Encoder-Aware Sequence-Level Knowledge Distillation for Low-Resource Neural Machine Translation

Menan Velayuthan[1], Nisansa de Silva[1], Surangika Ranathunga[2]
[1]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[2]School of Mathematical and Computational Sciences, Massey University,
Auckland, New Zealand
{velayuthan.22,NisansaDdS}@cse.mrt.ac.lk
S.Ranathunga@massey.ac.nz

Abstract

Domain adaptation in Neural Machine Translation (NMT) is commonly achieved through fine-tuning, but this approach becomes inefficient as the number of domains increases. Knowledge distillation (KD) provides a scalable alternative by training a compact model on distilled data from a larger model. However, we hypothesize that vanilla sequence-level KD primarily distills the decoder while neglecting encoder knowledge, leading to suboptimal knowledge transfer and limiting its effectiveness in low-resource settings, where both data and computational resources are constrained. To address this, we propose an improved sequence-level KD method that enhances encoder knowledge transfer through a cosine-based alignment loss. Our approach first trains a large model on a mixed-domain dataset and generates a Distilled Mixed Dataset (DMD). A small model is then trained on this dataset via sequence-level KD with encoder alignment. Experiments in a low-resource setting validate our hypothesis, demonstrating that our approach outperforms vanilla sequence-level KD, improves generalization to out-of-domain data, and facilitates efficient domain adaptation while reducing model size and computational cost.

1 Introduction

Domain adaptation in Neural Machine Translation (NMT) has been extensively studied (Saunders, 2022), but significant gaps remain, particularly for low-resource languages (Ranathunga et al., 2023). A common approach involves fine-tuning general-domain pre-trained models on specific target domains (Chu and Wang, 2018). 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.

To address these challenges, research has increasingly focused on developing a single model capable of handling multiple domains (Liang et al., 2024; Pan et al., 2021; Currey et al., 2020; Pham et al., 2019). One promising direction is knowledge distillation (KD) (Hinton, 2015), where domain-specific teacher models (expert models), typically deep neural networks, generate distilled target-side data for each domain. Currey et al. (2020) 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 may not generalize well to low-resource languages, as training domain-specific teacher models with limited data is inherently suboptimal. Furthermore, existing approaches heavily rely on sequence-level distillation (Kim and Rush, 2016), but its effectiveness in low-resource settings remains unclear. We hypothesize that sequence-level KD primarily distills the decoder while transferring limited encoder knowledge, leading to suboptimal knowledge transfer in encoder-decoder architectures (NLLB Team et al., 2024; Liu et al., 2020; Xue et al., 2021; Mohammadshahi et al., 2022). If this hypothesis holds, overlooking encoder knowledge may hinder the adaptability of distilled models across multiple domains. To address this, we propose a cosine-based alignment between the teacher and student encoders to enhance the effectiveness of distillation.

Building on the methodologies of Currey et al. (2020) and Chu et al. (2017), we propose a mixed-domain KD approach for low-resource NMT. Chu et al. (2017) demonstrated that training a single

Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), pages 161–170

model on a mixture of domains can facilitate domain adaptation in low-resource scenarios. Inspired by this, we first train a single teacher model (with six encoder and decoder layers) on a mixed-domain dataset. The teacher then generates distilled target text by decoding the source text from the original dataset. Combining these distilled sentence pairs with the original dataset, we construct a *Distilled Mixed Dataset* (DMD), following the approach of Currey et al. (2020).

Using this DMD, we train a randomly initialized student model (with three encoder and decoder layers) through sequence-level distillation, augmented with our proposed teacher-student encoder alignment (Stage 1). In Stage 2, we fine-tune the distilled models on two new domains to assess their adaptability.

Our contributions are as follows:

1. We validate the hypothesis that *sequence-level distillation primarily distills the decoder, potentially leading to suboptimal knowledge transfer*, and demonstrate that our encoder alignment method effectively mitigates this limitation.

2. We show that models trained using our proposed methodology generalize better to out-of-domain data compared to vanilla sequence-level distillation and its successful data-centric extensions, such as that of Currey et al. (2020).

