/

[43] A. El-Kishky, V. Chaudhary, A. Fan, S. Bhosale, H. Schwenk, Zhiyi, S. Goyal, M. Baines, O. Çelebi, G. Wenzek, A. Fernando, S. Ranathunga, N. Goyal, C. Gao, J. Chen, D. Ju, S. Kr-373, M. Ranzato, F. Guzmán, F. Guzmán, P.-J. Chen, M. Ott, G. Pino, P. Lample, and V. Koehn, "Sequence-to-Sequence Multilingual Pre-Trained Models: A Hope for Low-Resource Language Translation?" 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:252620124

[44] C. Cieri, M. Maxwell, S. Strassel, J. Devlin, M.-W. Chang, K. Lee, D. M. Eberhard, G. F. Simons, K. Kann, O. Lacroix, and A. Sø-599, "Exploring the low-resource transfer-learning with mt5 model," 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252547526

[45] T. J. Singh, S. R. Singh, and P. Sarmah, "Can big models help diverse languages? investigating large pretrained multilingual models for machine translation of Indian languages," in *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, J. D. Pawar and S. Lalitha Devi, Eds. Goa University, Goa, India: NLP Association of India (NLPAI), Dec. 2023, pp. 663–669. [Online]. Available: https://aclanthology.org/2023.icon-1.66/

[46] A. Galiano-Jiménez, F. Sánchez-Martínez, V. M. Sánchez-Cartagena, and J. A. Pérez-Ortiz, "Exploiting large pre-trained models for low-resource neural machine translation," in *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín,

M. Forcada, M. Popovic, C. Scarton, and H. Moniz, Eds. Tampere, Finland: European Association for Machine Translation, Jun. 2023, pp. 59–68. [Online]. Available: https://aclanthology.org/2023.eamt-1.7/

[47] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," 2015. [Online]. Available: https://arxiv.org/abs/1312.6211

[48] Y. Liang, F. Meng, J. Xu, J. Wang, Y. Chen, and J. Zhou, "Unified model learning for various neural machine translation," 2023. [Online]. Available: https://arxiv.org/abs/2305.02777

[49] J. Wu, Y. Liu, and C. Zong, "F-malloc: Feed-forward memory allocation for continual learning in neural machine translation," *arXiv preprint arXiv:2404.04846*, 2024.

[50] D. Britz, Q. Le, and R. Pryzant, "Effective domain mixing for neural machine translation," in *Proceedings of the Second Conference on Machine Translation*, O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, and J. Kreutzer, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 118–126. [Online]. Available: https://aclanthology.org/W17-4712/

[51] Z. Liu, G. I. Winata, and P. Fung, "Continual mixed-language pre-training for extremely low-resource neural machine translation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 2706–2718. [Online]. Available: https://aclanthology.org/2021.findings-acl.239/

[52] W.-J. Ko, A. El-Kishky, A. Renduchintala, V. Chaudhary, N. Goyal, F. Guzmán, P. Fung, P. Koehn, and M. Diab, "Adapting high-resource NMT models to translate low-resource related languages without parallel data," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 802–812. [Online]. Available: https://aclanthology.org/2021.acl-long.66/

[53] A. Bapna and O. Firat, "Non-parametric adaptation for 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)*, J. Burstein, C. Doran, and

T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1921–1931. [Online]. Available: https://aclanthology.org/N19-1191/

[54] R. Aharoni and Y. Goldberg, "Unsupervised domain clusters in pretrained language models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7747–7763. [Online]. Available: https://aclanthology.org/2020.acl-main.692/

[55] C. Escolano, M. R. Costa-Jussà, and J. A. R. Fonollosa, "From bilingual to multilingual neural-based machine translation by incremental training," *Journal of the Association for Information Science and Technology*, vol. 72, no. 2, pp. 190–203, 2021. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24395

