UOM-Constrained IWSLT 2024 Shared Task Submission - Maltese Speech Translation

Kurt Abela **Md Abdur Razzaq Riyadh** **Melanie Galea** **Alana Busuttil**

Roman Kovalev **Aiden Williams** **Claudia Borg**

kurt.abela@um.edu.mt, md.riyadh.23@um.edu.mt, melanie.galea.20@um.edu.mt,
alana.busuttil.20@um.edu.mt, roman.a.kovalev.23@um.edu.mt,
aiden.williams.19@um.edu.mt, claudia.borg@um.edu.mt,

Abstract

This paper presents our IWSLT-2024 shared task submission on the low-resource track. This submission forms part of the constrained setup; implying limited data for training. Following the introduction, this paper consists of a literature review defining previous approaches to speech translation, as well as their application to Maltese, followed by the defined methodology, evaluation and results, and the conclusion. A cascaded submission on the Maltese to English language pair is presented; consisting of a pipeline containing: a DeepSpeech 1 Automatic Speech Recognition (ASR) system, a KenLM model to optimise the transcriptions, and finally an LSTM machine translation model. The submission achieves a 0.5 BLEU score on the overall test set, and the ASR system achieves a word error rate of 97.15%. Our code is made publicly available[1].

1 Introduction

Speech Translation (ST) may be defined as the task of transforming audio in a source language to its transcription in a target language. ST is generally tackled through two main approaches: the first being an end-to-end approach; with the source language audio serving as input to the model, which in turn produces a transcription in the target language as output, the second being a pipeline or cascading approach; suggesting multiple systems with varying responsibilities, primarily generating ASR transcription and machine translation. A Meta-Net White Paper series confirms the Maltese language as low-resourced; meaning it has little support for speech technology, including translation tasks (Rosner and Joachimsen, 2012).

This paper introduces a cascading system that utilises an ASR system to generate transcriptions in the source language, a language model to improve the transcriptions and finally, a machine translation system to produce the transcription in the target language. The following sections define the current state of research into low-resource speech translation, followed by a methodology and discussion.

[1] https://github.com/melanie-galea/uom_constrained

2 Literature Review

The literature review focuses on previous attempts at Automated Speech Recognition (ASR) and Machine Translation (MT), in particular, when applied to the Maltese language. Furthermore, the main models attempted for this task are defined, these being: HMM, DeepSpeech 1 for ASR, LSTM and Transformers for MT.

2.1 Previous IWSLT Low-Resource Track Attempts

In 2023, the shared task set by IWSLT consisted of "benchmarking and promoting speech translation technology for a diverse range of dialects and low-resource language".

Among other attempts, QUESPA (E. Ortega et al., 2023a) submitted two cascade systems to the constrained setting, where ASR and MT were combined together in a pipeline. One of these cascade systems used wav2letter+ (Pratap et al., 2019) - a fast open-source speech recognition system; the other one was an implementation of a conformer architecture along with OpenNMT translation system (Klein et al., 2017), which was trained on constrained ST and MT data. Both of these models demonstrated relatively poor performance compared to the other submissions, with a BLEU score of less than 1.

Previous attempts in both constrained and unconstrained settings, proved that this task is still a major challenge. Using powerful massively pretrained ASR models; such as Wav2Vec 2.0, in combination with multilingual decoders has been an emerging trend, and oftentimes produces excellent

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 156–161

results. Training a self-supervised model and producing artificial supervision has proven to be an effective approach (Zanon Boito et al., 2022). Additionally, several methods were employed to improve the performance of cascade systems, such as voice activity detection for segmentation (Zhang et al., 2022; Ding and Tao, 2021), as well as training the ASR on synthetic data with noise filtering and domain-specific fine-tuning (Zhang et al., 2022).

2.2 HMM and DeepSpeech for Maltese ASR

Our work attempts two instruments for ASR: Hidden Markov Model and DeepSpeech 1. The former used to be a preferred method since the 1970s (Rabiner, 1989). As demonstrated by Ellis and Morgan (1999), the size of a model plays a significant role, especially when it comes to the quantity of training data and the trainable parameters. The latter was made difficult due to hardware and design limitations. A survey conducted by Nagpal et al. (2019) showed that deep learning approaches could still deliver effective results for ASR.

This led to the development of end-to-end supervised neural network models such as DeepSpeech 1 (Hannun et al., 2014) and then DeepSpeech 2 (Amodei et al., 2015), which had successfully outperformed Hidden Markov Models for English ASR. In speech-related fields, labelled data is usually referred to as annotated data. Although the use of large amounts of annotated data proved beneficial for these models, access to data at these scales became a limitation for development, especially in the case of low-resource languages. This led to the use of unannotated data in unsupervised training, with cases described in Lee et al. (2009)'s and Radford et al. (2016)'s work, or self-supervised learning, namely the work done on the Wav2Vec system (Schneider et al., 2019).

