Multi-Domain Neural Machine Translation with Knowledge Distillation for Low Resource Languages

By : Velayuthan Menan
Reg no: 238043D
Supervisor : Dr. Nisansa de Silva
Co-Supervisor: Dr. Surangika Ranathunga
Department: Computer Science & Engineering

What to expect?

- Motivation
- Introduction & Research Problem
- Background Knowledge
- Initial Failure and Cause
- Final Approach
- Results & Discussion
- Achievements & Contributions

Motivation

- **Machine Translation (MT)**: Reduces language barriers and fosters global knowledge sharing.
- **Domain-Specific NMT**: High demand as general NMT[1] systems have limited applications [2].
- **Challenges**:
  - Domain shifts degrade performance (e.g., Legal → Biomedical).
  - These challenges are more severe for Low-Resource Languages (LRLs).
- **Reasons**:
  - **Word Uniqueness**: Some words are specific to a domain.
    Example: "Benzodiazepine" → Biomedical domain.
  - **Contextual Ambiguity**: Words can have different meanings across domains.
    Example: "Conductor" → Electrical Engineering vs. Music.
    Example: "Administer" → Medical (treatment) vs. Political (governance).

Introduction

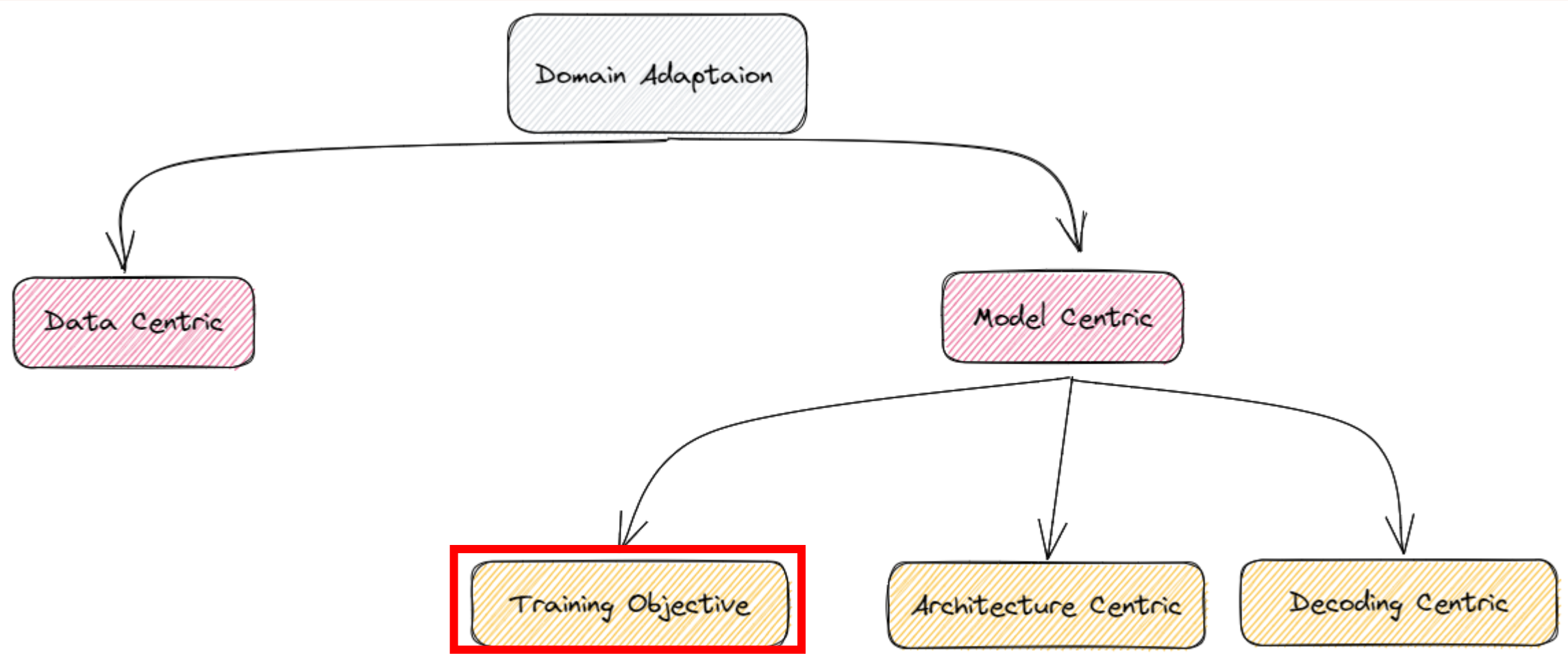
Neural Machine Translation

- MT is a problem creating a Probabilistic model. In essence we are finding a target sentence **y** given a source sentence **x** , mathematically put it is $argmax_y\ p(y|x)$ [1].
- In neural machine translation, we fit a parameterized model to maximize the conditional probability of sentence pairs using a parallel training corpus. (problem shifts to $argmax_{(\theta)}\ p(y|x)$)
- Although NMT has shown success when compared to it's predecessor Statistical Machine Translation (SMT)[3], their performance still falls short in low-resource and domain-specific adaptation scenarios[4].

Multi-Domain Adaptation

- Domain Adaptation is the problem of improving performance of a model trained on general domain data and test instances from a new domain [5].
- In Multi-Domain adaptation, we extend our approach by incorporating multiple domains into the system while ensuring that the model's performance remains consistently high across all domains.
- In this context, a "domain" is defined in accordance with the characterization provided in [2], which identifies a domain based on three distinct attributes.
  - Provenance is the source of the text.
  - Topic holds the subject of the text.
