BUINUS at IWSLT: Evaluating the Impact of Data Augmentation and QLoRA-based Fine-Tuning for Maltese to English Speech Translation

Filbert Aurelian Tjiaranata[1], Vallerie Alexandra Putra[2], Eryawan Presma Yulianrifat[1], Ikhlasul Akmal Hanif[1]
[1]Universitas Indonesia, [2]Bina Nusantara University
filbert.aurelian@ui.ac.id, vallerie.putra@binus.ac.id

Abstract

This paper investigates approaches for the IWSLT low-resource track, Track 1 (speech-to-text translation) for the Maltese language, focusing on data augmentation and large pre-trained models. Our system combines Whisper for transcription and NLLB for translation, with experiments concentrated mainly on the translation stage. We observe that data augmentation leads to only marginal improvements, primarily for the smaller 600M model, with gains up to 0.0026 COMET points. These gains do not extend to larger models like the 3.3B NLLB, and the overall impact appears somewhat inconsistent. In contrast, fine-tuning larger models using QLoRA outperforms full fine-tuning of smaller models. Moreover, multi-stage fine-tuning consistently improves task-specific performance across all model sizes.

1 Introduction

Despite increasing advances in multilingual technologies, the development of speech translation (ST) systems for low-resource languages continues to pose significant challenges. Maltese, though an official language of the European Union, exemplifies these difficulties. Currently, there are approximately 200 language resources available for Maltese, a relatively small amount, especially compared to the availability of resources for languages spoken in more populous countries (Rosner and Borg, 2022). With fewer than one million speakers and a scarcity of both transcribed speech and parallel text corpora, Maltese remains under-resourced in the context of speech and language processing. This paper describes our approach to the IWSLT 2025 Low-Resource Shared Task for the Maltese-English language pair.

Speech translation involves two main components: transcription and translation. For transcription, we primarily fine-tune Whisper (Radford et al., 2022), while for translation, we fine-tune NLLB (Team et al., 2022). For the larger NLLB model, we also incorporate QLoRA (Dettmers et al., 2023), one of the best parameter-efficient fine-tuning methods, to accommodate resource constraints (Han et al., 2024). However, we treat transcription mainly as a supporting infrastructure and focus the majority of our experimentation on the translation component.

Data augmentation techniques have become indispensable in machine translation, particularly for addressing the challenges posed by limited parallel data in low-resource languages (Hamed et al., 2023). The term "low-resource" refers to the limited availability of data for one of the languages—in this case, Maltese. A common strategy to mitigate this issue is data augmentation (Tang and Lepage, 2023; Takahagi and Shinnou, 2023), which aims to reduce the likelihood of the model encountering completely out-of-distribution data during translation (Wei et al., 2020; Wang et al., 2018).

Unlike most approaches that augment data by generating similar text, the method proposed in (Sánchez-Cartagena et al., 2021) introduces auxiliary tasks such as token swapping, sentence reversal, and the insertion of UNK tokens to enhance model performance. We found this approach promising and adapted it slightly. Specifically, we fine-tuned the NLLB model in two stages: first on both auxiliary tasks and the main translation task, and then on the main task alone to finalize the model.

2 System Overview

Our speech translation pipeline comprises three main components: transcription, machine translation, and data augmentation. Each component is optimized to address the specific challenges of low-resource translation settings.

