SemiAdapt: Semi-Supervised and Efficient LoRA-Based Domain Adaptation for Low-Resource Irish Machine Translation with Transformers

Josh McGiff* **and Nikola S. Nikolov**
TEIC Lab
Department of Computer Science and Information Systems
University of Limerick, Ireland
*josh.mcgiff@ul.ie

Abstract

Fine-tuning is widely used to adapt multilingual Transformer models for machine translation (MT) in specific domains. However, full-parameter fine-tuning of large multilingual models with billions of parameters is computationally expensive, thus creating a barrier to entry for researchers working on low-resource tasks such as Irish translation. Parameter-efficient fine-tuning (PEFT) addresses this by updating a fraction of the original model parameters, with the Low-Rank Adaptation approach (LoRA) introducing small, trainable adapter layers. We introduce SemiAdapt-Full and SemiAdapt-LoRA as semi-supervised approaches that leverage inferred domains to improve overall performance in MT. SemiAdapt-LoRA employs dynamic routing at inference time, eliminating the need to load multiple separately fine-tuned models. Instead, a single shared base model is maintained while lightweight domain-specific adapters, updating only 1.39% of the model parameters in our case, are activated dynamically. We demonstrate that SemiAdapt-Full can outperform full-model fine-tuning and SemiAdapt-LoRA can propel PEFT methods to compete with full-model fine-tuning. We further evaluate corpus-level domain fine-tuning and demonstrate that our embedding-based inference methods perform especially well on larger and noisier corpora. Code and training configurations are released to support reproducibility. Ultimately, our approach narrows the performance gap between PEFT and full-parameter fine-tuning, offering resource-constrained researchers a computationally efficient alternative.

Keywords: Parameter-Efficient Fine-Tuning, Low-Rank Adaptation, Neural Machine Translation, Domain Adaptation, Low-Resource Languages

1. Introduction

Since their introduction, Transformer architectures (Vaswani et al., 2017) have reshaped the landscape of natural language processing (NLP), achieving substantial improvements in diverse language modeling tasks such as text generation (Lewis et al., 2020). Furthermore, Transformer-based architectures have revolutionised neural machine translation (NMT), particularly for low-resource languages (LRLs) (Meta AI, 2024).

NMT is the process of using a neural network to encode a sentence in a source language and to decode a corresponding translated sentence in a target language (Bahdanau et al., 2014). Although rule-based (Castilho et al., 2017; España-Bonet and Costa-jussà, 2016) and statistical machine translation (Bojar et al., 2015; España-Bonet and Costa-jussà, 2016) has had historical success, NMT now dominates as the prevailing paradigm (Castilho et al., 2018).

Despite constant advancements in the field of natural language generation, NMT remains the most extensively studied generative language modelling task in LRL communities (McGiff and Nikolov, 2025). In an era where these advancements are empowering services such as ChatGPT and Claude, the prevalence of NMT in these language communities indicates that they are stuck playing catch-up. The domination of English language content on the internet (Common Crawl, 2025) extends to the widely popular generative services available online (Robinson et al., 2023). This existing bias in the data enables research and development for majority languages such as English, and presents major challenges for less-represented languages (McGiff and Nikolov, 2025). Given the large amount of unstructured data required to build coherent generative language models for a language, popular services such as ChatGPT are unable to accurately represent minority languages and their communities. Furthermore, the general lack of literature evaluating LLMs for these languages indicates that these language communities are often an afterthought for LLM-powered tools and services.

In the context of Irish language translation, previous work has assessed various statistical and NMT methods on their capacity to translate from English to Irish and vice versa (Dowling et al., 2018; Lankford et al., 2022b; Defauw et al., 2019). However, translating from English to Irish poses a greater challenge for large, multilingual models than the reverse direction (Tran et al., 2024; Lankford et al., 2022a). This asymmetry arises because models such as NLLB-200 (Meta AI, 2024) are trained to generate English text across hundreds of language pairs, giving the decoder extensive exposure to English, while LRLs such as Irish are rarely used as tar-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10208–10220
11–16 May 2026.

get languages and are typically only decoded in the English to Irish direction (Costa-Jussà et al., 2022). Consequently, the model's decoder learns richer and more stable representations for English than for Irish, thus making generation into Irish more difficult. These issues are further compounded by the Irish language's morphological complexity, vocabulary sparsity, and limited parallel data (Grönroos et al., 2020). Addressing these challenges in English to Irish translation remains an under-explored research area.

As a result, we introduce SemiAdapt-Full and SemiAdapt-LoRA as two efficient approaches for improving English to Irish translation. SemiAdapt-Full and SemiAdapt-LoRA are semi-supervised methods that involve zero-shot domain assignment to training data, followed by either full-model fine-tuning or the training of low-rank adaptation (LoRA) adapters. At inference time, domain embedding centroids are used to efficiently assign domains. We find that SemiAdapt-LoRA outperforms full-model fine-tuning on some domains and enables LoRA-based methods to achieve performance comparable to full-model fine-tuning. Although full-model fine-tuning outperforms SemiAdapt-LoRA on some domains, SemiAdapt-LoRA demonstrates that it can also benefit from our semi-supervised, inference-efficient strategy. The primary objective of this study is not to benchmark all domain adaptation techniques, but to examine whether semi-supervised domain routing combined with lightweight LoRA adapters can match or exceed the performance of full-parameter fine-tuning. Overall this paper produces:

- **SemiAdapt-Full** and **SemiAdapt-LoRA** as two efficient, semi-supervised, and embedding-informed approaches that can outperform standard fine-tuning for English to Irish translation.
- A comparative assessment of domain-based fine-tuning using **SemiAdapt-LoRA**, **LoRA**, **SemiAdapt-Full** and **full-model fine-tuning**.
- A comparative assessment of translation performance with **corpus-level** versus **sentence-level** domain labelling.
- A complete **open-source training framework** for English to Irish translation, including implementations of SemiAdapt-Full, SemiAdapt-LoRA, and fully fine-tuned NLLB-200 baselines, with code and configuration files released to support reproducibility[1].

