UM IWSLT 2024 Low-Resource Speech Translation: Combining Maltese and North Levantine Arabic

Sara Nabhani **Aiden Williams** **Miftahul Jannat**

Kate Rebecca Belcher **Melanie Galea** **Anna Taylor**

Kurt Micallef **Claudia Borg**

Department of Artificial Intelligence, University of Malta
{sara.nabhani.23, aiden.williams.19, miftahul.jannat.23, kate.belcher.23, melanie.galea.20, anna.taylor.23, kurt.micallef, claudia.borg}@um.edu.mt

Abstract

The IWSLT low-resource track encourages innovation in the field of speech translation, particularly in data-scarce conditions. This paper details our submission for the IWSLT 2024 low-resource track shared task for Maltese-English and North Levantine Arabic-English spoken language translation using an unconstrained pipeline approach. Using language models, we improve ASR performance by correcting the produced output. We present a 2 step approach for MT using data from external sources showing improvements over baseline systems. We also explore transliteration as a means to further augment MT data and exploit the cross-lingual similarities between Maltese and Arabic.

1 Introduction

There are a variety of challenges inherent in spoken language translation for low-resource languages. By definition, these languages have very limited data available to use for natural language processing (NLP) tasks. The majority of current work on NLP targets just 20 out of the approximately 7,000 languages used worldwide, leading to a significant gap in research and negative impacts on excluded speech communities (Joshi et al., 2020; Magueresse et al., 2020). While most machine learning tasks are performed using vast amounts of data, models for low-resource languages must be adapted to work with less data or other strategies must be employed to augment the existing data.

In this paper, we present our submission for the 2024 IWSLT low-resource shared task. Concretely, we submit two systems for speech translation to English, from Maltese and North Levantine Arabic. The main motivation for focusing on these two languages is their similarity to one another which we aim to exploit to improve the performance of our pipeline speech translation system.

As a well-known case of diglossia, the Arabic language has a notable distinction between the formal variety used in written communication, political speech, and the educational system – known as Modern Standard Arabic (MSA) – and the informal varieties primarily used in spoken communication – collectively referred to as Dialectal Arabic (DA). These dialects exhibit considerable diversity influenced by geographical and socio-economic factors, diverging significantly from MSA in phonology, morphology, lexicon, and syntax (Zbib et al., 2012). On the other hand, Maltese is a Semitic language derived from Siculo-Arabic (Borg and Azzopardi-Alexander, 1997), with a notable mutual intelligibility with Tunisian DA (Čéplö et al., 2016). Its evolution independently from the Arab world – particularly its substantial influence from Italian and English and its use of a modified Latin alphabet – makes it a distinct language and not an Arabic dialect.

We split the task of spoken language translation into two sequential tasks consisting of automatic speech recognition (ASR) and machine translation (MT). This process transforms speech in a low-resource language into text in a high-resource language, namely English for this shared task. At the same time, splitting the task into ASR and MT allows us to exploit existing multilingual models and source larger corpora for each sub-task to improve their performance in the target language. All of our code is made publicly available.[1]

2 Related Work

In order to examine past approaches to low-resource spoken language translation, this literature review includes an overview of previous IWSLT low-resource track submissions and our ASR and

[1] https://github.com/saranabhani/iwslt-2024-um-pipeline

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 145–155

MT systems as well as our innovative approach to data augmentation through transliteration.

2.1 Previous IWSLT Low-Resource Track Approaches

For the 2023 IWSLT Shared Task (Agarwal et al., 2023), Williams et al. (2023a) submitted five systems for Maltese-English spoken language translation as part of the low-resource track in the unconstrained setting. This marked the first time that Maltese was included in the IWSLT low-resource track campaign, with this submission being the sole entry in its category, making it a unique approach in this context. All of the systems employed a pipeline approach, making use of XLS-R (Conneau et al., 2020) for ASR and mBART-50 (Tang et al., 2020) for MT, fine-tuned using various training data. In their primary approach, their model was exclusively fine-tuned on Maltese data resulting in a BLEU score of 0.6. Contrasting with the Maltese-only model, they explored four alternative approaches by incorporating corpora from Arabic, French, Italian, or a combination of all three in conjunction with the Maltese data. The most successful configuration, with a BLEU score of 0.7, was achieved by fine-tuning the ASR system on a combination of Maltese data with 50 hours each of Arabic, French, and Italian data from the Common-Voice speech corpus (Williams et al., 2023a).

While our submission utilizes a pipeline approach, the alternative is an end-to-end system where a single neural network is trained to jointly perform both ASR and MT (Sethiya and Maurya, 2023). This approach offers several advantages by significantly reducing training time, allowing for quicker development of models, and necessitating lower memory resources compared to other methods, which can be particularly beneficial for environments with constraints on computational processing power. Additionally, by integrating ASR and MT into an end-to-end system, it mitigates the risk of errors propagating from the ASR output to the MT input, which is a common problem in pipeline systems (Sethiya and Maurya, 2023). However, speech translation systems that operate end-to-end require parallel data containing both speech audio signals on the source side and translated transcriptions on the target side. Acquiring such parallel data can pose challenges, even for languages with readily available components for pipeline-based systems. Consequently, the pipeline approach is often deemed more feasible and realistic (Alves et al., 2020).

