UoM-DFKI submission to the low resource shared task

Rishu Kumar[◇] **Aiden Williams*** **Claudia Borg*** **Simon Ostermann**[◇]

[◇]DFKI, *University of Malta
rishu.Kumar@dfki.de, aiden.williams.19@um.edu.mt
claudia.borg@um.edu.mt, simon.ostermann@dfki.de

Abstract

This system description paper presents the details of our primary and contrastive approaches to translating Maltese into English for IWSLT 24. The Maltese language shares a large vocabulary with Arabic and Italian languages, thus making it an ideal candidate to test the cross-lingual capabilities of recent state-of-the-art models. We experiment with two end-to-end approaches for our submissions: the Whisper and wav2vec 2.0 models. Our primary system gets a BLEU score of 35.1 on the combined data, whereas our contrastive approach gets 18.5. We also provide a manual analysis of our contrastive approach to identify some pitfalls that may have caused this difference.

1 Introduction

In this paper, we describe the UoM-DFKI submission to the Dialectical and Low-Resource track of the IWSLT 2023 evaluation campaign, focusing on the unconstrained approach for the Maltese to English track. Maltese is considered a hybrid language, with most vocabulary coming from Arabic, Italian, and English. While there is a major overlap with Arabic, Maltese data uses Latin instead of Arabic script. Our main focus for this submission is using publicly available multilingual models to exploit the multilingual capabilities of models, given the interesting mixture of vocabulary in the Maltese language.

For this paper, we focus on end-to-end approaches for spoken language translation (SLT), namely with Whisper (Radford et al., 2022) and wav2vec 2.0 xls-r (Baevski et al., 2020; Babu et al., 2021). We use the Whisper-based model as our primary submission and the wav2vec 2.0 model as the contrastive approach. The Whisper system is pre-trained on 680,000 hours of speech data using an encoder-decoder method. A substantial amount of the training data, nearly one-fifth, is English audio, and 9,000 hours is Maltese. (Radford et al., 2022) claim that with 41 hours of Maltese translation data, the Whisper model is able to achieve roughly 14 BLEU points. In this paper, we use the data released for this task to fine-tune the Whisper model further. There are various Whisper models with varying parameter sizes. (Williams et al., 2023b) shows how, with larger parameters, the Whisper architecture performs better in the ASR setting for Maltese ASR. For this work, we decided to use the most recent Whisper model; the largest model is Whisper-large-v3. Our approach for wav2vec 2.0-based models also consisted of using an encoder-decoder approach, namely SpeechEncoderDecoder framework (Chan et al., 2015; Wang et al., 2021), as made available on HuggingFace (Wolf et al., 2020). We worked with three different models as our decoder for our contrastive approaches, namely BERT (Devlin et al., 2019) and mBART fine-tuned for machine translation from different languages into English (Tang et al., 2020).

2 Literature Review

The IWSLT Low-resource and Dialectical shared task increased the number of language pairs they released data for in 2023. In the 2022 edition of the workshop, (Anastasopoulos et al., 2022) released the data for teams to develop systems to transcribe and translate the low-resource language pairs of Tamasheq-English and Tunisian Arabic-French. In the 2023 edition of the task, however, Agarwal et al. (2023) extended the task to include the language pairs Irish-English, Maltese-English, Pashto-French and Quechua-Spanish.

In 2022, three teams submitted models for Tamasheq-English: ON-TRAC (Zanon Boito et al., 2022), TalTech and GMU. ON-TRAC also submitted to the Tunisian Arabic-French pair, like CMU (Yan et al., 2022) and JHU (Yang et al., 2022) did. In 2023, GMU submitted models

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 328–333

for Irish-English, Marathi-Hindi, Pashto-French and Tamasheq-French (Mbuya and Anastasopoulos, 2023), Alexa AI submitted models for Marathi-Hindi and Tamasheq-French (Shanbhogue et al., 2023), ON-TRAC submitted for Tamasheq-French and Pashto-French (Laurent et al., 2023), NAVER submitted for Tamasheq-French and Quechua-Spanish (Gow-Smith et al., 2023), BUT (Kesiraju et al., 2023) and SRI-B (Radhakrishnan et al., 2023) submitted only for Marathi-Hindi, QUESPA submited for Quechua-Spanish (E. Ortega et al., 2023) and UM-DFKI submitted for Maltese-English (Williams et al., 2023a).