This paper aims to arm LRL modelling researchers with an evaluation of two novel and efficient fine-tuning approaches for English to Irish translation that could be extended to other tasks that benefit from domain adaptation. In contrast to assuming that dataset-level domain labels are sufficiently granular, our sentence-level domain label assignment results in better translation performance across domains. Therefore, SemiAdapt-Full and SemiAdapt-LoRA can empower researchers working on LRLs to build language models such as NMT systems that are both efficient and robust across domains, thus reducing reliance on large labelled datasets and full-parameter fine-tuning.

[1] Code available on GitHub: https://github.com/JoshMcGiff/SemiAdapt

2. Related Work

2.1. Multilingual Models

LLMs and their capacity to generalise across multiple languages at once have been exploited to enhance language generation performance across tasks such as LRL translation (McGiff and Nikolov, 2025; Cahyawijaya et al., 2021; Wongso et al., 2023; Tanwar and Majumder, 2020; Liu, 2022). Furthermore, building language models with groups of linguistically-related languages has been equated with augmenting the training dataset by 33% (Cahyawijaya et al., 2021). Although some research suggests that smaller multilingual models can be outperformed by monolingual models for LRLs such as Slovene, multilingual models tend to perform better at scale (Ulčar and Robnik-Šikonja, 2023). However, further research is required to both verify this generalisation for different language groups and to identify the inflection point where multilingual translation models outperform translation models for a given pair of languages.

2.2. Neural Machine Translation

In terms of Irish language machine translation research, Irish is mostly paired with English for training models (Dowling et al., 2018; Lankford et al., 2022b; Defauw et al., 2019). However, given that training language models of linguistically-related languages can boost modelling performance (Cahyawijaya et al., 2021), Irish could be better aligned with other Goidelic languages such as Scottish Gaelic and Manx (Anderson et al., 2024). To our knowledge, there are no published research efforts exploring large Goidelic language family-based models. That said, code-mixing of English and Irish is very common in Ireland, with both languages often being used in the same sentence (Laoire, 2016). The pervasive influence of English on Irish, a legacy of Ireland's colonial and post-colonial linguistic history, has led to frequent code-mixing and borrowing in modern Irish language use (Hickey, 2009). Consequently, the strong relationship between the languages necessitates the

integration of English in Irish translation systems, and enables English to Irish NLP in general.

Despite the fact that Irish is regarded as a LRL (McGiff and Nikolov, 2025; Meta AI, 2024), a growing body of work has explored machine translation for the language, spanning statistical, recurrent, and transformer-based approaches. During the mid-2010s, some work focused on enhancing statistical machine translation (SMT) methods for English to Irish translation (Arcan et al., 2016). The first NMT systems for the language pair were produced in 2018, with their LSTM-based results underperforming compared to SMT models at the time (Dowling et al., 2018).

More recent work (Dowling et al., 2020; Defauw et al., 2019; Lankford et al., 2022a, 2021) has been dominated by the advent of the Transformer architecture (Vaswani et al., 2017). Two approaches have since used the OpenNMT ecosystem with the Transformer architecture as a base (Dowling et al., 2020; Defauw et al., 2019). Defauw *et al.* explore the positive impact of back-translation of domain-specific monolingual data on translation performance. Additionally, they also indicate that misalignment detection-based filtering of synthetic sentence pairs can produce higher BLEU scores (Defauw et al., 2019). Although Defauw *et al.* explore domain-specific translation, their approach only focuses on the legal domain and infers domain labels at a corpus level. Other papers focus on specific domains such as health (Lankford et al., 2022a) and legislation without exploring parameter-efficient fine-tuning (PEFT) such as LoRA (Lankford et al., 2021), or a wider set of domains at once.

In terms of English to Irish translation performance, it is challenging to compare existing studies as many of them report results such as BLEU on different datasets and domains. Dowling *et al.* report BLEU scores between 31.9-33.9 across various NMT setups on 1,500 random in-domain sentences sourced from datasets mostly found on the OPUS platform (Dowling et al., 2020). Langford *et al.* achieve BLEU scores between 53.4 and 60.5 when exploring the effect of byte-pair encoding vocabulary size on Transformer-powered translation (Lankford et al., 2022b). Tung *et al.* highlight that English to Irish translation is poor on Llama 2-13B as a baseline with a BLEU score of 3.25 (Tran et al., 2024). However, their approach (UCCIX) of fine-tuning LLama 2-13B for the pair achieves a BLEU score of 33.34 on the 500 LoResMT English to Irish evaluation set.

Alternatively, Dowling *et al.* perform human evaluation on OpenNMT-based English to Irish translations (Dowling et al., 2020). Surprisingly, their study was the first to include professional translators in evaluating English to Irish machine translated text. They found that a small group of four translators indicated that NMT was the most accurate in comparison with SMT methods, even when this perception did not fully align with their post-editing experience or fluency preferences. Their findings support the idea in generative language modelling that automatic scores do not necessarily reflect human experience or preference (Mathur et al., 2020; Escribe, 2019).

2.3. Domain Adaptation

Domain adaptation in machine translation has been explored through several established paradigms. Multi-domain NMT approaches train a single model across domains using explicit domain tags appended to source sentences, enabling domain control without architectural modifications (Kobus et al., 2017; Stergiadis et al., 2021). Mixed fine-tuning methods further improve robustness by oversampling in-domain data during training to mitigate catastrophic forgetting while retaining general-domain performance (Chu et al., 2017a). Model merging techniques combine independently fine-tuned checkpoints through parameter averaging to produce unified models (Xu et al., 2025; Gao et al., 2022). However, limited work has explored the integration of model merging techniques with LoRA adapters for efficient domain adaptation in neural machine translation.