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 269–273

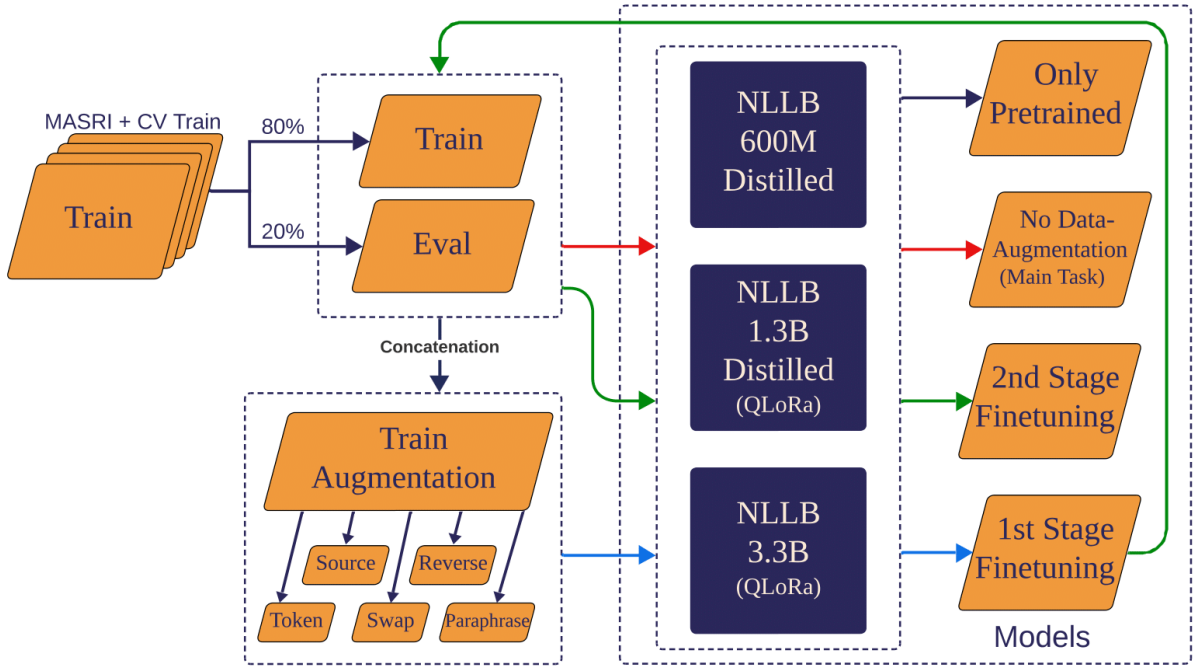

Figure 1: illustrates our end-to-end training pipeline for the translation task. The process begins with the official training data (MASRI + CV Train), which is split into 80% for training and 20% for evaluation. The data is further passed through a data augmentation module.

2.1 Dataset

For this study, we restrict our training data to the official dataset released for the IWSLT 2025 shared task, which consists of approximately 14 hours of speech data. In addition to this, we leverage the pretrained capabilities of Whisper (Radford et al., 2022) and NLLB (Team et al., 2022), both of which were trained on large-scale multilingual corpora. However, we do not incorporate any external datasets beyond what was used during the pretraining of these models.

2.2 Data Splitting and Evaluation Strategy

The dataset is divided into training and validation sets using an 80:20 split. Each speech instance is aligned with its corresponding transcription, and each transcription is paired with a translation from Maltese to English. The Whisper model is fine-tuned using the speech-transcription pairs to perform automatic speech recognition. Separately, the NLLB models are fine-tuned on the Maltese-English text pairs for machine translation. It is important to note that the NLLB models operate exclusively on text and do not utilize any speech data during training.

For evaluation, we use the development set. Evaluation is conducted on both individual components and the complete end-to-end pipeline. Specifically, we assess transcription and machine translation quality independently, as well as the overall performance by feeding Whisper-generated transcriptions into the translation model. This evaluation reflects real-world usage and system robustness.

Performance is measured using the COMET metric (Rei et al., 2020), which provides a semantically-informed evaluation of translation quality. Notably, COMET is also the official evaluation metric used in the shared task competition, ensuring alignment between our development-time evaluation and the final scoring criteria.

2.3 Transcription

We use the Whisper large-v3 model (Radford et al., 2022) for speech transcription. Whisper provides state-of-the-art performance in multilingual speech recognition and serves as a reliable backbone for converting audio input into text.

2.4 Machine Translation

For translation, we employ three variants of the NLLB model (Team et al., 2022): the 600M distilled, 1.3B distilled, and 3.3B versions. The 600M model is fine-tuned directly, while the 1.3B and 3.3B models are fine-tuned using QLoRA (Dettmers et al., 2023) to facilitate efficient adaptation under limited computational resources.