3. We demonstrate that models trained using our methodology adapt effectively to new domains when fine-tuned, outperforming those trained using vanilla sequence-level distillation and its data-centric extensions, such as Currey et al. (2020).

4. We perform our experiments in a resource-constrained setting, with a limited training dataset of 50K sentences and a compute-poor environment, and demonstrate that our proposed method performs effectively in this scenario.

5. We conduct an ablation study on the alignment loss weighting parameter α in a bona fide low-resource language setting (English–Sinhala), and find that the optimal value is dataset-dependent. Within the tested range ($\alpha = 1$ to $\alpha = 7$), a moderate setting yields the best in-domain performance for this specific dataset.

2 Background

2.1 Domain Adaptation for NMT

Catastrophic forgetting (Goodfellow et al., 2015; Chu et al., 2017), the degradation of model performance on previous tasks when continually trained on new tasks, is a major challenge in adapting new domains to an existing NMT system. A simple yet effective approach to address this issue was demonstrated by Chu et al. (2017) and Liang et al. (2023), where the model is re-trained from scratch using data from all domains. While this method may not scale well as the number of domains increases, due to the requirement of having all domain data available, it could still be feasible in low-resource scenarios where datasets are typically small. However, adapting a single domain is insufficient for applications requiring NMT systems to handle multiple domains. Maintaining separate expert systems for each domain becomes prohibitively expensive as the number of domains grows. Multi-domain NMT systems offer a more scalable alternative (Wu et al., 2024; Pham et al., 2019; Britz et al., 2017).

Approaches to domain adaptation can be broadly categorized into two types: (i) data-centric methods (Kim and Rush, 2016; Currey et al., 2020; Liu et al., 2021; Ko et al., 2021), which focus on leveraging data to improve adaptability, and (ii) model-centric methods (Bapna and Firat, 2019; Aharoni and Goldberg, 2020; Escolano et al., 2021; Cao et al., 2021), which involve modifications to model architectures or training processes. In our work, we adopt a hybrid approach. We draw inspiration from data-centric methods such as Currey et al. (2020) and Kim and Rush (2016), while incorporating a novel model-centric alignment method to enhance KD for NMT.

2.2 Knowledge Distillation for NMT.

Knowledge Distillation (KD) (Hinton, 2015), a framework for transferring knowledge from a teacher model (a large, slow model) to a student model (a smaller, faster model), has become a widely adopted technique for compressing models into more efficient forms. Its impact in NLP became evident with the introduction of DistilBERT (Sanh et al., 2019; Jiao et al., 2020), a compressed version of BERT (Devlin et al., 2019). Despite its success, KD presents challenges in

sequence-to-sequence tasks. To address these challenges, Kim and Rush (2016) extended KD to Neural Machine Translation (NMT) by introducing sequence-level distillation, which enables KD to be applied effectively to sequence-to-sequence learning tasks (Sutskever et al., 2014).

Leveraging sequence-level distillation, Currey et al. (2020) demonstrated its utility for multi-domain adaptation in NMT. Their work showed that a single student model could outperform larger teacher models dedicated to individual domains. Furthermore, Liang et al. (2024) investigated domain adaptation using continual learning (Silver et al., 2013), where they incrementally expanded the domain coverage of an existing translation model. In this context, sequence-level distillation proved instrumental in retaining knowledge of older domains while adapting to new ones. However, these methods have not been extensively studied in resource-constrained environments, such as compute-poor settings, low-resource languages, or combinations of both. In our work, we specifically target both scenarios together, addressing the challenges of compute constraints and low-resource languages simultaneously

3 Methodology

Our methodology is based on the hypothesis that *sequence-level distillation primarily distills the decoder, potentially leading to suboptimal knowledge transfer*. To address this, we propose aligning the student encoder with the teacher encoder using a cosine-based loss function (Barz and Denzler, 2019), improving upon vanilla sequence-level distillation. Our approach consists of the following steps:

- **Step 1**: Randomly initialize a large teacher model and train it to convergence on the given data.
- **Step 2**: Use the teacher model to decode the source side of the training data, generating "distilled" target data.
- **Step 3**: Combine the distilled data with the original data to create the *distill mixed dataset* (DMD).
- **Step 4**: Randomly initialize a smaller student model and train it on the DMD to convergence, applying cosine embedding loss to align the student encoder with the teacher encoder.