For the 2023 IWSLT Quechua-Spanish speech translation task in the low-resource track, E. Ortega et al. (2023) utilized a variety of systems both constrained and unconstrained, with one of the few pipeline-based methods submitted for this task. The primary constrained system employed a direct speech translation model based on the Fairseq speech-to-text (S2T) framework (Wang et al., 2020). To create audio representations, this system made use of log mel-scale filter banks for features and a transformer for translations. With a BLEU score of 1.25, their primary system surpassed the performance of the pipeline alternatives in the constrained setting. On the other hand, the primary unconstrained system employed a pipeline approach on the additional 60 hours of speech data made available, where speech transcriptions were generated using a pretrained XLS-R based multilingual model augmented by a fine-tuned language model (Park et al., 2019), and translations were generated using the fine-tuned Flores-101 model from Guzmán et al. (2019). The unconstrained pipeline approach performed much better with a BLEU score of 15.36 for the primary model. Their findings reveal that the use of a pretrained language model with fine-tuning is necessary for cascaded spoken language translation (ASR and MT combined in a pipeline) in low-resource scenarios for Quechua to Spanish translation. This work further demonstrates the immense value of access to additional data, which yielded nearly 14 BLEU points improvement for the unconstrained task when applied to both ASR and MT systems compared to the limited data used in the constrained setting (E. Ortega et al., 2023). Accordingly, our approach also utilizes an unconstrained pipeline of an XLS-R-based ASR model and a fine-tuned pretrained MT model considering it was found to have the best results for this submission.

2.2 Automatic Speech Recognition

Due to their extensive multilingual pretraining, Wav2Vec 2.0 models (Baevski et al., 2020) are able to acquire and utilize cross-lingual speech representations to improve accuracy for ASR. The XLS-R model presented by Babu et al. (2022) underwent pretraining of Wav2Vec 2.0 models for as many as 128 distinct languages including Maltese, using 436,000 hours of unannotated speech data from diverse sources including the Mozilla Common Voice (Ardila et al., 2020), BABEL (Gales et al., 2014),

and Multilingual LibriSpeech (Pratap et al., 2020) speech corpora. Specifically, this system incorporates 9,000 hours of unannotated Maltese speech sourced from the Voxpopuli corpus (Wang et al., 2021). Notably, the largest model is pretrained using a cumulative total of 56 thousand hours of speech data (Conneau et al., 2020). This represents an increase in both the amount of data and the languages covered.

A common practice in the field of ASR is to use a language model to reduce errors in the generated transcription. This technique was used in the development of Wav2Vec 2.0 (Baevski et al., 2020) and Deepspeech 2 (Amodei et al., 2016). Leveraging an external language model trained on domain-specific textual data has the potential to increase the accuracy of ASR systems by minimizing errors in content.

Moreover, due to the scarcity of high-quality labelled data in DA, the models based on XLS-R emerge as optimal solutions for leveraging available datasets and adapting ASR to distinct Arabic variants through fine-tuning, as highlighted by Waheed et al. (2023) in their work on VoxArabica. These models not only capitalize on existing resources but also offer the adaptability to accommodate the nuances of various Arabic dialects, thus addressing the challenges associated with the limited availability of labelled data for DA.

2.3 Machine Translation

Past approaches to multilingual neural machine translation treat it as a sequence-to-sequence task, where an encoder is utilized to process an input sequence in the source language and a decoder is used to generate the corresponding output sequence in the target language. With massively multilingual translation, a model undergoes training on multiple translation directions simultaneously. While this approach can facilitate advantageous cross-lingual transfer among related languages, it also carries the risk of amplifying interference between unrelated languages.

In this work we make use of the NLLB model (NLLB Team et al., 2022) for MT. It uses a single SentencePiece model to tokenize the text sequences by training it across all languages using a total of 100M sentences sampled from primary bitext data. For equitable representation of low-resource languages, high-resource ones are downsampled and low-resource ones are upsampled, using a sampling temperature of five. The resulting vocabulary size of the trained SentencePiece model is 256,000, ensuring comprehensive representation across the diverse range of supported languages. The choice of this model is highly motivated by its inclusion of a large number of languages, notably Maltese and North Levantine Arabic (NLLB Team et al., 2022).