Task	Type	Augmented Training sample
original training sample	source target	roberto ma kienx jidher inkwitat daqs kuġinuh dwar dan Roberto didn't seem as worried as his cousin about this
swap	target	Roberto about seem his worried as as cousin didn't this
token	target	UNK UNK UNK as worried UNK his UNK about this .
source	target	roberto ma kienx jidher inkwitat daqs kuġinuh dwar dan
reverse	target	this about cousin his as worried as seem didn't Roberto
rephrase	target rephrased	but Joe Calleja did not let him continue But Joe Calleja wouldn't let him go on

Table 1: A Maltese–English, word-aligned training sample (first row) and the result of applying the transformations described in Sec. 2.5 using hyperparameter = 0.4 for the swap task and = 0.5 for the token task. Words modified by each transformation are coloured; for swap, a different colour identifies each pair of words that are swapped together; for rephrased, a different colour identifies each pair of words rephrased.

2.5 Data Augmentation

We follow the multi-task data augmentation (MTL DA) framework proposed by Sánchez-Cartagena et al. (2021) (Sánchez-Cartagena et al., 2021), where several auxiliary tasks are defined to modify target sequences in ways that challenge the decoder and reinforce encoder reliance. Among the auxiliary tasks the *swap* and *token* are controlled by a hyperparameter α, which determines the proportion of tokens in the target sentence that are affected. For instance, in the *swap* task, α defines the fraction of target words whose positions are altered; similarly, in the *token* task, it defines the proportion of target words replaced by the `[UNK]` symbol.

Swap	Random swapping of target tokens to disrupt sequence order.
Token	Replacement of target tokens with the `[UNK]` symbol.
Source	Copying the source sentence to the target side.
Reverse	Reversal of the target token order.
Paraphrase	As an additional augmentation method, we employ a paraphrasing approach using NLLB for back-translation, which translates the target sentence to Italian and then back to English.

Although we did not conduct hyperparameter tuning in our setup, we adopted $\alpha = 0.4$ for the *swap* task and $\alpha = 0.5$ for the *token* task, which fall within the optimal range (typically $\alpha \in [0.1, 0.9]$) explored in the original study. These values were chosen based on their reported performance and balancing between task disruption and learnability as described in (Sánchez-Cartagena et al., 2021). The choice allows us to benefit from the task's intended regularization effect without introducing excessive noise.

Fine-tuning is conducted in two stages: an initial phase on a mixture of the main and auxiliary tasks, followed by a final phase focused solely on the primary translation task.

3 Results

Model Size and Fine-Tuning Strategy. Our results indicate that the larger 3.3B NLLB model, fine-tuned using QLoRA, outperforms the smaller 600M model that is fully fine-tuned. While the larger models achieve higher overall performance after fine-tuning, this may partly reflect its stronger baseline performance. The performance gain from fine-tuning is actually greater for the smaller 600M model, suggesting that smaller models benefit more directly from the fine-tuning process, while larger models rely more on their pretrained capacity.

Effect of Data Augmentation. For the 3.3B model, none of the tested data augmentation techniques such as paraphrasing, token swapping, UNK token insertion, or sentence reversal led to noticeable gains, with the highest improvement being

Model	Baseline	1st-Stage (DA)					2nd-Stage (DA)				
	(No DA)	Swap	Token	Source	Reverse	Paraphrase	Swap	Token	Source	Reverse	Paraphrase
NLLB 3.3B pretrained	0.8056	–	–	–	–	–	–	–	–	–	–
NLLB 1.3B distilled	0.8018	–	–	–	–	–	–	–	–	–	–
NLLB 600M distilled	0.7858	–	–	–	–	–	–	–	–	–	–
NLLB 3.3B fine-tuned	0.8323	0.8320	0.8310	0.8275	0.8316	0.8196	0.8322	0.8323	0.8320	0.8321	**0.8324**
NLLB 1.3B fine-tuned	0.8275	–	–	–	–	–	–	–	–	–	–
NLLB 600M fine-tuned	0.8223	0.8235	0.8221	0.8198	0.8229	0.8186	0.8240	**0.8250**	0.8229	0.8249	0.8241
Whisper to NLLB 3.3B fine-tuned	0.7602	**0.7608**	0.7595	0.7512	0.7601	0.7499	0.7604	0.7602	0.7607	0.7601	0.7601
Whisper to NLLB 600M fine-tuned	0.7472	0.7507	0.7477	0.7452	0.7495	0.7481	0.7501	0.7495	0.7493	**0.7529**	0.7503