  - Genre stands orthogonal to topic, consisting of function, register, syntax and style.
- Domain adaptation has been a thoroughly researched field.

Categorization of Domain Adaptation [6]



Current Line of Thoughts

- Most of the promising results rely on either ensembling, a priori domain clustering in order to add domain tags and introducing a new domain specific gating vector [7].
- Other techniques delve into the architecture level, by introducing adapters in between layers to enhance adaptation [8].
- Another simple solution includes having a dedicated model for each domain.
- Methods mentioned above introduce either high model complexity or constraints such as prior knowledge of the domain the input text belongs.
- The best technique that does not add more complexity or prior classification, either using supervised or unsupervised methods, is based on fine-tuning on the concatenation of all in-domain data.

Fine Tuning

- Fine tuning has proven effectiveness in transferring between similar tasks [9], [10], [11].
- We can perform domain adaptation by fine tuning a model trained on large out-of-domain data with in-domain data.
- Two key issues fine tuning adheres to are,
  - Over-fitting when the in-domain data is small.
  - Catastrophic forgetting happens when out-of-domain translation is degraded.
- Authors Chu et al [13] show a simple yet effective way of overcoming the above-mentioned issues by "domain mixing" (here the out-of-domain data is mixed with in-domain data during fine tuning).

Challenges in Domain Adaptation

- Although Domain adaptation for NMT is a thoroughly researched field, but it is not a complete field (meaning there is not a single solution without tradeoffs)
- Challenge 1: Having dedicated models for different domain and different languages will provide a scalability issue.
- Challenge 2: As new domain gets added, we may have to train models from scratch, which requires data storages of previous domains and will incur high computational cost during training.
- Challenge 3: Most real-world datasets (web crawled datasets) may not be one hundred percent pure, there will be presence of other domains as well.
- We try to address Challenge 1 and 3 in our research.

Research Problem

How to design a Multi-Domain NMT system for Low Resource Languages which has the ability to scale well towards new domains while having low/no degradation of in performance on the existing domains ?