These acoustic models have the capacity to generate understandable transcriptions at the character level. Yet, these transcriptions often harbour inaccuracies, such as substituting phonetically similar words or misspelling due to language orthography idiosyncrasies. Consequently, enhancing ASR systems by incorporating an external language model trained on domain-specific text can boost their performance. A common strategy employed is the use of a simple n-gram model. The KenLM language modelling tool (Heafield, 2011) is able to achieve high processing efficiency and language modelling quality by assigning a score to a sequence of *n* words. This is particularly useful in ASR to be able to select between multiple possible candidates through beam search. Deepspeech supports the integration of KenLM language models to enhance the quality of the ASR output.

2.3 Machine Translation

While some systems have reached human parity in certain domains in machine translation, this is yet to be achieved for low-resource languages (Hassan et al., 2018). The primary challenge lies in parallel data scarcity. The efforts in solving this issue focus on various other aspects such as exploiting shared language features between a high and low resource language as well as techniques for data augmentation.

Her and Kruschwitz (2024) used German-Bavarian parallel data to train a transformer model and then used that to back-translate and augment the training set. They then used that data to fine-tune a German-French neural translation model given its similarity to the source language. Nzeyimana (2024) focused on improving the performance of machine translation models by improving predictions of the morphological features. Their method was based on the fact that sub-word tokenizers split the words on a surface level and are prone to losing morphological features. Encoding morphological features as input to the model improves performance. E. Ortega et al. (2023b) used the Transformer architecture (Vaswani et al., 2017) to develop a machine translation system as part of their pipeline for an automatic speech recognition system for Quechua to Spanish.

Before Transformers (Vaswani et al., 2017), RNN (Rumelhart et al., 1986) was widely used for natural language processing. RNN with self-attention proved quite effective for machine translation, achieving state-of-the-art performance (Sutskever et al., 2014), (Bahdanau et al., 2014).

Research on machine translation for Maltese is quite limited. One of the earliest works was on statistical machine translation where the authors focused their attention on phrase extraction for proper phrase alignment (Rosner and Bajada, 2007). In their work on Maltese automatic speech recognition (ASR), Williams et al. (2023) leveraged the pre-trained mBART model. However, their system was evaluated as a whole (ASR - MT) and does not represent the model's true capability for machine translation on Maltese as the input is the ASR output.

3 Methodology

3.1 Automatic Speech Recognition

3.1.1 Hidden Markov Models

Hidden Markov Models (HMMs) were trained for ASR (Rabiner, 1989) on the MASRI dataset, totaling to 6 hours and 39 minutes. The model was trained using Mel Frequency Cepstral Coefficient (MFCC) features derived from the WAV files and their corresponding verbatim transcription.

3.1.2 DeepSpeech

A DeepSpeech v1 (Hannun et al., 2014) model was trained on both MASRI and CV datasets, containing 6 hours 39 minutes and 5 hours 11 minutes respectively, totalling nearly 12 hours. The model was trained using Maltese WAV files and their corresponding verbatim transcription. Development, Testing and Training csv files therefore contain the WAV file dataset root, its corresponding transcription, and file size in bytes. The text was pre-processed; characters cases were converted to lower-case, non-alphabetic characters removed except for the hyphen and apostrophe. Accented letters were included in order to better support the model's understanding of pronunciation. An alphabet was created including special Maltese characters.

The training code was cloned through the git DeepSpeech branch, and all required dependencies were installed. Finally, the training, development and testing files, along with a layer size of 64 units wide, rather than the default 2048. The dropout was set to 0.4, and a batch size of 100 was used to train the model. The model was trained for 250 epochs. The hyper-parameters were set with the limited data-set size in mind. The relatively smaller size of the model parameters was beneficial for our experiment; it is usually the case that a larger parameter size causes the model to over-fit when trained on a small training set such as ours.

The DeepSpeech model was selected over the HMM due to higher performance. The HMM scored a WER of 112.33%, whilst DeepSpeech model scored a WER of 97.15%.

3.2 KenLM

Initial experimentation involved investigating the impact of both word-level and character-level n-grams on a set of erroneous test data. Upon examining the alterations made by KenLM on this sample dataset, it was deduced that a word-level KenLM model was more apt for the task.