The training schema is illustrated in Figure 1. As shown in the figure, the output of the final layer of the encoder is passed through a mean pooling unit to obtain a single vector representation. We chose mean pooling because BehnamGhader et al. (2024) demonstrated that it performs best for sequence-to-sequence models.

The final loss consists of two components: (1) the cosine embedding loss (L_1) and (2) the negative log-likelihood loss (L_2). Since the DMD is used, L_2 includes the standard negative log-likelihood loss for NMT (L_{NLL}) and the sequence-level negative log-likelihood loss ($L_{\text{SEQ-KD}}$). For details on sequence-level negative log-likelihood loss, refer to Section 3.2 of Kim and Rush (2016). The total loss is defined as:

$$\text{L}_{\text{total}} = \alpha \cdot L_1 + L_2$$

Here, α is an attenuation factor used to control the contribution of the cosine embedding loss. α is treated as a hyperparameter, and its value is varied in increments of 0.5 to determine the optimal setting.

4 Experimentation Details

4.1 Datasets

For our experiments, we selected the German-to-English language direction and created training, development, and test sets using six domains: **europarl** (**parl**) (Koehn, 2005), **law** (JRC-Acquis) (Tiedemann, 2012), **medical** (**med**) (EMEA corpus) (Tiedemann, 2012), **news commentary** (**news**) (Tiedemann, 2012), **open subtitles** (**opensub**) (Lison and Tiedemann, 2016), and **Ted2020** (**ted**) (Reimers and Gurevych, 2020). All datasets were sourced from OPUS[1] (Tiedemann, 2012).

For our ablation study on the weighting parameter α, we utilize a real low-resource language setting: English–Sinhala. We select three available domains for evaluation—**CCAligned** (ccalign) (El-Kishky et al., 2020), **OpenSubtitles** (opensub) (Lison and Tiedemann, 2016), and **SITA** (gov) (Fernando et al., 2020), the latter of which was constructed from government documents of Sri Lanka.

To emulate a low-resource setting, we randomly sampled and deduplicated 50K sentences for the training set and 1K sentences each for the development and test sets. Sentences were selected with