Background Knowledge

Knowledge Distillation in a nutshell

Linear Algebra

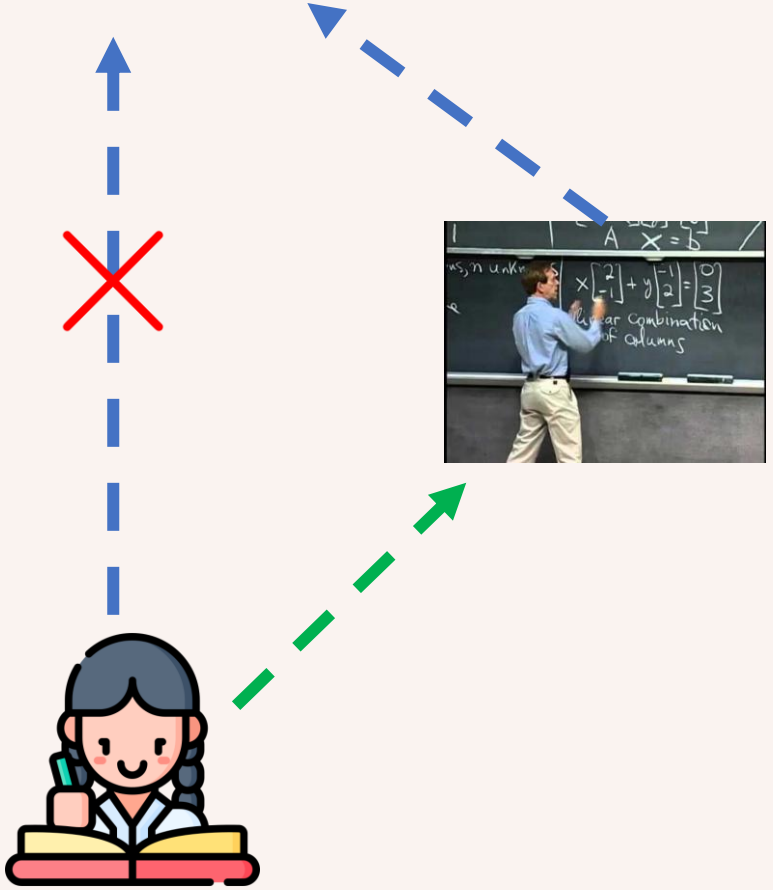
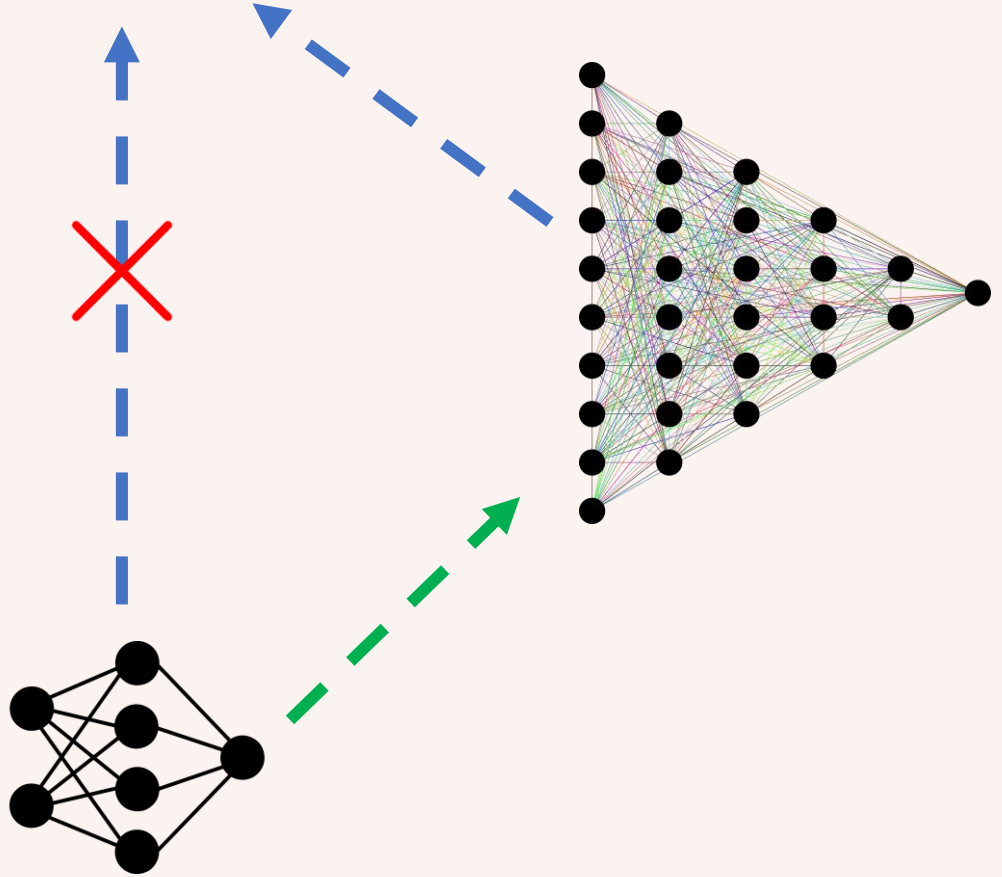