The KenLM toolkit (Heafield, 2011) was used to train a probabilistic 3-gram model on the Korpus Malti v4.0 Shuffled training subset (Micallef et al., 2022) [2], which is resource referred to by the shared task organizers. Before training said model, the corpus was pre-processed to not include punctuation, apart from the hyphen and apostrophe, with all text lowercased. The KenLM model, once trained, served as a tool to decode the ASR output, employing a beam search algorithm. This process converted probabilities into textual transcripts, which were subsequently delivered by the system.

3.3 Machine Translation

All models are built using the Fairseq (Ott et al., 2019) library. The Fairseq library allows for easy implementation of a MT system through CLI commands, meaning minimal code is needed to create a fully working MT system.

Three different architectures were experimented with, namely Transformer (base), Transformer (large) and an LSTM. The base transformer version (Vaswani et al. (2017)) has six encoder and decoder layers with 512 dimensions each. There are eight attention heads for both the encoders and decoders, with 2048 dimensions for each. The large version of the transformer architecture has 1024 dimensions for each layer and 4096 dimensions for each attention head. There are also 16 attention heads in total. Thirdly, an LSTM architecture (Hochreiter and Schmidhuber, 1997) was used, which consists of a single-layer bidirectional encoder-decoder model with a hidden size of 512 for both the encoder and decoder.

LSTMs have generally fallen out of favour recently due to Transformers achieving better results. However, it was hypothesised that given the lack of data, LSTMs may still prove to be just as effective in this scenario. This is due to the fact that Transformers require a lot of data to be effective, and in low-resource settings such as this one, older techniques such as LSTM may perform better (Przystupa and Abdul-Mageed, 2019).

The data was pre-processed by training a SentencePiece tokenizer from scratch on the given training set. The training set was then pre-processed using this tokenizer.

[2] https://huggingface.co/datasets/MLRS/korpus_malti/viewer/shuffled

The 3 defined models (LSTM, base Transformer and large Transformer) were trained with the same hyperparameters. We performed an evaluation of all three models on the dev set and achieved the results in Table 1. It was ultimately concluded that LSTMs performed best. The LSTM model was therefore selected, and further hyperparameter tuning was performed for improved results.

Table 1: Results of the different architectures on the development set

Architecture	BLEU	CHRF-2
LSTM	**25.76**	**44.57**
Transformer (Base)	24.54	44.15
Transformer (Large)	25.20	43.57

For hyperparameter tuning, Akiba et al. (2019) was used to find optimal values for learning rate, dropout, warm-up duration and weight decay. The optimal learning rate was found to be 0.003 and dropout at 0.04. The learning rate scheduler was set with warm-up updates of 8522. Each model was trained for a maximum of 1,000,000 steps but all of them converged much sooner. The final LSTM model was trained for 4 minutes with early stopping. The training was stopped early when the validation BLEU did not improve for 10 steps.

4 Evaluation and Results

This section presents and discusses the models' results. The official results for our constrained task submission are presented in Tables 2 and 3. The final pipeline result was significantly influenced by the ASR performance. It can be extrapolated that the high Word Error Rate (WER) of the ASR is a result of the limited training data, which was not adequate to train a capable ASR system. Incoherent speech recognition outputs were considered 'out of domain' by the machine translation system since it was trained on meaningful data.

Table 2: Official results for the constrained task - BLEU score

Test Set	BLEU score
CV	0.6
Masri	0.2
Overall	0.5

To further evaluate and understand the results, specific outputs of both the ASR as well as the MT system were analysed.

Table 3: Official results for the constrained task - Word Error Rate

Test Set	Word Error Rate
CV	97.0%
Masri	97.43%
Overall	97.15%

Table 4: Results of the pipeline system with and without the use of a KenLM.

Test Set		BLEU	CHRF-2
Without KenLM	CV	0.48	15.79
	Masri	**0.23**	14.50
With KenLM	CV	**0.52**	**15.97**
	Masri	0.21	**14.73**

It may be noted that the ASR output is relatively poor; with most outputs consisting of invalid Maltese words. In addition to using Deepspeech 1, we made use of a KenLM trained specifically for this task, but whilst some improvements were seen, as illustrated in Table 4, it was not enough to compensate for the model's inability to accurately transcribe the Maltese language.

To further illustrate the ASR issues, the first audio file of the CV test set was transcribed by the model as: *"dan ma sarqat"*. The first two words were predicted correctly, however, the last word was invalid. The correct transcription should have been: *"dan ma sar qatt"*, meaning *"this was never done"*.

Since this is a pipeline setup, the resulting transcription was passed to the MT system. The translated output was *"this doesn't happen to him"*. The output here was not surprising, since the two words that the ASR got correct (*dan ma*) roughly mean *he has never [...]*. Since the word that the ASR got incorrect does not exist in the Maltese language (*sarqat*), it is likely that the MT system treated it as an unknown token.