[1] https://opus.nlpl.eu/

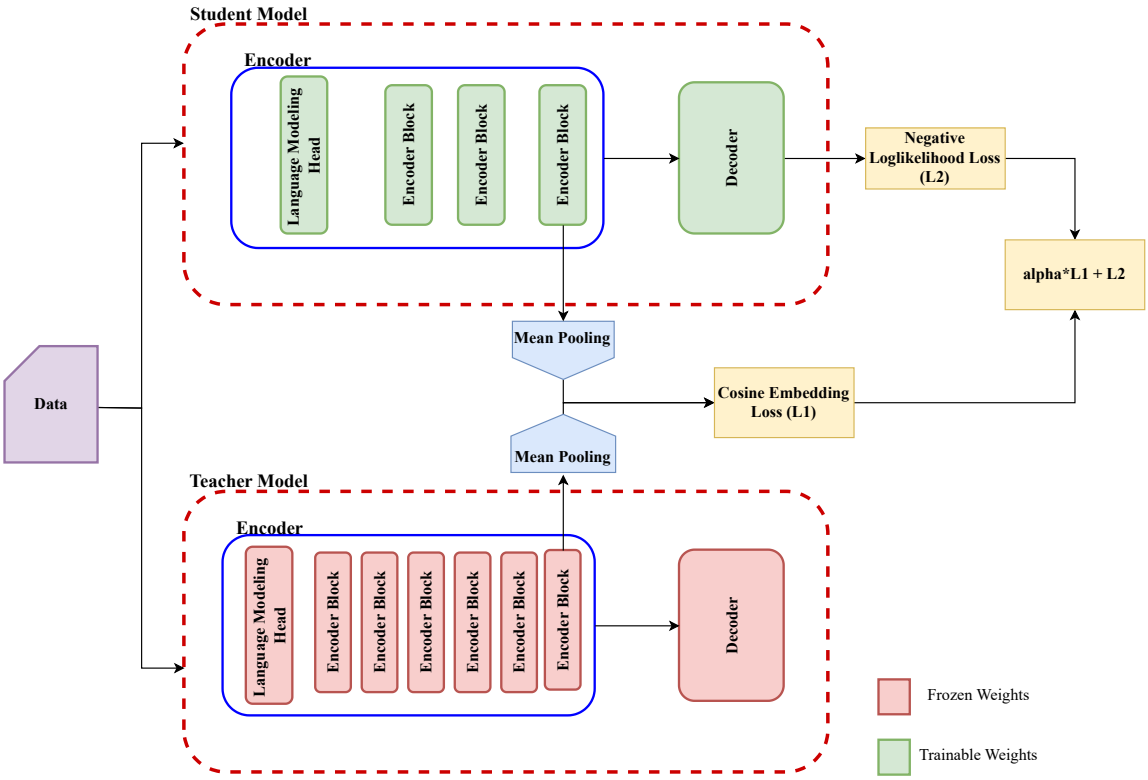

Figure 1: The figure illustrates our proposed method. The teacher model's weights are frozen, while the student model's weights are learnable. The final loss is computed as the sum of two components: (1) the cosine embedding loss between the mean-pooled final encoder layer outputs of the teacher and 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* (refer to §3).

word counts between 4 and 120.

The experiments were conducted in two stages:

1. In Stage 1, we tested our proposed improvement to sequence-level distillation. The training data consisted of a combined dataset of *parl*, *law*, *news*, and *med*. Additionally, we evaluated the generalization ability of these models to out-of-domain data using the Flores200(Flores) (NLLB Team et al., 2024) development-test set.

2. In Stage 2, we assessed whether distilled models perform better in domain adaptation by fine-tuning the models from Stage 1 on *opensub* and *ted* datasets. We followed the standard vanilla fine-tuning method in this stage.

4.2 Model Configurations

To evaluate the impact of our proposed methodology, we trained five different models under varying configurations:

- **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).

All-Domain refers to the combined dataset, formed by concatenating all available domains. L-ADO serves as the teacher model for all small models that utilize KD. The Distilled Mixed Dataset (DMD) is constructed by combining the original

dataset with the distilled dataset, where the latter is generated by the teacher model decoding the original source text. The large models consist of six encoder and decoder layers, while the small models have three.

4.3 Implementation Details

Hardware Specifications: All experiments were conducted on a single machine with an Intel i9-9900K CPU, 64GB of RAM, and an Nvidia Quadro RTX 6000 (24GB VRAM).

Software Specifications: All models and training code were developed using the HuggingFace (HF) Transformers (Wolf et al., 2020) library. For evaluation, we use chrF score from the `evaluate`[2] library of HF.

Models: To emulate a compute-constrained environment and meet the requirement of having the teacher model present during student training, we selected T5-small (77M parameters), a variant of the T5 model (Raffel et al., 2019), as the teacher. The student model was created by halving T5-small, removing three layers each from the encoder and decoder. All models were randomly initialized to avoid influence from pre-trained knowledge.

Training Details: All models were trained for a maximum of 100 epochs with early stopping (patience: 4). A learning rate of 2×10^{-4} was used across all experiments. The batch size was set to 64, with gradient accumulation of 2, resulting in an effective batch size of 128. The maximum sequence length was set to 120.

Inference & Generation Details: During inference, a maximum sequence length of 120 was used, and beam search (Graves, 2012) with a beam size of 5 was employed. The same settings were applied for generating target data for KD.

α	**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 1: 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 Flores200 development-test set.

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

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. Refer to §4.2 for configuration naming conventions.

5 Results and Discussion

To evaluate the effectiveness of our proposed approach, we conduct experiments in two distinct settings. The primary evaluation is carried out on the German–English language pair in a simulated low-resource setting (§5.1), where we have access to multiple domains and can systematically control the amount of training data. This setup allows us to isolate and analyze the core contributions of our method under controlled conditions. Additionally, we conduct an ablation study on the English–Sinhala language pair to examine the effect of the weighting parameter α under a bona fide low-resource scenario with limited parallel data and minimal domain diversity (§5.2). This analysis helps assess the robustness of our method when applied to real-world low-resource language settings.