Image Classification



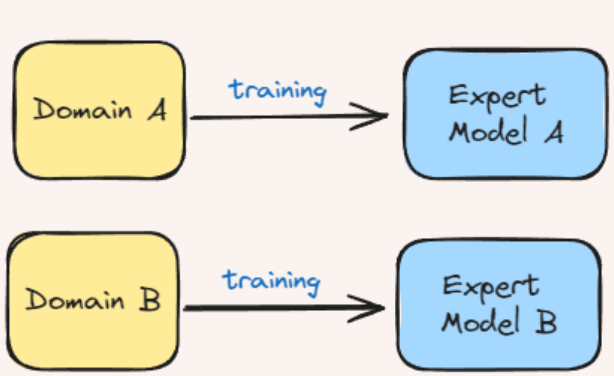
Initial Failure and Cause

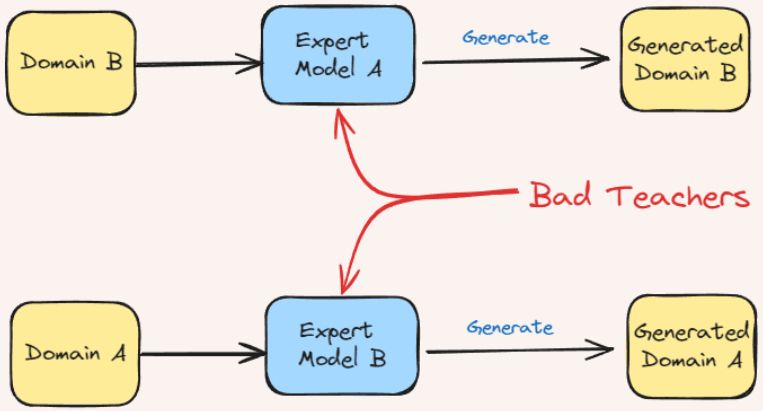
Initial Approach

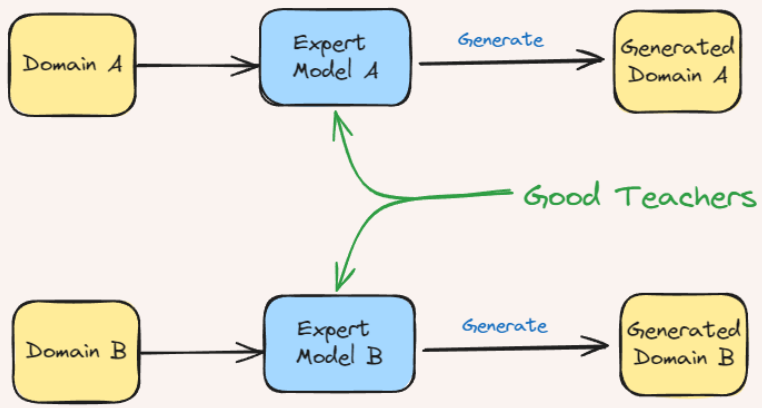
- Adding novelty to our approach, we try to answer the question "**Can models learn from bad teachers?**"
- Some domains can hinder the performance of other domains, they may not be compatible with them. In other words, two domains can have high divergence between each other.
- In our novel approach we try to reduce the impact of divergent domains on each other. We achieve this by making the models learn from bad teachers.
- We update our approach by,
  - Making bad teachers generate data for target domains. (will be further explained with illustration in the coming slides)
  - Making good teachers generate for their own domain.
  - Train the model using the data generated using embedding alignment.

