

LB/TH/24/2025
TH5868

**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 Component 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 Component 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

Declaration of the Candidate & Supervisor	i
Dedication	ii
Acknowledgement	iii
Abstract	iv
Table of Contents	v
List of Figures	viii
List of Tables	ix
List of Abbreviations	ix
1 Introduction	1
1.1 Background and Motivation	1
1.2 Research Problem	2
1.3 Research Objectives	3
1.4 Contributions	4
1.5 Publications	4
1.6 Other Contributions	5
1.7 Thesis Structure	5
2 Literature Review	8
2.1 Overview	8
2.2 Neural Machine Translation (NMT)	8
2.2.1 An Overview of NMT	8
2.2.2 Common Architectures in NMT	9
2.3 Domain Adaptation for NMT	11
2.3.1 An Overview of Domain Adaptation	11
2.3.2 Challenges in Domain Adaptation	11
2.3.3 Multi-domain Adaptation Strategies	13
2.4 Knowledge Distillation in NMT	15
2.4.1 An Overview of Knowledge Distillation	15

2.4.2	Knowledge Distillation for NMT	15
2.4.3	Sequence-Level Distillation	16
2.5	Embedding Alignment	17
2.5.1	Alignment Methods In Natural Language Processing (NLP)	17
2.5.2	Cosine-based Alignment Methods	18
2.6	Summary and Research Gaps	19
3	Methodology	20
3.1	Motivation and Hypothesis	20
3.2	Initial Proposed Approach	20
3.2.1	Overview of the Initial Approach	20
3.2.2	Limitations of the Initial Approach	22
3.2.3	Transition to the Final Methodology	22
3.3	Overview of Our Final Approach	22
3.4	Limitations of Sequence-Level Distillation	23
3.5	Cosine-Based Encoder Alignment	26
3.6	Choice of Mean Pooling for Encoder Representations	26
3.7	Loss Function and Optimization	26
3.8	Constructing the Distilled Mixed Dataset (DMD)	27
3.9	Summary	27
4	Experimental Setup	29
4.1	Experimental Framework	29
4.2	Hardware and Software Configurations	29
4.2.1	Hardware Specifications	29
4.2.2	Software Specifications	29
4.3	Datasets	30
4.3.1	Dataset Selection and Preprocessing	30
4.3.2	Experimental Stages	30
4.4	Training and Inference Details	30
4.4.1	Model Selection	30
4.4.2	Training Setup	31
4.4.3	Inference and Generation	31

4.5	Model Configurations	31
4.6	Summary	32
5	Results and Discussion	33
5.1	Discussion on the Initial Approach	33
5.2	Discussion of our Final Approach	35
5.2.1	Hyperparameter Selection: Impact of α	35
5.2.2	Stage 1: Multi-Domain Performance and Generalization	36
5.2.3	Stage 2: Domain Adaptation Performance	36
5.3	Evaluation on Bona Fide Low-Resource Language	37
5.4	Summary of Findings	39
5.5	Alignment Between Objectives and Contributions	39
6	Conclusion	41
	References	43

LIST OF FIGURES

Figure	Description	Page
Figure 3.1	Illustration of the steps involved in the initial proposed approach.	21
Figure 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 <i>distill mixed dataset</i> .	23

LIST OF TABLES

Table	Description	Page
Table 4.1	Domain-specific datasets used in our experiments.	30
Table 4.2	Model configurations used in our experiments.	31
Table 5.1	ChrF scores for teacher models on different test domains. The lowest-performing model for each domain is bolded.	33
Table 5.2	Identified bad teachers for each domain.	34
Table 5.3	ChrF scores of models trained with various configurations, evaluated on in-domain and out-of-domain test sets.	34
Table 5.4	Comparison of pooling methods based on average ChrF score for model trained on single individual domain.	34
Table 5.5	ChrF scores of our model trained with different α values, evaluated on in-domain test sets (med, parl, law, news) and the out-of-domain flores development-test set.	35
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.	36
Table 5.7	ChrF scores for Stage 1 models fine-tuned on single domains (Open Subtitles and Ted2020) to evaluate domain adaptation.	37
Table 5.8	ChrF scores of our model trained on the English–Sinhala language pair with different α values using the distilled dataset, evaluated on three in-domain test sets and the out-of-domain Flores200 development-test set.	37
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.	38
Table 5.10	Alignment between research objectives, methodological approaches, and corresponding deliverables.	40

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 Trans- lation
RNN	Recurrent Neural Networks
SMT	Statistical Machine Translation