5.1 Main Results: Simulated Low-Resource Setting

Table 1 shows a general trend of increasing ChrF scores across all domains as the attenuation factor (α) increases, except for the law domain and the out-of-domain Flores200 test set. Based on these results, we select $\alpha = 2.0$ as the optimal setting. Given the extensive training time required for each experiment (12+ hours per run), we evaluated α within a limited range.

Table 2 demonstrates that our proposed alignment methodology, used in S-DMD-Align, consistently outperforms all small models across different configurations, including the standard sequence-level distillation model (S-ADD) (Kim and Rush, 2016). Notably, S-DMD-Align also surpasses the large model (L-ADO), reaffirming Hinton (2015)'s observation on the effectiveness of knowledge distillation. These results validate our hypothesis that *sequence-level distillation primarily distills the decoder, leading to suboptimal knowledge transfer*.

Our approach achieves superior performance

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. Refer to §4.2 for configuration naming conventions.

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.

across all domains while utilizing a more compact small model with half the encoder and decoder layers, making it well-suited for deployment in resource-constrained environments. In contrast, the S-ADD model underperforms compared to the large model (L-ADO), indicating that vanilla sequence-level distillation leads to inadequate knowledge transfer. Interestingly, S-ADD achieves results comparable to the straightforward combined-domain training approach (S-ADO), suggesting that S-ADO is a more practical alternative to S-ADD due to its simpler training process.

Beyond multi-domain performance, our method enhances generalization to out-of-domain data, as evidenced by the Flores evaluation. The S-DMD-Align model surpasses the large model (L-ADO) by **+2.01** and the student model S-ADD by **+2.67** on Flores200, demonstrating that knowledge distillation can effectively enable multi-domain adaptation even in low-resource and compute-constrained settings.

To further assess the domain adaptation capability of the distilled models, we fine-tune the models from Stage 1 (refer to §4.1) on individual domains (opensub and ted). Table 3 presents the results of this Stage 2 experiment. Consistent with our findings from Stage 1, S-DMD-Align outperforms all baselines, including the large model (L-ADO), confirming that distilling models using our proposed methodology, followed by fine-tuning, enhances domain adaptation.

5.2 Ablation on α in a Real Low Resource Setting

In this ablation study, we examine the effect of the weighting parameter α in our proposed method using the English–Sinhala language pair, which represents a bona fide low-resource scenario. Due to the scarcity of parallel data and domain coverage, we limit our evaluation to three in-domain datasets and one out-of-domain test set. Since our approach is primarily designed as an extension of vanilla sequence-level knowledge distillation, we consider both the standard sequence-level distillation (S-ADD) and the no-distillation baseline as points of comparison. This focused setup allows us to analyze the relative contribution of α under practical low-resource constraints.

Table 4 reports ChrF scores across various values of the weighting parameter α, evaluated on three in-domain datasets (ccalign, opensubs, and gov) and one out-of-domain dataset (Flores). For the purpose of hyperparameter tuning, we use only the distilled dataset, as it is smaller and more efficient to experiment with. Once the optimal value of α is identified, we retrain the model on the full DMD dataset for comparison against standard baselines (Table 5). We observe that there is no consistent or monotonic trend as α increases—performance fluctuates across domains rather than improving uniformly. Nonetheless, the choice of $\alpha = 4.0$ consistently yields the best in-domain performance across all three datasets, indicating that our proposed objective benefits from higher alignment weighting in low-resource, domain-limited settings.

Interestingly, the out-of-domain Flores evaluation achieves its highest score at $\alpha = 1.0$, where the NMT loss and the cosine embedding alignment loss are equally weighted. Moreover, the optimal value of α appears to be dataset-dependent, as our earlier experiments on the German–English pair revealed a different sensitivity to α compared to the English–Sinhala setting explored in this study.