2.4. Low-Resource Challenges

While the term “low-resource” typically refers to a lack of data for training language models, it also extends to limitations in computational capacity and access to skilled researchers (Ogueji et al., 2021). A recent systematic review on language modelling for LRLs found that pre-training from scratch was rare, with the majority of studies opting to fine-tune models or develop prompt-engineering strategies for existing models such as ChatGPT (McGiff and Nikolov, 2025). This indicates that limited access to computational resources could pose a barrier to entry for researchers wishing to focus on minority languages for their communities.

Given these computational barriers, LoRA offers a parameter-efficient approach that can significantly reduce the computational requirements for fine-tuning language models (Hu et al., 2022). LoRA works by freezing pre-trained model weights and injecting trainable low-rank decomposition matrices into selected target modules such as the self-attention layers (Zhang et al., 2023). The original LoRA approach reduces the GPU memory required to fine-tune GPT-3 175B to roughly one-third of the original demand, while decreasing the number of trainable parameters by a factor of 10,000 (Hu et al., 2022). LoRA does not necessarily compromise on performance either with the model quality being

similar or better when fine-tuning models such as RoBERTa and GPT-3. As a result, LoRA offers the often computationally-constrained LRL researchers an approach to fine-tuning larger parameter models or many models in parallel on the same hardware.

Furthermore, the modular design of LoRA adapters allows them to be swapped in and out without altering the underlying base model (Hu et al., 2022). The modular design pairs well with domain adaptation as some form of domain detection can be used to route inputs to an appropriate LoRA adapter, thus enabling efficient adapter swapping without the need to load multiple large-parameter models during inference. Although, some studies introduce different complexities and inefficiencies by training models to predict the correct adapter or domain for a given input (Tian et al., 2025; Feng et al., 2024). In terms of NMT, numerous studies have empirically shown that domain adaptation improved translation performance (Freitag and Al-Onaizan, 2016; Chu et al., 2017a), particularly for LRLs (Marashian et al., 2025). As a result, LoRA adapters could be used to bridge the gap in computational constraints for researchers looking to build NMT models for LRLs.

3. Experimental Setup

3.1. Data

We source parallel sentences for English to Irish translation from a variety of sources, using the OPUS platform (Tiedemann, 2012). Furthermore, we mine 813k additional sentence pairs from alternative sources such as the State Examinations Commission and Foclóir (State Examinations Commission, 2025; Foras na Gaeilge, 2013). We automatically extracted parallel English to Irish text from Irish school exam papers by aligning corresponding English and Irish language PDF versions at the page and text-block level using PyMuPDF[2]. Text blocks were normalised by position and matched based on spatial proximity. We excluded language subjects as bilingual copies do not exist or contain irrelevant languages as opposed to English or Irish text content. This new dataset almost triples the existing OPUS-sourced sentence count and brings the total dataset to 1.32 million sentences. We report additional dataset-based features in Table 1

Several studies reveal that web-crawled corpora contain useless text for LRLs, with only 59% of Irish text in mC4 being deemed correct and natural language (Kreutzer et al., 2022; Caswell et al., 2020). Consequently, we exclude web-crawled English to Irish text data from our experiments in order to maintain the linguistic integrity and naturalness of the material.

It can be noted that there are on average 11.92 tokens per English sentence and 12.89 tokens per Irish sentence in the dataset. Interestingly, Irish sentences are about 8% longer than corresponding English sentences.

The general domain dominates as the largest domain with 813k sentence pairs sourced from an official Irish-English bilingual dictionary and Irish exam papers (Foras na Gaeilge, 2013; State Examinations Commission, 2025). We find that the vast majority of corpora are either related to Irish or European-level legislation, and therefore are considered in the legal domain. DGT, EUConst, Gaois, DECP and EuBookshop are the five sources of legal text content, thus making the legal domain the second biggest domain with 478k sentence pairs. Encyclopedic and medical domains follow suit with 19k and 8.6k sentences respectively.

Dataset	Sent.	EN	GA	Domain
DGT (Tiedemann, 2012)	173k	4.1M	4.4M	Legal
EUConst (Tiedemann, 2012)	6.7k	0.14M	0.14M	Legal
Gaois (Gaois research group, 2021)	99k	1.5M	1.6M	Legal
Wikimedia (Tiedemann, 2012)	19k	0.49M	0.52M	Encycl.
LoResMT (Ojha et al., 2021)	8.6k	0.13M	0.15M	Medical/COVID-19
DECP (Koehn, 2005)	103k	1.0M	1.1M	Legal
EUBookshop (Skadiņš et al., 2014)	96k	2.2M	2.3M	Legal
SEC (State Examinations Commission, 2025)	212k	2.5M	2.7M	General
Foclóir (Foras na Gaeilge, 2013)	601k	3.7M	4.1M	General
Total	**1.32M**	**15.7M**	**17.0M**	—

Table 1: English to Irish parallel corpus statistics. Sentence counts shown in thousands (k) with English (**EN**) and Irish (**GA**) token counts displayed in millions (M).

3.2. Base Model

All our experiments are conducted on nllb-200-distilled-600M (Meta AI, 2024) as a base model. This model is the smallest of the public NLLB-200 checkpoints released by Meta. We chose this base model to explore how a relatively small multilingual LLM can perform under a do-more-with-less approach. This configuration enables experimentation on modest hardware, thus offering practical insights for LRL researchers and communities operating under resource constraints. The models were fine-tuned using the Hugging Face transformers framework (Wolf et al., 2020) across two NVIDIA A100 80GB GPUs. We report BLEU (Papineni et al., 2002) as it remains the standard automatic evaluation metric in English to Irish MT studies

[2]PyMuPDF: https://github.com/pymupdf/PyMuPDF

(Tran et al., 2024; Lankford et al., 2022b; Dowling et al., 2020). Although off-the-shelf learned metrics such as COMET (Rei et al., 2020) have shown strong correlation with human judgements for high-resource pairs, several studies report reduced reliability for low-resource pairs and highlight distributional and language-coverage issues (Wang et al., 2024; Zouhar et al., 2024; Falcão et al., 2024). Future work will incorporate human evaluation to complement automatic metrics and provide a more comprehensive assessment of translation quality.