These teams employed various techniques, ranging from traditional cascade systems to various end-to-end architectures. Many teams leveraged large pre-trained models, including XLS-R, mBART, Wav2Vec 2.0 and HuBERT. The Alexa AI team tried an interesting approach by focusing on data augmentation, ensemble modelling and post-processing techniques to improve their results. Transformer models for MT were popular across the board. The NAVER submissions obtained particularly good results in their respective language pairs by using pre-trained ASR and MT models from the NLLB project (Team et al., 2022), which include both Tamasheq and Quechua in their training, showing the importance of language diversity in multilingual models.

3 Dataset

In this section, we briefly describe the dataset used to fine-tune our systems. We include a description of the dataset used to fine-tune the mBART50 many-to-one model (Tang et al., 2020) and the dataset released for this shared task.

mBART50 many-to-one (Tang et al., 2020) utilized and released the ML50 dataset for fine-tuning the mBART model for translating in 50 languages. It uses English as a pivot language to collect parallel data for 49 other languages from sources such as IWSLT, TED, WAT, etc. It is also noted that the 49 languages selected for the dataset are based on language family, available mono-lingual data and parallel data. This, in turn, means that the dataset is not balanced and results in better performance improvements for high-resource languages compared to low-resource languages.

We used the data released for this shared task to fine-tune our models. Namely, the two training sets created from the Common Voice and MASRI Maltese speech corpora. Subsets from these larger corpora were extracted, 5 hours and 11 minutes from the verified Common Voice data and 6 hours and 39 minutes from the MASRI-Headset corpus. The transcription of each sample was translated. Fine-tuning Whisper for speech translation requires audio for input and the transcription of that audio in sentence form as a target. We pre-processed the input text so that numbers were written in words and no punctuation or capitalization was included.

Given how they were acquired, we note the difference between the subset released from the MASRI corpus and the CommonVoice dataset. While the MASRI corpus provides clean and nearly noise-free audio samples, CommonVoice samples vary in terms of different noises and the quality of audio-capturing devices.

4 Experiments

In this section, we briefly describe different experiments we conducted for our submissions, including those we did not submit for evaluation. First, we describe our experiments with the wav2vec2-xls-r (Babu et al., 2021) model, followed by the Whisper-based (Radford et al., 2022) models. We utilized HuggingFace (Wolf et al., 2020) libraries for our experiments.

4.1 SpeechEncoderDecoder models

A SpeechEncoderDecoder model is an encoder-decoder-based model used for spoken language translation or transcription, where the encoder is used to process the speech, and a language model as a decoder generates the text in the target language. In our experiments, we use the wav2vec2-xls-r model with 2B parameters as our encoder, with BERT (Devlin et al., 2019), and mBART50 (Tang et al., 2020) as the decoder following the approach in Wang et al..

4.1.1 BERT based decoder

We utilize the base BERT (Devlin et al., 2019) model as our baseline model for our SpeechEncoderDecoder approach. Namely, we use `bert-large-uncased`[1] as our language model for the decoding since the evaluation strategy does not factor in casing or punctuations. For training and inference, we add cross attention to our decoder using the `BertConfig` class from the transformers

[1]https://huggingface.co/google-bert/bert-large-uncased

Submission Name	BLEU	ASR WER
KIT.st-unconstrained-Primary	58.9	0.0835
KIT.st-unconstrained-Contrastive1	55.2	
KIT.st-unconstrained-Contrastive2	56.2	
UM.st-unconstrained-Primary	52.4	0.1431
UM.st-unconstrained-Contrastive1	52.4	0.1431
UM.st-unconstrained-Contrastive2	52.3	0.1431
UM.e2e-unconstrained-Primary	35.1	
UM.e2e-unconstrained-Contrastive1	18.5	

Table 1: Official results for the IWSLT'24 shared task, as released by organizers.

library. We did not submit results from this experiment as the models failed to produce any output during inference.