We pick two optimal α values (1 and 4) from the results in Table 4 and evaluate them against the

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. Refer to §4.2 for a description of the configuration naming conventions.

standard baselines, including sequence-level distillation (S-ADD). In Table 5, we find that our model with $\alpha = 4$ outperforms all baselines by a considerable margin, except in the gov domain, where L-ADO marginally outperforms our model by 0.24 ChrF points. It is also worth noting that vanilla sequence-level distillation (S-ADD) lags behind the other baselines, suggesting that this approach is suboptimal in resource-constrained settings.

When comparing $\alpha = 1$ and $\alpha = 4$, although $\alpha = 1$ yielded the best performance on Flores during hyperparameter tuning, we observe that when trained on the DMD dataset, $\alpha = 4$ consistently outperforms $\alpha = 1$ across all domains. This highlights the importance of utilizing a combined data setting (original + distilled) during the training of an alignment-based model. Finally, we observe that in the low-resource language direction, greater importance is attributed to encoder alignment, as reflected by the superior performance of the model with $\alpha = 4$.

6 Conclusion

Our experiments validate the hypothesis that *sequence-level distillation primarily distills the decoder, potentially leading to suboptimal knowledge transfer*, and demonstrate the efficacy of our proposed methodology in addressing this limitation. We show that vanilla sequence-level distillation often produces underperforming models, making it unsuitable for resource-constrained settings involving low-resource languages and compute-poor environments. In contrast, models trained using our approach exhibit superior generalization to out-of-domain data and demonstrate enhanced capabilities as domain adapters.

By conducting all experiments on a single machine under constrained resources, we highlight the practicality of our method and aim to inspire further research on domain adaptation through knowledge distillation, even in resource-limited settings. Our findings underline the potential of knowledge distillation as a viable strategy for achieving effective multi-domain adaptation. Furthermore, our ablation study on the alignment loss weighting (α) in the English–Sinhala direction reveals that the optimal setting is dataset-dependent, with each language direction benefiting from a different choice of α.

7 Limitations

Our experiments were conducted under strict resource constraints, including a compute-poor environment and a limited training dataset of 50K sentences. While this setting highlights the practicality and efficiency of our method, it may limit the generalizability of our findings to larger-scale or high-resource scenarios. Additionally, our approach has only been evaluated on models trained from scratch; it remains an open question whether the proposed alignment objective would yield similar benefits when applied to pretrained models or larger architectures. We leave this exploration for future work.

Acknowledgements

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

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. 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)*, pages 1921–1931, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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*.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark. Association for Computational Linguistics.

Yue Cao, Hao-Ran Wei, Boxing Chen, and Xiaojun Wan. 2021. 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*, pages 3964–3974, Online. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. 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)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

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

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Carlos Escolano, Marta R. Costa-Jussà, and José A. R. Fonollosa. 2021. From bilingual to multilingual neural-based machine translation by incremental training. *Journal of the Association for Information Science and Technology*, 72(2):190–203.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. *arXiv preprint arXiv:2011.02821*.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2015. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Preprint*, arXiv:1312.6211.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *Preprint*, arXiv:1211.3711.

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

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

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.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. 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)*, pages 802–812, Online. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Yunlong Liang, Fandong Meng, Jiaan Wang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2024. 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)*, pages 10914–10928, Bangkok, Thailand. Association for Computational Linguistics.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2023. Unified model learning for various neural machine translation. *Preprint*, arXiv:2305.02777.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8348–8359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. 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)*, pages 244–258, Online. Association for Computational Linguistics.

MinhQuang Pham, Josep Crego, François Yvon, and Jean Senellart. 2019. Generic and specialized word embeddings for multi-domain machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

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

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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

Daniel L Silver, Qiang Yang, and Lianghao Li. 2013. Lifelong machine learning systems: Beyond learning algorithms. In *2013 AAAI spring symposium series*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Junhong Wu, Yuchen Liu, and Chengqing Zong. 2024. F-malloc: Feed-forward memory allocation for continual learning in neural machine translation. *arXiv preprint arXiv:2404.04846*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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, pages 483–498, Online. Association for Computational Linguistics.