4. Experiments

4.1. SemiAdapt-LoRA: Semi-Supervised LoRA Fine-Tuning

This experiment involves collating the training text content from the sources mentioned in Table 1 and using a zero-shot natural language inference classifier to assign domain labels on a sentence level. 90% of the available English to Irish parallel data are used for fine-tuning and the remaining 10% are reserved for evaluation. Meta's bart-large-mnli model (Lewis et al., 2020) enables noise reduction when creating domain-based dataset splits as not every sentence pair is particularly suited to the general domain of its source. We hypothesise that this preprocessing technique will form better domain groupings and consequently improve domain adaptation performance. This hypothesis is explored further in Section 4.3.

We define the domains for this experiment as general, legal, medical/COVID-19 and wiki/news. We experiment with two different approaches for domain classification with the zero-shot classifier. We create the first training split by providing the zero-shot classifier with all four domains. Alternatively, we create a second split using the same classifier but we exclude the general domain and apply a confidence threshold of 0.45 to the other domains; if no domain exceeds this threshold, the instance is assigned to the general domain.

The English sentences are assigned a domain using the zero-shot classifier and the parallel dataset is subsequently split by domain. The data undergo a process of deduplication, randomisation, and line splitting to account for rows containing multiple sentences. Finally, the training dataset splits are tokenised using the nllb-200-distilled-600M tokeniser (Meta AI, 2024).

We initially select the general domain as the starting point for training as NLLB-200 performs poorly on English to Irish translation by default. That is, we fine-tune entirely on the general domain split for each configuration. General domain fine-tuning was performed using the default `AdamW` optimiser provided in the `Transformers` library (Wolf et al., 2020) with a learning rate of 5×10^{-5}. Training ran for three epochs with a per-device batch size of 4 and gradient accumulation over 4 steps.

We subsequently fine-tune LoRA adapters for each domain, with all the adapters for a single configuration fitting and simultaneously training on a single A100 GPU. For the LoRA-based fine-tuning experiments, we applied low-rank adapters where the configuration used a rank of $r = 16$, scaling factor $\alpha = 16$, and a dropout rate of 0.1. Adapters were inserted into the attention projection layers (`q_proj`, `k_proj`, `v_proj`, `out_proj`) as well as the feed-forward layers (`fc1`, `fc2`). Bias terms were kept frozen, and the task type was set to `SEQ_2_SEQ_LM` to support encoder-decoder training. Fine-tuning was performed using the `AdamW` optimiser with a learning rate of 1×10^{-4}, weight decay of 0.01, and mixed-precision (FP16) training for efficiency. Each model was trained for three epochs with a per-device batch size of 4 and gradient accumulation over 4 steps.

The evaluation step is where the LoRA and the SemiAdapt-LoRA approaches diverge. SemiAdapt-LoRA involves computing centroid embeddings for each domain in the training split and comparing them with the embeddings of input sentences in the evaluation dataset. Unlike the LoRA approach of reusing the zero-shot classifier to label input sentences, SemiAdapt-LoRA is semi-supervised and only uses a classifer to label training data. SemiAdapt-LoRA does not enlist a classification model at inference time. Specifically, the input embeddings are compared to each domain centroid using cosine similarity and assigned a label depending on the closest domain. Domains are essentially assigned based on the semantic similarity of a given input sentence and the average representation for each domain label. We report the results of these LoRA-based experiments in Table 2 and Figure 3.

SemiAdapt-LoRA with domains assigned by the zero-shot classifier outperforms the other approaches on each of the specific domains. Most notably, SemiAdapt-LoRA achieves substantial improvements over the LoRA alternative, with gains of 11 BLEU in the medical domain, 1.7 BLEU in the wiki/news domain, and 3.5 BLEU in the legal domain. Results vary by method for defining domains. The four domains classified without a confidence threshold results in the best results for domain-specific translation.

On the other hand, the three-domain approach where the general domain is labelled based on a confidence threshold, outperforms the domain-specific models on the general domain. This was somewhat surprising given that the confidence threshold-informed general domain split was half the size of the general domain chosen by the

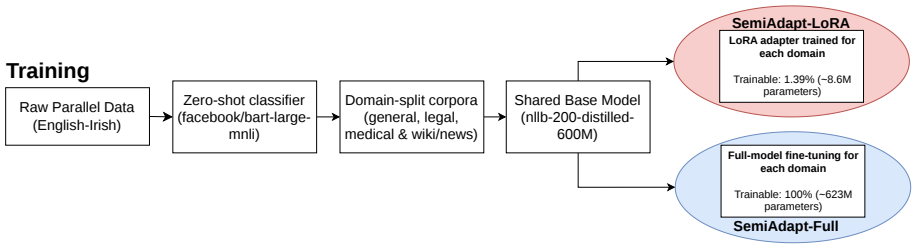

Figure 1: Overview of the training pipeline for SemiAdapt-LoRA and SemiAdapt-Full.

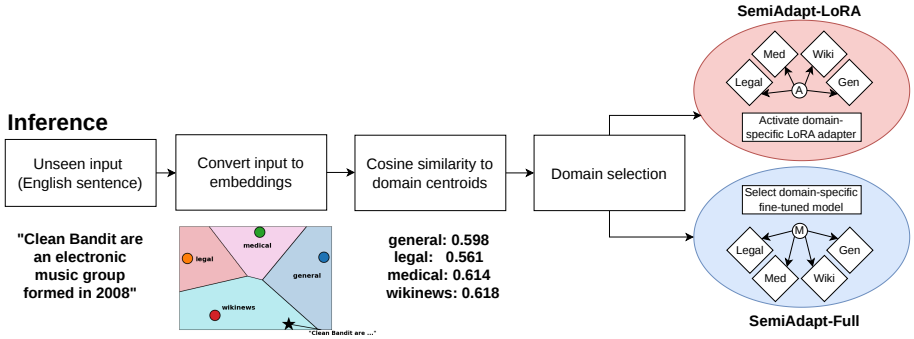

Figure 2: Overview of the inference-time routing and domain selection mechanism.