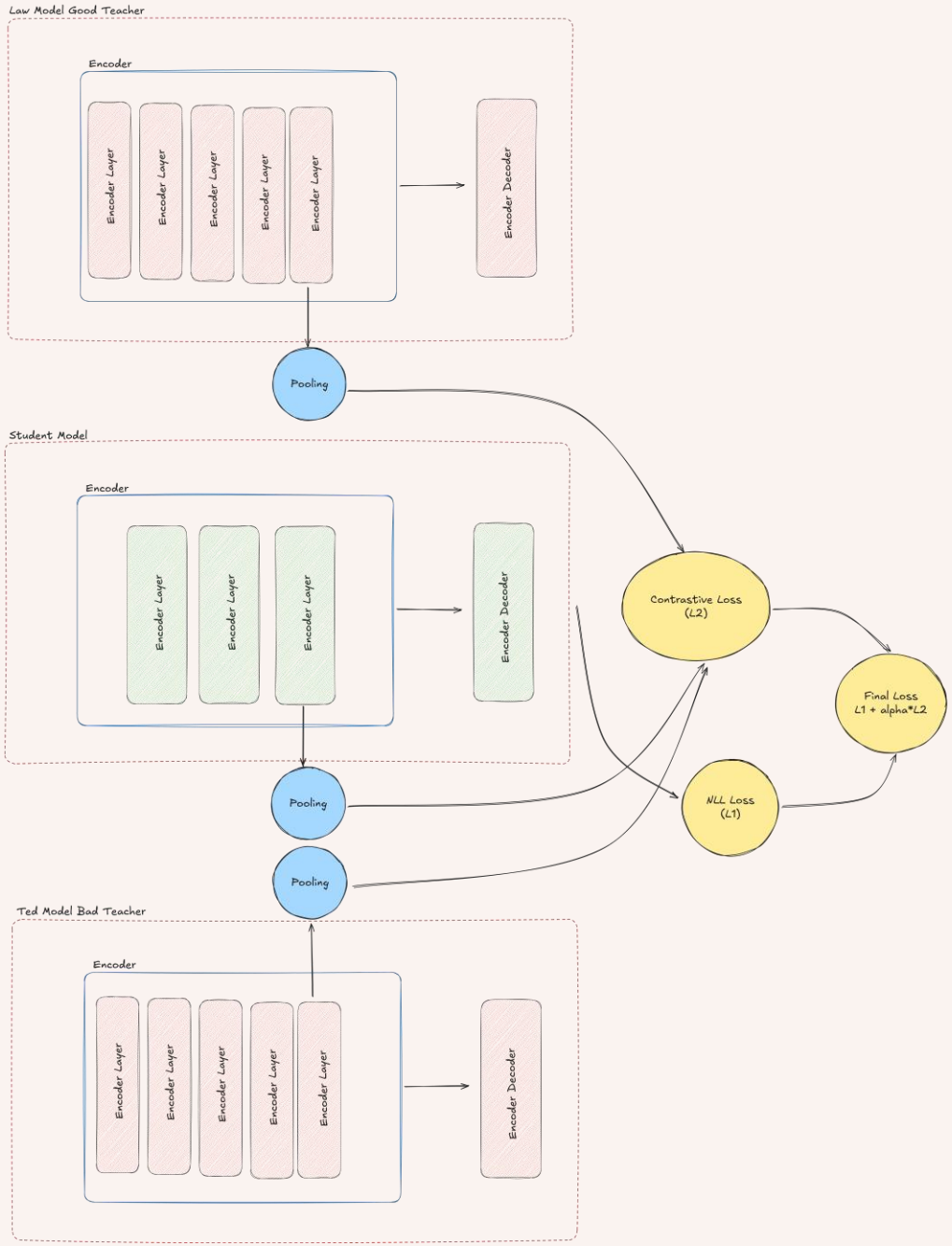
What are bad teachers and how do we find them?

- We define a bad teacher for a given domain (say A), as the expert model which is trained on a different domain (say B) and produces the least performance for the given domain (A).








Law Model Good Teacher
Encoder
Encoder Layer
Encoder Layer
Encoder Layer
Encoder Layer
Encoder Layer
Encoder Decoder
Pooling
Student Model
Encoder
Encoder Layer
Encoder Layer
Encoder Layer
Encoder Decoder
Pooling
Pooling
Contrastive Loss
(L2)
NLL Loss
(L1)
Final Loss
L1 + alpha*L2
Ted Model Bad Teacher
Encoder
Encoder Layer
Encoder Layer
Encoder Layer
Encoder Layer
Encoder Layer
Encoder Decoder

Results & Discussion

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

Domains	Bad Teacher
Law	Ted Teacher
Medicine	Ted Teacher
News	Medicine Teacher
Ted	Medicine Teacher