[56] Y. Cao, H.-R. Wei, B. Chen, and X. Wan, "Continual learning for neural machine translation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 3964–3974. [Online]. Available: https://aclanthology.org/2021.naacl-main.310/

[57] B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn, "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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2062–2068. [Online]. Available: https://aclanthology.org/N19-1209/

[58] D. Saunders and S. DeNeefe, "Domain adapted machine translation: What does catastrophic forgetting forget and why?" in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12 660–12 671. [Online]. Available: https://aclanthology.org/2024.emnlp-main.704/

[59] E. Hasler, T. Domhan, J. Trenous, K. Tran, B. Byrne, and F. Hieber, "Improving the quality trade-off for neural machine translation multi-domain adaptation," in *Proceedings of the 2021 Conference on Empirical Methods*

*in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8470–8477. [Online]. Available: https://aclanthology.org/2021.emnlp-main.666/

[60] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, p. 3521–3526, Mar. 2017. [Online]. Available: http://dx.doi.org/10.1073/pnas.1611835114

[61] J. Park, J. Song, and S. Yoon, "Building a neural machine translation system using only synthetic parallel data," 2017. [Online]. Available: https://arxiv.org/abs/1704.00253

[62] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96. [Online]. Available: https://aclanthology.org/P16-1009/

[63] S. Tars and M. Fishel, "Multi-domain neural machine translation," *arXiv preprint arXiv:1805.02282*, 2018.

[64] K. Song, Y. Zhang, H. Yu, W. Luo, K. Wang, and M. Zhang, "Code-switching for enhancing nmt with pre-specified translation," 2019. [Online]. Available: https://arxiv.org/abs/1904.09107

[65] Z. Yang, W. Chen, F. Wang, and B. Xu, "Generative adversarial training for neural machine translation," *Neurocomputing*, vol. 321, pp. 146–155, 2018.

[66] S. Tars and M. Fishel, "Multi-domain neural machine translation," in *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, J. A. Pérez-Ortiz, F. Sánchez-Martínez, M. Esplà-Gomis, M. Popović, C. Rico, A. Martins, J. Van den Bogaert, and M. L. Forcada, Eds., Alicante, Spain, May 2018, pp. 279–288. [Online]. Available: https://aclanthology.org/2018.eamt-main.26/

[67] H. Jiang, C. Liang, C. Wang, and T. Zhao, "Multi-domain neural machine translation with word-level adaptive layer-wise domain mixing," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association

for Computational Linguistics, Jul. 2020, pp. 1823–1834. [Online]. Available: https://aclanthology.org/2020.acl-main.165/

[68] J. Lee, H. Kim, H. Cho, E. Choi, and C. Park, "Specializing multi-domain NMT via penalizing low mutual information," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 10 015–10 026. [Online]. Available: https://aclanthology.org/2022.emnlp-main.680/

[69] W. Wang, Y. Tian, J. Ngiam, Y. Yang, I. Caswell, and Z. Parekh, "Learning a multi-domain curriculum for neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7711–7723. [Online]. Available: https://aclanthology.org/2020.acl-main.689/

[70] A. Sharaf, H. Hassan, and H. Daumé III, "Meta-learning for few-shot NMT adaptation," in *Proceedings of the Fourth Workshop on Neural Generation and Translation*, A. Birch, A. Finch, H. Hayashi, K. Heafield, M. Junczys-Dowmunt, I. Konstas, X. Li, G. Neubig, and Y. Oda, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 43–53. [Online]. Available: https://aclanthology.org/2020.ngt-1.5/

[71] E. Stergiadis, S. Kumar, F. Kovalev, and P. Levin, "Multi-domain adaptation in neural machine translation through multidimensional tagging," in *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, J. Campbell, B. Huyck, S. Larocca, J. Marciano, K. Savenkov, and A. Yanishevsky, Eds. Virtual: Association for Machine Translation in the Americas, Aug. 2021, pp. 396–420. [Online]. Available: https://aclanthology.org/2021.mtsummit-up.27/

[72] A. Cooper Stickland, A. Berard, and V. Nikoulina, "Multilingual domain adaptation for NMT: Decoupling language and domain information with adapters," in *Proceedings of the Sixth Conference on Machine Translation*, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, and C. Monz, Eds. Online: Association for Computational Linguistics, Nov. 2021, pp. 578–598. [Online]. Available: https://aclanthology.org/2021.wmt-1.64/

[73] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, and C. Xu, "Distilling object

detectors via decoupled features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2154–2164.