4.1.2 mBART based decoder

For our mBART-based decoding approach, we utilize the model fine-tuned for translating from 49 languages to English as released by (Tang et al., 2020). Since Maltese has a large vocabulary that is shared with Arabic and Italian, we decided to use this model instead of the vanilla mBART model. We indicate outputs from this system as the contrastive system for our work.

4.2 Whisper model

For our main system, we fine-tuned the `whisper-large-v3` model on the released dataset. As mentioned in previous sections, we also utilize several pre-processing steps for our dataset while fine-tuning.

5 Results & Discussion

In this section, we discuss the results from both submissions. As per the participation instruction, the results are reported individually for the CommonVoice subset, MASRI subset and the combined testset.

Table 1 provides the official results for different submission to the Maltese->English track for the IWSLT Low-Resource SLT shared task. Our `whisper` based submission performed consistently better than our `SpeechEncoderDecoder` model based on wav2vec2-xls-r (Babu et al., 2021) and mBART (Tang et al., 2020).

A rudimentary manual analysis of our constrained system shows a common theme of repeated phrases across some bad translations. For example, for the file MSRTS_M_03_TS_00016.wav, our contrastive system produced "`our words are not 'as it were', the people's words are not 'as they should be', our words are not 'as they should be', our words are not 'as they should be'`", whereas for the file MSRTS_M_09_TS_00008.wav, it produced "`and he comes running`" repeated 8 times. We did not find any conclusive pattern of this repetition based on the output text length, as in some instances, we find that only a sub-phrase is repeated one or more times towards the end of the output. We experimented with different output token lengths while debugging this behaviour, but it did not yield any conclusive reason, as it was present while using different inference strategies as well. Another approach to fix this behaviour would be a post-fix approach where we automatically fix the output with repeated substring search. In this study, we did not utilize such an approach and left it for future work.

We analyzed the performance of our primary submission method using speech in a code-switched conversation (Hindi and English) and found Whisper auto-translating the Hindi part to English in a few instances when we put the input language as "en". The nature of these fixes is not deterministic in this preliminary experiment, as we saw different segments translated in different runs. However, due to the end-to-end nature of our approaches, we are uncertain if this is the case with our model as well. We attribute the improved performance of the Whisper model to the increased pre-training of the model on more data than wav2vec2-xls-r. However, without inspection of the data and the high domain sensitivity in Maltese, it remains difficult to quantify the effect.

We also note that our models' performance on the testset closely resembles the results we obtained on the dev set during our training. Our primary model scored **35.9** on the dev set, whereas it scored **35.1** on the testset; similarly, our contrastive model scored **18.5** on both dev and test splits.

6 Conclusion & Future work

In our end-to-end translation system experiments, we report that the Whisper-based model outperforms our SpeechEncoderDecoder model. The performance of our contrastive model is much worse for the MASRI subset than that of the CommonVoice subset. We report that multi-lingual pre-training and fine-tuning can provide good-quality translation output in an end-to-end approach. We also report that since Whisper is already trained in a semi-supervised manner, the model output had

to be re-processed to produce the ideal results for this work. Overall, the results are much better compared to Williams et al. (2023a) for IWSLT 2023. However, we note that in the previous edition, the test dataset drastically differed in quality and domain compared to the test set for this year's shared task. However, it is not possible to draw a parallel for this comparison as the test set in the IWSLT'23 edition consisted of a podcast episode, which is more colloquial in nature. It also suffered from poor ASR outputs as there were instances of speakers talking over each other. Another hypothesis is that while much of the training data for the MT part of the previous submission contains legal domain data, which has more influence from Italian, the colloquial speak has more influence from Arabic.

Based on the performance of our SpeechEncoderDecoder model, we hypothesize that data augmentation and combining parallel data from Arabic and Italian may improve the models' performance. We aim to extend this study with an analysis of gain/drop in performance when using a fine-tuned mBART50 as a translation system from Arabic and Italian to English, compared to using the same model as the decoder in this encoder-decoder setting. We also aim to investigate the auto-translation capabilities of Whisper-based models by using them in a pipeline-based approach as well.