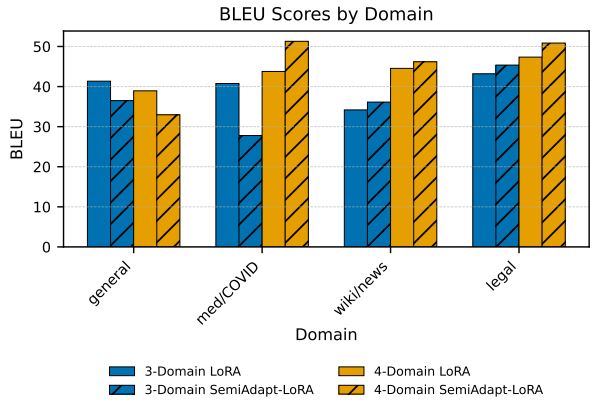

Figure 3: Comparison of BLEU scores by domain for LoRA and SemiAdapt-LoRA models trained on three and four domains.

zero-shot classifier. This likely indicates that the SemiAdapt-LoRA method of assigning domains by comparing input embeddings to domain embedding centroids, struggles with noisy general data. Combining a base model fine-tuned on the entire dataset with SemiAdapt-LoRA for domain adaptation requires further exploration. Although, it is clear that the SemiAdapt-LoRA semantic similarity approach can outperform the supervised approach of using a classifier at inference time, if suitable domain groupings are selected.

4.2. SemiAdapt-Full: Semi-Supervised Full-Model Fine-Tuning

This experiment follows a similar setup to the previous SemiAdapt-LoRA experiment, where full-model fine-tuning, as opposed to adapter fine-tuning, is completed on each domain. Similarly to SemiAdapt-LoRA, SemiAdapt-Full follows a semi-supervised approach by using domain labels sourced from Meta's zero-shot classification model (Lewis et al., 2020) for training and by subsequently using domain embedding centroids to efficiently determine domain labels at inference time. This experiment also explores the impact of the number of domains on SemiAdapt-Full performance. However, unlike SemiAdapt-LoRA, the alternative SemiAdapt-Full approach requires training all of the model's learnable parameters and therefore cannot support multiple models being trained on a single A100 GPU at once.

Similarly to SemiAdapt-LoRA, we choose the general domain as a base for fine-tuning. In other words, we fine-tuned on the general domain before fine-tuning on specific domains such as legal and medical domains. SemiAdapt-Full fine-tuning was performed using the default `AdamW` optimiser provided in the `Transformers` library with a learning rate of 5×10^{-5}. Training ran for three epochs with a per-device batch size of 4 and gradient accumulation over 4 steps. Unlike the SemiAdapt-LoRA configuration, which used PEFT to adapt only low-rank adapter layers, the models in this experiment were fully fine-tuned. This means all model param-

eters were updated during training.

The evaluation step is where full-model fine-tuning and SemiAdapt-Full diverge. Full-model fine-tuning is evaluated on an evaluation dataset that has been labelled with a zero-shot classification model, whereas SemiAdapt-Full uses domain embedding centroids to apply labels at inference time. We include this experiment to explore if regular full-model fine-tuning can also benefit from semi-supervised domain assignment, as seen with LoRA-based models. The inclusion of this full-model fine-tuning setup helps to measure the upper bound of domain-specific translation performance and to provide a comparison point against PEFT methods such as LoRA. We report the results of these full-model fine-tuning experiments in Table 2 and Figures 4 and 5.

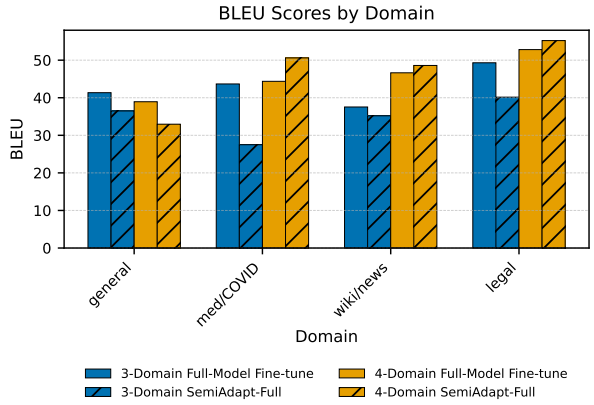

Figure 4: BLEU scores by domain comparing full-model fine-tuning and SemiAdapt-Full models trained on three and four domains.

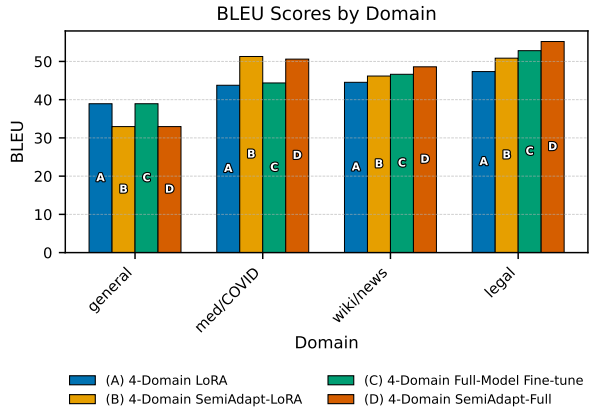

Figure 5: BLEU scores by domain comparing full-model fine-tuning and LoRA-based models trained on four domains.

The results show that four-domain SemiAdapt-Full outperforms the other approaches on each specific domain. Most notably, SemiAdapt-Full achieves considerable improvements over full-model fine-tuning with zero-shot domain classification at inference time on the same medical domain split with a 6.3 increase in BLEU score. More moderate improvements are achieved on wiki/news and legal domains with increases of 2 and 2.4 BLEU respectively. Although, it performs the worst on the general domain, thus indicating that this semi-supervised domain adaptation approach, in addition to SemiAdapt-LoRA, could benefit from a base model fine-tuned on the entire dataset.