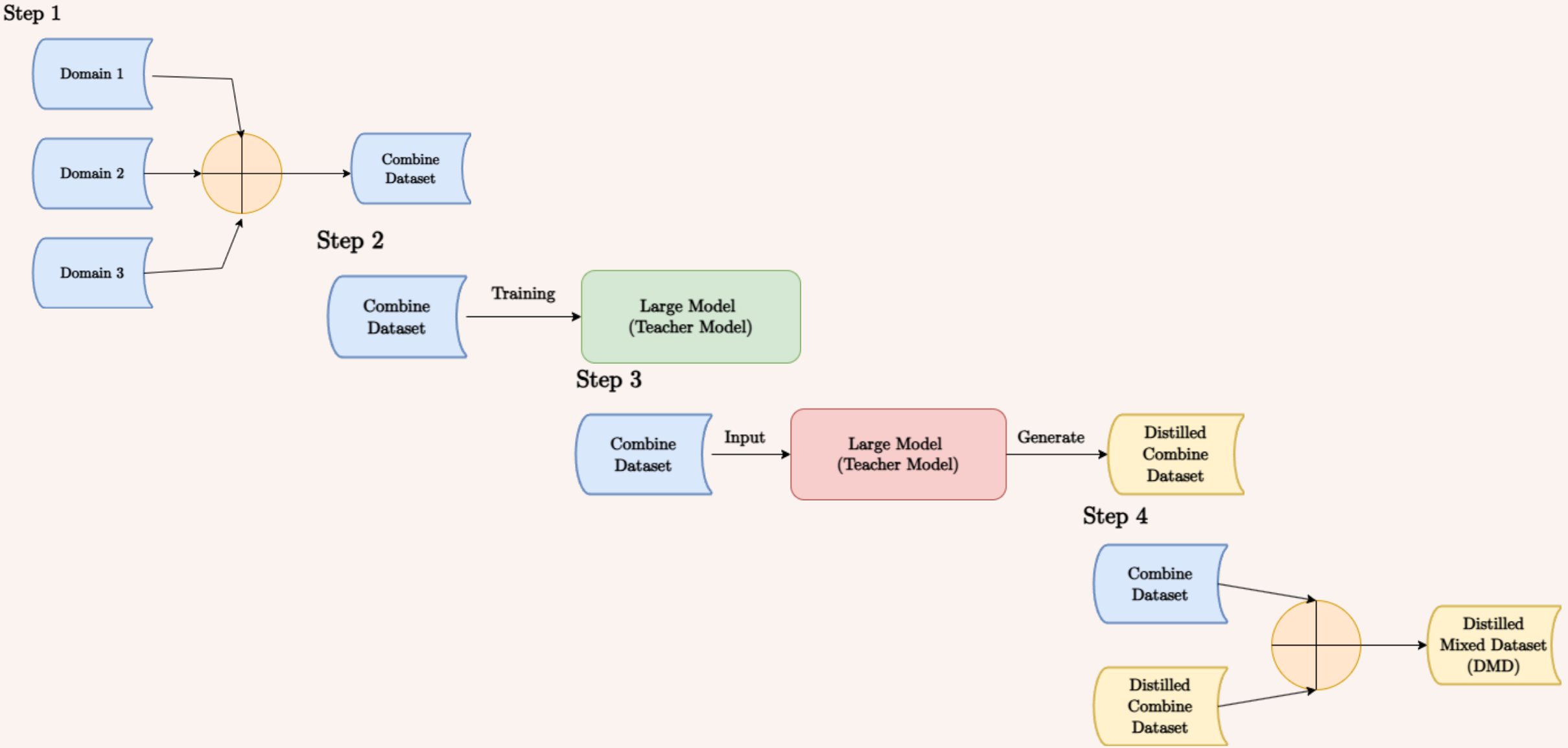
Model	med	parl	law	news	Flores
L-ADO	63.27	56.33	63.73	53.80	50.89
S-ADO	62.31	55.66	62.39	53.28	50.23
S-ADD	62.36	56.08	62.92	53.87	50.90
S-DMD-NoAlign	61.38	55.49	61.89	52.91	49.88
Our Approach	58.93	54.55	58.34	50.88	47.01