Furthermore, using language-specific adapters to leverage models trained only on ASR or NMT data enables SLT in low-resource contexts. Previous work on this area (Escolano et al., 2021), (Le et al., 2021), including previous submissions to IWLST (Gow-Smith et al., 2023) achieved high BLEU scores and found that this method works particularly well in low-resource contexts. We also aim to explore this approach in the future with related languages such as Italic and Arabic, as it shows promise for Maltese-English SLT.

Acknowledgments

We acknowledge the LT-Bridge Project (GA 952194).

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondr ˇej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Esteve, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, David Javorsky, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr Ojha, John E Ortega, Proyag Pal, Juan Pino, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, and Matthias Sperber. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 Evaluation Campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2015. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.

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

John E. Ortega, Rodolfo Zevallos, and William Chen. 2023. 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.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Carlos Segura. 2021. Enabling Zero-shot Multilingual Spoken Language Translation with Language-Specific Encoders and Decoders. ArXiv:2011.01097 [cs].

Edward Gow-Smith, Alexandre Berard, Marcely Zanon Boito, and Ioan Calapodescu. 2023. NAVER LABS Europe's Multilingual Speech Translation Systems for the IWSLT 2023 Low-Resource Track. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 144–158, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Santosh Kesiraju, Karel Beneš, Maksim Tikhonov, and Jan Černocký. 2023. BUT Systems for IWSLT 2023 Marathi - Hindi Low Resource Speech Translation Task. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 227–234, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Antoine Laurent, Souhir Gahbiche, Ha Nguyen, Haroun Elleuch, Fethi Bougares, Antoine Thiol, Hugo Riguidel, Salima Mdhaffar, Gaëlle Laperrière, Lucas Maison, Sameer Khurana, and Yannick Estève. 2023. ON-TRAC Consortium Systems for the IWSLT 2023 Dialectal and Low-resource Speech Translation Tasks. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 219–226, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight Adapter Tuning for Multilingual Speech Translation. ArXiv:2106.01463 [cs].

Jonathan Mbuya and Antonios Anastasopoulos. 2023. GMU Systems 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 269–276, Toronto, Canada (in-person and online). Association for Computational Linguistics.

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.

Balaji Radhakrishnan, Saurabh Agrawal, Raj Prakash Gohil, Kiran Praveen, Advait Vinay Dhopeshwarkar, and Abhishek Pandey. 2023. SRI-B's Systems for IWSLT 2023 Dialectal and Low-resource Track: Marathi-Hindi Speech Translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 449–454, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Akshaya Vishnu Kudlu Shanbhogue, Ran Xue, Soumya Saha, Daniel Zhang, and Ashwinkumar Ganesan. 2023. Improving Low Resource Speech Translation with Data Augmentation and Ensemble Strategies. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 241–250, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

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. *arXiv preprint*. ArXiv:2207.04672 [cs].

Changhan Wang, Anne Wu, Juan Miguel Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. 2021. Large-scale self- and semi-supervised learning for speech translation. In *Interspeech*.

Aiden Williams, Kurt Abela, Rishu Kumar, Martin Bär, Hannah Billinghurst, Kurt Micallef, Ahnaf Mozib Samin, Andrea DeMarco, Lonneke van der Plas, and Claudia Borg. 2023a. 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.

Aiden Williams, Andrea Demarco, and Claudia Borg. 2023b. The applicability of Wav2Vec2 and Whisper for low-resource Maltese ASR. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 39–43.

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.

Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiatong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig, and Shinji Watanabe. 2022. CMU's IWSLT 2022 Dialect Speech Translation System. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 298–307, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Jinyi Yang, Amir Hussein, Matthew Wiesner, and Sanjeev Khudanpur. 2022. JHU IWSLT 2022 Dialect Speech Translation System Description. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 319–326, Dublin, Ireland (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.