Overall, these results reveal that SemiAdapt-Full-based semi-supervised domain adaptation can also improve translation performance when performing full-model fine-tuning with NLLB-200. However, the substantially higher computational cost of full-model fine-tuning, compared to LoRA-based approaches such as SemiAdapt-LoRA, may limit its practicality for researchers working with constrained resources or minority language settings. These results justify experimenting with SemiAdapt-Full on language models for other downstream tasks such as question answering.

Additionally, inter-experiment analysis reveals that SemiAdapt-Full consistently surpasses SemiAdapt-LoRA and LoRA in the wiki/news and legal domains. However, SemiAdapt-Full falls slightly short (0.7 BLEU) of SemiAdapt-LoRA on the medical domain, despite the SemiAdapt-LoRA adapter being 1.39% (8.65M parameters) of the NLLB-200 model's full size (623.72M parameters). Therefore, it is worth noting that the semi-supervised approach helps bridge the gap between LoRA and full-model fine-tuning, with SemiAdapt-LoRA almost on par with fine-tuning entirely on the wiki/news domain, and less than two BLEU short of the legal domain. Not only does this indicate that SemiAdapt-LoRA can help PEFT compete with full-model fine-tuning, but it also highlights that a semi-supervised embedding-based approach can cut out the need for a traditional classifier at inference time.

As a result, SemiAdapt-LoRA presents a particularly suitable approach for resource-constrained domains such as LRL modelling. Alternatively, SemiAdapt-Full presents an inference-efficient approach that can improve domain adaptation and general translation performance for regular fine-tuning. Additionally, this approach could be applied to identify domain groupings within out-of-domain datasets, thereby extending the benefits of domain adaptation reported in the SemiAdapt-LoRA and SemiAdapt-Full experiments.

4.3. Corpus-Level Domain Fine-Tuning

In this experiment, we explore our hypothesis from Experiment 4.1 where we suggest that sentence-level domain labelling could form better domain

groupings and potentially improve domain adaptation performance. The previous SemiAdapt-LoRA and SemiAdapt-Full experiments involve splitting the dataset by domain using a zero-shot classifier on each sentence.

In this experiment, we infer and assign domain labels on a corpus level. In other words, we assume every sentence pair in a corpus reflects the domain label identified in Table 1.

For example, the LoResMT English to Irish dataset related to COVID-19 (Ojha et al., 2021) contains sentences such as:

> EN: *What is added by this report?*
>
> GA: *Cad a chuireann an tuarascáil seo leis?*

Although this example sentence is sourced from a medical/COVID-19 dataset, it does not specifically or exclusively apply to the medical domain. The sentence is more suited to a general, domain-agnostic label. Therefore, these inaccuracies support our hypothesis that sentence-level labeling could reduce noise in the dataset splits and consequently improve domain adaptation and translation performance.

The exam and dictionary corpora are assigned the general domain due to the out-of-domain nature on their text content. We merge all legal sources into a single legal corpus and assume that the Wikimedia and LoResMT datasets correspond to the wiki/news and medical domains, respectively. This time we use 90% of the available English to Irish parallel data from each domain split for fine-tuning and reserve the remaining 10% for evaluation.

Unlike the SemiAdapt-LoRA and SemiAdapt-Full approaches, the domain is inferred by the dataset source and therefore does not require any classification process. We experiment with full-model fine-tuning and with using LoRA to do PEFT on each domain. We subsequently reuse the training parameters for the full-model fine-tuning and LoRA-based approaches from the previous experiments. The results of these corpus-level domain fine-tuning experiments are reported in Table 2 and Figure 6.

The results of this experiment indicate that model performance varied widely across domains when domain labels were assigned at a dataset level. We report that there is a negative correlation between dataset size and BLEU score performance. Both full-model fine-tuning and LoRA-based training methods perform best on the medical and wiki/news domains, i.e. on the domains with the least amount of sentence pairs. Although the full-model fine-tuning approach achieves BLEU scores of 52 and 38 on the medical and wiki/news domains respectively, its poor performance on the significantly larger general and legal domains suggests

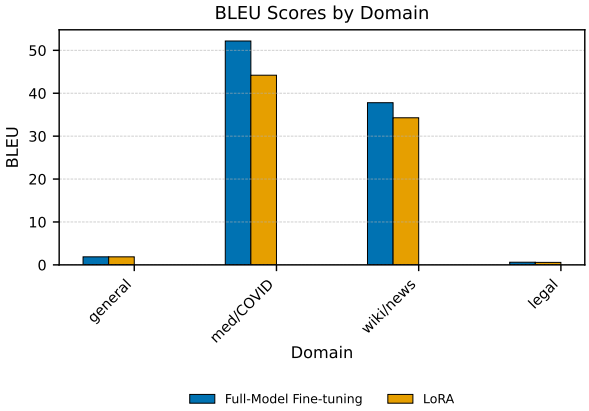

Figure 6: BLEU scores by domain comparing full-model fine-tuning and LoRA-based models on the corpus-level domain fine-tuning split.

the model memorises domain-specific phrases as opposed to learning generalisable translation mappings within the domain itself. This suggests that large, noisy datasets could benefit from a more granular approach to domain labelling and grouping. This suggestion is supported by our results from the SemiAdapt-LoRA and SemiAdapt-Full experiments, where better translation performance was recorded across each of the domains.

4.4. Full-Dataset Fine-Tuning (Baseline)

This experiment involves combining each domain split from the four-domain SemiAdapt-LoRA and SemiAdapt-Full experiment training dataset. Unlike our previous experiments, the NLLB-200 base model is fine-tuned directly on this combined, domain-agnostic collection of English to Irish parallel sentences. We also reuse the evaluation set from the previous four-domain experiments, as this enables a direct comparison between this full-dataset fine-tuning approach and previous experiment methods. We use the same model parameter configuration as the aforementioned SemiAdapt-Full experiment. We report the results of this full-dataset fine-tuning experiment and a comparison of other methods in Table 2 and Figure 7.