Cause For Failure

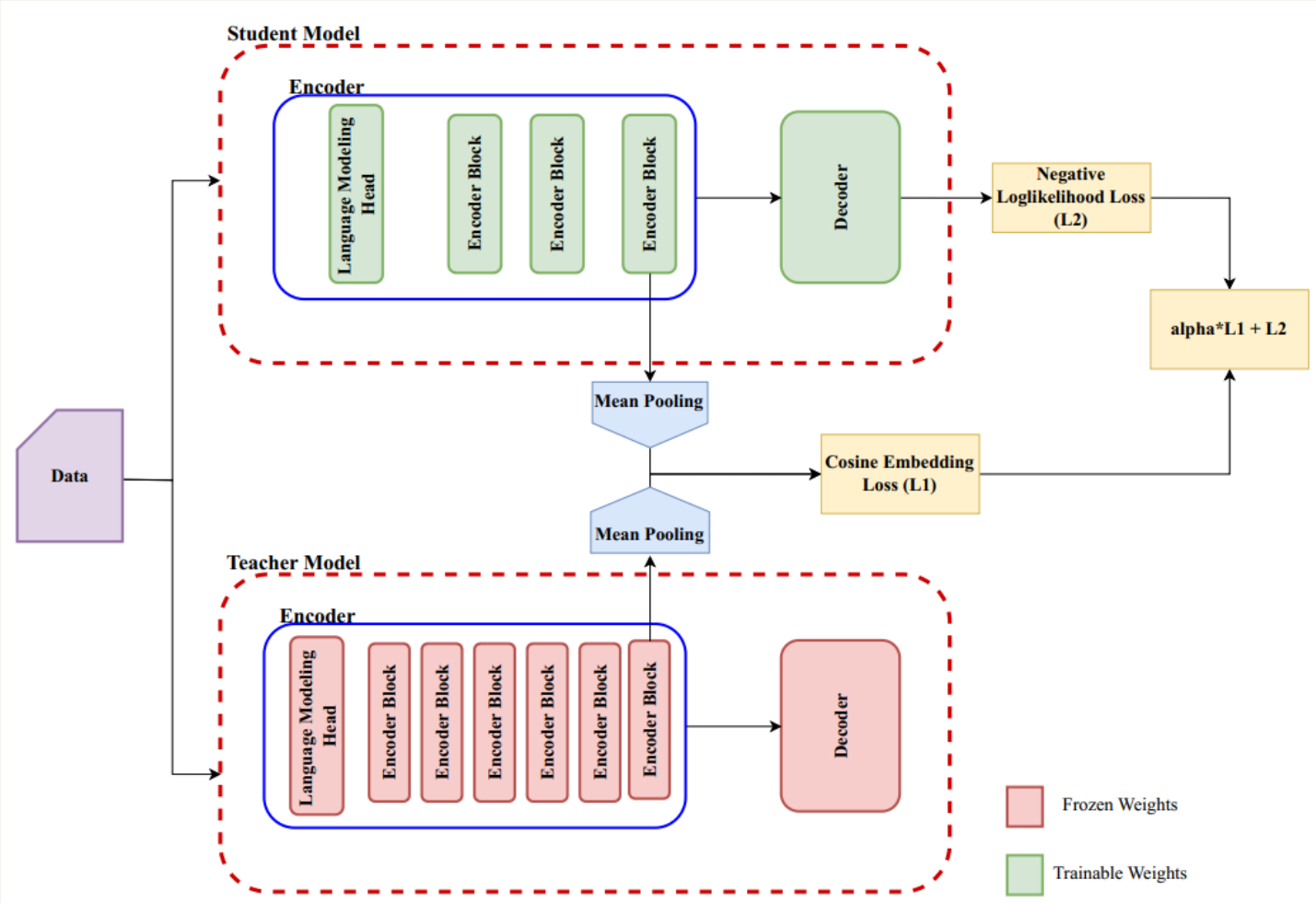
- “Too many cooks spoil the broth”
- Each domain has its own teacher and all teachers are trained independently of each other.
- This independence makes embedding space to be learnt independently.
- Due to this, the chance of getting overlapped embeddings are higher than getting embeddings which are contrastive (when we apply good teacher bad teacher contrastive learning).
- A solution will be finding objective function which trains all the teachers together in a contrastive manner during the teacher training phase.
  - This will be not be feasible in a low compute environment.

Final Approach

Distilled Mixed Dataset(DMD) Creation



Proposed Teacher-Student Encoder Alignment



$$L_{total} = \alpha \cdot L_1 + L_2$$

- Here α is the attenuation factor used to control the contribution of the cosine embedding loss.
- We chose mean pooling based on BehnamGhader et al. (2024).
- A cosine-based loss function (Barz and Denzler,2019) was used for encoder alignment between teacher and student.

Experimental Setup

We perform our experiment in two experimental setting

1. **Main study:** On German-English direction by emulating low resource language setting.
2. **Ablation study:** On bona fide low resource language setting using English-Sinhala direction.

In our Main study, we perform experiments in two stages,

- **Stage1**:
  - We train on 4 domains: europarl (Koehn, 2005), law (Tiedemann, 2012), medical (Tiedemann, 2012) and news commentary(Tiedemann, 2012).
  - We test on the above-mentioned domains and one out-of-domain dataset, Flores200 (NLLB Team et al., 2024).
- **Stage 2**:
  - We further fine-tune the models and baselines from Stage 1 on two new domains: Open Subtitles (Lison and Tiedemann, 2016) and TED2020 (Reimers and Gurevych, 2020).

Experimental Setup Contd..