[74] A. Moslemi, A. Briskina, Z. Dang, and J. Li, "A survey on knowledge distillation: Recent advancements," *Machine Learning with Applications*, vol. 18, p. 100605, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827024000811

[75] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," 2015. [Online]. Available: https://arxiv.org/abs/1412.6550

[76] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7130–7138.

[77] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," 2019. [Online]. Available: https://arxiv.org/abs/1904.05068

[78] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

[79] D. L. Silver, Q. Yang, and L. Li, "Lifelong machine learning systems: Beyond learning algorithms," in *2013 AAAI spring symposium series*, 2013.

[80] W. Jooste, R. Haque, and A. Way, "Knowledge distillation: A method for making neural machine translation more efficient," *Information*, vol. 13, no. 2, p. 88, 2022.

[81] Y. Li, C. Zhao, and C. Caragea, "Improving stance detection with multi-dataset learning and knowledge distillation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6332–6345. [Online]. Available: https://aclanthology.org/2021.emnlp-main.511/

[82] A. Ganesan, F. Ferraro, and T. Oates, "Learning a reversible embedding mapping using bi-directional manifold alignment," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021,

pp. 3132–3139. [Online]. Available: https://aclanthology.org/2021.findings-acl.276/

[83] S. Wu and M. Dredze, "Do explicit alignments robustly improve multilingual encoders?" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4471–4482. [Online]. Available: https://aclanthology.org/2020.emnlp-main.362/

[84] K. Huang, X. Guo, and M. Wang, "Towards efficient pre-trained language model via feature correlation distillation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 16 114–16 128. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/34260a400e39a802961470b3d3de99cc-Paper-Conference.pdf

[85] P. O. Aboagye, Y. Zheng, M. Yeh, J. Wang, Z. Zhuang, H. Chen, L. Wang, W. Zhang, and J. Phillips, "Quantized Wasserstein Procrustes alignment of word embedding spaces," in *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, K. Duh and F. Guzmán, Eds. Orlando, USA: Association for Machine Translation in the Americas, Sep. 2022, pp. 200–214. [Online]. Available: https://aclanthology.org/2022.amta-research.15/

[86] S. Cao, N. Kitaev, and D. Klein, "Multilingual alignment of contextual word representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.03518

[87] J. Zhang, C. Dong, X. Duan, Y. Zhang, and M. Zhang, "Third-party aligner for neural word alignments," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3134–3145. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.228/

[88] P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados, and S. Reddy, "Llm2vec: Large language models are secretly powerful text encoders," 2024. [Online]. Available: https://arxiv.org/abs/2404.05961

[89] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

*Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6/

[90] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2214–2218. [Online]. Available: https://aclanthology.org/L12-1246/

[91] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of Machine Translation Summit X: Papers*, Phuket, Thailand, Sep. 13-15 2005, pp. 79–86. [Online]. Available: https://aclanthology.org/2005.mtsummit-papers.11/

[92] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 923–929. [Online]. Available: https://aclanthology.org/L16-1147/

[93] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4512–4525. [Online]. Available: https://aclanthology.org/2020.emnlp-main.365/

[94] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "Scaling neural machine translation to 200 languages," *Nature*, vol. 630, no. 8018, pp. 841–846, 2024. [Online]. Available: https://doi.org/10.1038/s41586-024-07335-x

[95] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified

text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204838007