The results show that full-dataset fine-tuning yields the strongest performance on the general and wiki/news domains. In contrast, both SemiAdapt-LoRA and SemiAdapt-Full outperform it on the medical domain, with SemiAdapt-Full also leading on the legal domain. Nevertheless, LoRA and full-domain fine-tuning remain within five BLEU of the full-dataset fine-tuning baseline on the general domain, and SemiAdapt-Full is similarly close on wiki/news.

These results indicate that full-dataset fine-tuning can be outperformed by domain-specific fine-

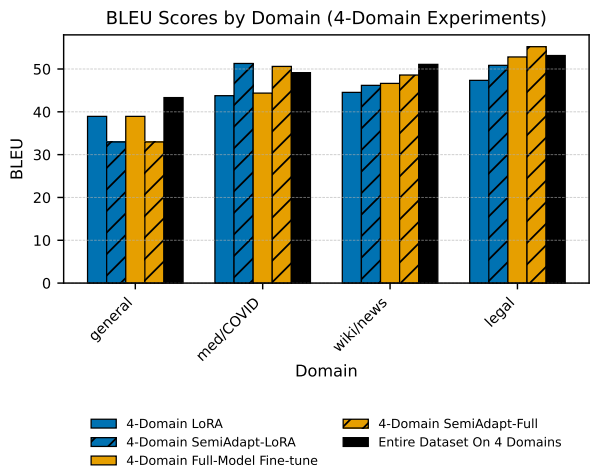

Figure 7: BLEU scores by domain comparing full-dataset fine-tuning, full-model domain fine-tuning, and LoRA-based models on the four-domain split.

Model	General	Medical	WikiNews	Legal
3-domain models				
LoRA	41.35	40.76	34.18	43.20
Full-model domain FT	41.35	43.67	37.54	49.30
SemiAdapt-LoRA	36.49	27.78	36.13	45.36
SemiAdapt-Full	36.49	27.50	35.20	40.16
4-domain models				
LoRA	38.94	43.77	44.54	47.35
Full-model domain FT	38.94	44.37	46.64	52.81
SemiAdapt-LoRA	32.96	51.29†	46.18†	50.85†
SemiAdapt-Full	32.96	50.61	48.59	**55.21**
Corpus-level domain fine-tuning				
LoRA	1.87	44.20	34.28	0.58
Full-model fine-tune	1.87	**52.18**	37.79	0.62
Full-dataset fine-tuning				
Full-model fine-tune	**43.33**	49.15	**51.08**	53.15

Table 2: BLEU scores by domain for different fine-tuning configurations. Bold indicates the best score per domain. † marks parameter-efficient models performing within 5 BLEU of the best model for that domain.

tuning. Moreover, they show that SemiAdapt-LoRA can either surpass or closely match the performance of full-model fine-tuning on both the entire dataset and domain-specific splits. This finding suggests that SemiAdapt-LoRA offers a practical advantage for LRL research, as it enhances PEFT methods to achieve performance comparable to, or even exceeding, that of full-model fine-tuning. However, the strong performance of the full-dataset fine-tuning approach on the general domain supports our earlier argument that SemiAdapt-LoRA and SemiAdapt-Full could benefit from using a full-dataset fine-tuned model as a foundation for further adaptation.

5. Conclusion

In this paper, we introduced two novel domain adaptation techniques, SemiAdapt-LoRA and SemiAdapt-Full, in addition to a suite of open-source models for English to Irish translation. We demonstrated that semi-supervised sentence-level domain adaptation enables LoRA-based models to match the performance of full-model fine-tuning. This in itself should empower resource-constrained LRL researchers to train language models that compete with full-sized models and to train multiple adapters at once in parallel. We have also shown that SemiAdapt-Full outperforms standard domain-based fine-tuning, suggesting that leveraging domain embedding centroids for sentence-level domain labeling can enhance full-model fine-tuning as well as parameter-efficient approaches such as SemiAdapt-LoRA.

For future work, we plan to extend our investigation of SemiAdapt-LoRA and SemiAdapt-Full to other LRLs and additional downstream tasks, such as dialogue generation. In addition, we wish to experiment further with different heuristics for defining domains. Within the scope of neural machine translation, we also intend to explore sentence filtering techniques to better leverage web-crawled Irish text.

Furthermore, we will continue experimenting with domain adaptation strategies, including combining SemiAdapt-LoRA and SemiAdapt-Full with a fully fine-tuned base model. We hope that the introduction of SemiAdapt-LoRA and SemiAdapt-Full will empower researchers working on LRLs such as Irish to develop more accessible and efficient language technologies. These tools can help address technological language inequality, which is being intensified by computationally expensive LLM-based systems that predominantly benefit majority languages with disproportionate digital representation.

6. Limitations

This work focused on higher-quality, non-web-crawled data sources; future work will evaluate the methods on additional datasets. Although a substantial portion of web-crawled text for Irish is not natural or linguistically correct, excluding it limits the available training data. We therefore plan to investigate quality assessment and filtering methods to better leverage web-crawled sentence pairs. Additionally, while we compare against full-model fine-tuning and standard LoRA, broader comparisons with alternative domain adaptation strategies such as multi-domain tagging or mixed fine-tuning remain an important direction for future work. These experiments do not report human evaluation of SemiAdapt-based translations. We leave human

evaluation to future work. Finally, we did not explore the Irish to English translation direction, as decoding into English is generally stronger in large multilingual models, but future work will address this.

7. Ethics Statement

This work contributes to Irish language technology by proposing parameter-efficient domain adaptation strategies tailored to low-resource settings. Due to redistribution restrictions on certain training materials, trained model weights are not publicly released; however, implementation code and training configurations are made available to support transparency and reproducibility. By focusing on computationally efficient adaptation methods, we aim to lower barriers to participation in machine translation research for under-resourced languages and mitigate technological inequalities in current large-scale language models.

8. Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann - Research Ireland under Grant number 18/CRT/6223.

9. Bibliographical References

Cormac Anderson, Sacha Beniamine, and Theodorus Fransen. 2024. Goidelex: A lexical resource for old irish. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 1–10.

Mihael Arcan, Caoilfhionn Lane, Eoin Ó Droighneáin, and Paul Buitelaar. 2016. IRIS: English-Irish machine translation system. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 566–572, Portorož, Slovenia. European Language Resources Association (ELRA).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. In *Translation quality assessment: From principles to practice*, pages 9–38. Springer.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, (108).

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017a. 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, Raj Dabre, and Sadao Kurohashi. 2017b. 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.

Common Crawl. 2025. Common Crawl—Open Repository of Web Crawl Data. Accessed: 2025-07-29.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Arne Defauw, Sara Szoc, Tom Vanallemeersch, Anna Bardadym, Joris Brabers, Frederic Everaert, Kim Scholte, Koen Van Winckel, and Joachim Van den Bogaert. 2019. Developing a neural machine translation system for irish. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 32–38.

Meghan Dowling, Sheila Castilho, Joss Moorkens, Teresa Lynn, and Andy Way. 2020. A human evaluation of English-Irish statistical and neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 431–440, Lisboa, Portugal. European Association for Machine Translation.

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20, Boston, MA. Association for Machine Translation in the Americas.

Marie Escribe. 2019. Human evaluation of neural machine translation: The case of deep learning. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 36–46.

Cristina España-Bonet and Marta R Costa-jussà. 2016. Hybrid machine translation overview. In *Hybrid approaches to machine translation*, pages 1–24. Springer.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. 2024. Mixture-of-loras: An efficient multitask tuning method for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11371–11380, Torino, Italy. European Language Resources Association (ELRA). © 2024 ELRA Language Resource Association, CC BY-NC 4.0.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Yingbo Gao, Christian Herold, Zijian Yang, and Hermann Ney. 2022. Revisiting checkpoint averaging for neural machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 188–196, Online only. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2020. Transfer learning and subword sampling for asymmetric-resource one-to-many neural translation. *Machine Translation*, 34(4):251–286.

Tina Hickey. 2009. Code-switching and borrowing in irish 1. *Journal of Sociolinguistics*, 13(5):670–688.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Séamus Lankford, Haithem Afli, Órla Ní Loinsigh, and Andy Way. 2022a. gaHealth: An English–Irish bilingual corpus of health data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6753–6758, Marseille, France. European Language Resources Association.

Séamus Lankford, Haithem Afli, and Andy Way. 2022b. Human evaluation of english–irish transformer-based nmt. *Information*, 13(7):309.

Seamus Lankford, Haithem Alfi, and Andy Way. 2021. Transformers for low-resource languages: Is féidir linn! In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 48–60, Virtual. Association for Machine Translation in the Americas.

Siobhán Ní Laoire. 2016. Irish-english code-switching: a sociolinguistic perspective. In *Sociolinguistics in Ireland*, pages 81–106. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. From priest to doctor: Domain adaptation for low-resource neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Josh McGiff and Nikola S Nikolov. 2025. Overcoming data scarcity in generative language modelling for low-resource languages: A systematic review. *arXiv preprint arXiv:2505.04531*.

Meta AI. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st workshop on multilingual representation learning*, pages 116–126.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Emmanouil Stergiadis, Satendra Kumar, Fedor Kovalev, and Pavel Levin. 2021. Multi-domain adaptation in neural machine translation through multi-dimensional tagging. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 396–420, Virtual. Association for Machine Translation in the Americas.

Ashwani Tanwar and Prasenjit Majumder. 2020. Translating morphologically rich indian languages under zero-resource conditions. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(6).

Yimin Tian, Bolin Zhang, Zhiying Tu, and Dianhui Chu. 2025. Adapters selector: Cross-domains and multi-tasks lora modules integration usage method. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 593–605.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang Nguyen. 2024. Irish-based large language model with extreme low-resource settings in machine translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 193–202.

Matej Ulčar and Marko Robnik-Šikonja. 2023. Sequence-to-sequence pretraining for a less-resourced slovenian language. *Frontiers in Artificial Intelligence*, 6:932519.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi,

Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Hassan Ayinde, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Ochieng', Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

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.

Wilson Wongso, Ananto Joyoadikusumo, Brandon Scott Buana, and Derwin Suhartono. 2023. Many-to-many multilingual translation model for languages of indonesia. *IEEE Access*, 11:91385–91397.

Sen Xu, Yi Zhou, Wei Wang, Jixin Min, Zhibin Yin, Yingwei Dai, Shixi Liu, Lianyu Pang, Yirong Chen, and Junlin Zhang. 2025. Tiny model, big logic: Diversity-driven optimization elicits large-model reasoning ability in vibethinker-1.5b.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

10. Language Resource References

Foras na Gaeilge. 2013. *Foclóir.ie: English–Irish Dictionary*. Developed by the Lexicography Department, Foras na Gaeilge. Online English–Irish dictionary containing example sentences used research purposes. Publicly accessible, not redistributable.

Gaois research group. 2021. *Gaois: Bilingual Corpus of English–Irish Legislation*. Maintained by the Gaois research group, Fiontar & Scoil na Gaeilge, Dublin City University.

Koehn, Philipp. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*.

Ojha, Atul Kr. and Liu, Chao-Hong and Kann, Katharina and Ortega, John and Shatam, Sheetal and Fransen, Theodorus. 2021. *Findings of the LoResMT 2021 Shared Task on COVID and Sign Language for Low-resource Languages*. Association for Machine Translation in the Americas.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

State Examinations Commission. 2025. *Irish State Examination Papers*. Official bilingual examination materials used for research purposes. Publicly accessible, not redistributable.

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).