In our ablation study

- For our ablation study we use a bona fide low resource language direction English-Sinhala.
- We study the α (hyperparameter) impact on the domain adaptation.
- For this experiment we,
  - We train 3 in-domain datasets : CCAligned (El-Kishky et al., 2020), OpenSubtitles (Lison and Tiedemann, 2016), and SITA (**gov**) (Fernando et al., 2020)
  - We test on the above-mentioned domains and one out-of-domain dataset, Flores200 (NLLB Team et al., 2024)

- All experiments were performed on randomly initialized T5-small (Raffel et al., 2019), architecture.
- All Teacher models had 6 encoder and decoder layers, while all student models had 3 encoder and decoder layers.
- All experiments we performed on single Nvidia QuadroRTX 6000 (24GB VRAM).
- Other details on the Experimental Setup can be found in Section 4 of our paper.

Naming Conventions

- **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 (vanilla sequence-level distillation (Kim and Rush, 2016)).
- **S-DMD-NoAlign**: Small model trained on the Distilled Mixed Dataset (DMD) without teacher-student encoder alignment (as followed in (Currey et al., 2020)).
- **S-DMD-Align**: Small model trained on the DMD with teacher-student encoder alignment(using the proposed methodology).

Main Study: Simulated Low Resource Setting (German-English)

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 2: ChrF scores of models trained with various configurations, evaluated on in-domain test sets (med, parl, law, news) and the out-of-domain Flores200 development-test set.

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 3: ChrF scores for Stage 1 models fine-tuned on single domains (Open Subtitles and Ted2020) to evaluate domain adaptation. Each model is fine-tuned on an individual domain and evaluated on its corresponding test set.

Ablation on α in a Real Low Resource Setting (English-Sinhala)

α	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 4: 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.

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: 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.

Achievements & Contributions

Results From All Progress Reviews

Review	Grade
Progress Review 1	Very Good
Progress Review 2	Very Good
Progress Review 3	Very Good

Publications During My Masters

- Title: Encoder-Aware Sequence-Level Knowledge Distillation for Low-Resource Neural Machine Translation
  - Authors: **Menan Velayuthan**, Dilith Jayakody, Nisansa de Silva, Aloka Fernando, Surangika Ranathunga.
  - Accepted at LoResMT @ NAACL
- Title: Back to the Stats: Rescuing Low Resource Neural Machine Translation with Statistical Methods.
  - Authors: **Menan Velayuthan**, Dilith Jayakody, Nisansa de Silva, Aloka Fernando, Surangika Ranathunga.
  - Accepted at WMT 2024 (h5-index 43)
- Title: Quality Does Matter: A Detailed Look at the Quality and Utility of Web-Mined Parallel Corpora
  - Authors: Surangika Ranathunga and Nisansa de Silva and **Menan Velayuthan** and Aloka Fernando and Charitha Rathnayake.
  - Accepted at EACL 2024(CORE Rank: A, h5-index 56).
  - Won the Best Low Resource Paper award.

Other Contributions (Selected)

- Conducted a Tech Talk Titled “How to act data rich in a data poor country?” at IESL under the guidance of Dr Nisansa De Silva and Prof Chandana Gamage.
- Successfully conducted “Sequence to Sequence Learning: Hands on Tutorial” at MERCon 2023.[Github Link]
  This was conducted by Dr. Uthaya, Dr. Nisansa, Sritharan Braveenan and myself.
- Successfully conducted the following sessions to Dr. Nisansa’s research group (2023-2025),
  - Distilling the Knowledge in a Neural Network. [Youtube],[Github]
  - Graph Neural Networks - Part 1. [Youtube],[Github]
  - Successfully conducted session to Dr. Nisansa’s research group and final year students on “Understanding Low Rank Adapters”[Youtube][Github Link]

Acknowledgement

- I would like to thank NLPC for providing me with GPU for all my experiments.
- I would like to thank my supervisors Dr Nisansa & Dr Surangika for their continued support and guidance.
- I would like to extend my gratitude to the Research Coordinator Dr Kutila and my Progress Reviews panelists and my Thesis Defense panelists for their time and guidance.

References

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. J. Artif. Int. Res., 75.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1317–1327, Austin,Texas. Association for Computational Linguistics.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Anna Currey, Prashant Mathur, and Georgiana Dinu.2020. Distilling multiple domains for neural machine translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4500–4511, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 2765–2781,Mexico City, Mexico. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. In First Conference on Language Modeling.

Björn Barz and Joachim Denzler. 2019. Deep learning on small datasets without pre-training using cosine loss. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1360–1369.

Thank you!