Table 2: COMET scores for two-stage fine-tuning. The first three rows show pretrained NLLB models without fine-tuning. The next three rows show the NLLB models after fine-tuning. Baseline: no data augmentation; 1st-Stage: scores with data augmentation (DA); 2nd-Stage: another fine-tuning without DA.

just 0.001 COMET points from paraphrasing. The 600M model, on the other hand, showed slightly better results, with consistent but small improvements across all methods, reaching up to 0.0026 COMET points. While the gains for the smaller model are more apparent, they remain modest. These results suggest that data augmentation may be more useful for smaller models, which benefit more from the added variability due to their limited capacity. This aligns with prior findings (Sánchez-Cartagena et al., 2021), where augmentation strategies had a greater effect on less-pretrained models.

Impact of Multi-Stage Fine-Tuning. The two-stage fine-tuning approach, where models are first trained on a mix of auxiliary and primary translation tasks and then fine-tuned solely on the main task, resulted in performance improvements across all model sizes. This shows that a final alignment phase focused on the primary objective enhances model performance and task-specific adaptation.

End-to-End Performance. Table 2 shows that feeding Whisper transcriptions into the NLLB models lowers COMET by around 0.07–0.076 points across all settings. This degradation is most likely caused by transcription errors from the ASR stage, which the MT component cannot fully recover from. Notably, the highest end-to-end COMET score achieved was 0.7608, obtained using Whisper to NLLB 3.3B fine-tuned model with swap-based augmentation in the first stage. For the official test set submission in the unconstrained setting, this same system achieved 45.4 BLEU and 65.11 chrF.

4 Limitations

The limitation of our study is the lack of extensive qualitative analysis due to limited language proficiency. Since we do not fully understand the language in the dataset, our analysis primarily relies on quantitative methods.

5 Conclusion

In this paper, we explore the use of pretrained models—Whisper for ASR and NLLB for MT—alongside data augmentation and parameter-efficient fine-tuning methods. Our experiments show that fine-tuning larger NLLB models using QLoRA outperforms full fine-tuning on smaller models. Two-stage fine-tuning also provides consistent performance improvements across model sizes. In contrast, data augmentation offers only marginal benefits, limited to the smaller 600M model, and the improvements appear inconsistent.

These findings highlight the promise of scalable fine-tuning techniques for translation in low-resource settings. However, our focus on MT fine-tuning overlooks the more significant impact of ASR errors, which remain a primary source of performance degradation in the end-to-end pipeline. This suggests that future research should prioritize improvements in the ASR component. Further work could also explore more targeted data augmentation strategies, end-to-end fine-tuning approaches, and incorporate qualitative evaluations with native speakers to better capture translation quality nuances.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Injy Hamed, Nizar Habash, and Thang Vu. 2023. Data augmentation techniques for machine translation of code-switched texts: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 140–154, Singapore. Association for Computational Linguistics.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

Mike Rosner and Claudia Borg. 2022. Report on the maltese language. Language Technology Support of Europe's Languages in 2020/2021. Available online at `https://european-language-equality.eu/deliverables/`.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2021. Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8502–8516, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kyosuke Takahagi and Hiroyuki Shinnou. 2023. Data augmentation by shuffling phrases in recognizing textual entailment. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 194–200, Hong Kong, China. Association for Computational Linguistics.

Wenyi Tang and Yves Lepage. 2023. A dual reinforcement method for data augmentation using middle sentences for machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 48–58, Macau SAR, China. Asia-Pacific Association for Machine Translation.

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. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Luxi Xing, and Weihua Luo. 2020. Uncertainty-aware semantic augmentation for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2724–2735, Online. Association for Computational Linguistics.