Admittedly, this was one of the few examples that the ASR system performed well in. Results were exceptionally poor when a named entity was included. For example, the name *Simon Busuttil* was outputted as *sajminbużutiel*. This is expected due to the small size of the training data. Apart from this, the ASR model struggled to understand when a word starts and ends. In most cases, the output sounds phonetically similar to what the actual transcription should be, however, the spelling is incorrect. For example, the word *mhux* was tran-

scribed as *mux*, which is understandable as the 'h' is silent. Overall, the ASR model performed poorly, with most resulting sentences not resembling the actual transcription, highlighted by the 97.15% WER.

These errors naturally propogated to the MT system. Since the dataset of the MT system is also constrained and very limited in nature, it did not have the implicit understanding of the language to identify the typos written by the ASR system (such as *mux* instead of *mhux*). This is even harder with phonetic misspellings. Generally, the MT system output a (seemingly) random response since the input given by the ASR system is equally poor.

Overall, it is evident that an increase in training data would have yielded better results. The ASR set-up makes it difficult to evaluate the MT system alone, given the model pipelining and overall poor performance.

5 Conclusion and Future Work

This paper presents the different approaches to ST for low-resource languages under constrained settings. A short overview of previous research into challenges associated with speech translation was presented, as well as specific attempts and pipelines used for the task. The final pipeline consisted of a DeepSpeech 1 model, KenLM model and LSTM model, each fine-tuned for the task at hand. The final results show that the constrained setting has an extreme impact on the models performance, with a final WER of 97.15%. The very poor ASR performance highlights the challenges present in low-resource settings. Future work on ASR includes the use of higher-quality training data, as well as dealing with named entities in the data itself. It is also suspected that pre-trained models would likely yield better results in low-resourced environments, helping to compensate for data scarcity.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *Preprint*, arXiv:1907.10902.

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. 2015. Deep speech 2: End-to-end speech recognition in english and mandarin. *Preprint*, arXiv:1512.02595.

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

Liang Ding and Dacheng Tao. 2021. The USYD-JD speech translation system for IWSLT2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 182–191, Bangkok, Thailand (online). Association for Computational Linguistics.

John E. Ortega, Rodolfo Zevallos, and William Chen. 2023a. QUESPA submission for the IWSLT 2023 dialect and low-resource speech translation tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 261–268, Toronto, Canada (in-person and online). Association for Computational Linguistics.

John E. Ortega, Rodolfo Zevallos, and William Chen. 2023b. QUESPA Submission for the IWSLT 2023 Dialect and Low-resource Speech Translation Tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 261–268, Toronto, Canada (in-person and online). Association for Computational Linguistics.

D. Ellis and N. Morgan. 1999. Size matters: an empirical study of neural network training for large vocabulary continuous speech recognition. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, volume 2, pages 1013–1016 vol.2.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *Preprint*, arXiv:1412.5567.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Wan-Hua Her and Udo Kruschwitz. 2024. Investigating Neural Machine Translation for Low-Resource Languages: Using Bavarian as a Case Study. *arXiv preprint*. ArXiv:2404.08259 [cs].

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Honglak Lee, Yan Largman, Peter Pham, and Andrew Y. Ng. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, page 1096–1104, Red Hook, NY, USA. Curran Associates Inc.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.

Shaveta Nagpal, Munish Kumar, · Maruthi, Rohit Ayyagari, and · Kumar. 2019. A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering*.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. *arXiv preprint*. ArXiv:2404.02392 [cs].

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2019. Wav2letter++: A fast open-source speech recognition system. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235.

L.R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. *Preprint*, arXiv:1511.06434.

Michael Rosner and Jo-Ann Bajada. 2007. Phrase extraction for machine translation. Accepted: 2017-10-17T13:30:37Z Publisher: University of Malta. Faculty of ICT.

Mike Rosner and Jan Joachimsen. 2012. *Il-Lingwa Maltija Fl-Era Diġitali – The Maltese Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika*, 71:599–607.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

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.

Aiden Williams, Kurt Abela, Rishu Kumar, Martin Bär, Hannah Billinghurst, Kurt Micallef, Ahnaf Mozib Samin, Andrea DeMarco, Lonneke Van Der Plas, and Claudia Borg. 2023. UM-DFKI Maltese Speech Translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 433–441, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, and Yannick Estève. 2022. ON-TRAC consortium systems for the IWSLT 2022 dialect and low-resource speech translation tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 308–318, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Pan Deng, Mohan Shi, Yifan Song, Dan Liu, Junhua Liu, and Lirong Dai. 2022. The USTC-NELSLIP offline speech translation systems for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 198–207, Dublin, Ireland (in-person and online). Association for Computational Linguistics.