2.4 Transliteration

From a simplified linguistic perspective, Maltese can be regarded as a variant of Arabic with a significant level of code-switching to Italian and a modified Latin alphabet. Past work suggests that transliterating Maltese could serve as a viable strategy for benefiting from cross-lingual similarities with Arabic (Micallef et al., 2023). In the approach taken by Micallef et al. (2023), the transliteration process involves two main steps: mapping and ranking. Initially, Maltese text tokens and characters in Latin script are mapped to one or more corresponding alternatives in Arabic script. Subsequently, a separate component either ranks these alternatives or employs a deterministic hard-coded baseline.

This approach is further developed in Micallef et al. (2024) by taking a mixed pipeline and integrating a combination of transliteration and translation based on the etymology of Maltese words. This is motivated by the results of Micallef et al. (2023), where the advantages of transliterating Arabic-origin words were limited by the corresponding disadvantages of distancing Italian and English-origin words from their etymological source through transliteration. A mixed pipeline gave promising results on downstream tasks, establishing the technique as a competitive approach for Maltese NLP tasks.

3 Automatic Speech Recognition

3.1 Data Sources

For Maltese, we use the training sets provided by the shared task namely Common Voice 7.0 (Ardila et al., 2020) and MASRI (Hernandez Mena et al., 2020). The speech corpus is made up of around 50 hours of Maltese speech data.

To train our Arabic ASR system, we opted to use a 50-hour subset from the Common Voice project (Ardila et al., 2020), as this would contain roughly the same data that we used for Maltese ASR. While a training set for North Levantine Arabic would have been preferred, there was no data provided for the shared task, nor were we able to find ASR data for North Levantine Arabic. Furthermore, even

though a Tunisian Arabic training set could be used, we did not make use of this to train, since North Levantine Arabic is more closely related to MSA than Tunisian Arabic (Kwaik et al., 2018).

3.2 Approach

For the ASR component of the pipeline, we continue to build off of previous work done for both DA and Maltese ASR. As concluded in Williams et al. (2023b), fine-tuning the Wav2Vec 2.0 XLS-R model (Babu et al., 2022) with around 50 hours of Maltese speech data produces the best Maltese ASR model to date and was used in the IWSLT 2023 submission by Williams et al. (2023a). A similar XLS-R based ASR model is employed for DA by leveraging data for MSA.

In addition, for Maltese, we incorporate language models with the ASR system to get more accurate speech transcriptions. For this we use n-gram models built using the KenLM language modelling toolkit (Heafield, 2011), which assign scores to sequences of words. This aids in selecting the best candidates through beam search for improved ASR output. We use KenLM mainly as it has been used for other state-of-the-art ASR publications as well as in previous work on Maltese in particular. We make use of the 6-gram word-level LM produced by Hernandez Mena et al. (2020) as a baseline to compare our own KenLM n-gram models which was trained on Korpus Malti v3.0 (Gatt and Čéplö, 2013). We produce 2 additional word-level n-gram language models for Maltese: a 3-gram and a 4-gram, both trained on the Korpus Malti v4.1 Shuffled train dataset[2] (Micallef et al., 2022). We note that Korpus Malti v4 used here is substantially larger than the v3 used for the 6-gram baseline.

3.3 Results and Discussion

Table 1 shows the WER score for all languages considered on the shared task development set. For both Maltese and North Levantine Arabic, a single model is trained, but for Maltese we show the models' performance without adding a language model as well as incorporating each language model.

For Maltese, we see that all models perform comparably, but models using a language model give better results. In addition, when using the 3-gram and 4-gram models, these give better results than the 6-gram model, which we attribute to the larger data used to train the former models.

[2] https://huggingface.co/datasets/MLRS/korpus_malti/tree/4.1.0/data/shuffled

Data	Language Model	Dev Set WER ↓
CV+MASRI	-	0.12
CV+MASRI	3-gram	**0.10**
CV+MASRI	4-gram	**0.10**
CV+MASRI	6-gram	0.11

(a) Maltese

Data	Language Model	Dev Set WER ↓
Common Voice	-	**1.08**

(b) North Levantine Arabic

Table 1: Speech Recognition Results

The overall performance of our Arabic approach was limited by the lack of North Levantine Arabic speech data, which severely impacted the accuracy of the ASR system when tested on Levantine data. We provide a brief qualitative error analysis of the Arabic ASR outputs to highlight this.

We looked at a sample of the ASR output generated from North Levantine Arabic audio data using our model trained on MSA. The analysis of specific examples reveals various errors that significantly impact the usability of the ASR system for Levantine speech recognition. Table 2 shows a few examples of the output, highlighting various inconsistencies with the reference text.

Phonetic errors were a common issue across the examples. For instance, in Sample 2c, the system outputted "سأدمت" for "فقدمت" (I applied) likely due to the similar pronunciation of "س" and "ف" at the start of word, and the dialect-specific pronunciation of "ق" (qaf) as a glottal stop [ʔ] in Levantine Arabic, which is similar to "ء" (hamza). Additionally, in Sample 2a, segmentation errors featured prominently, as seen where "هييكان" should have been segmented into "هي كان" (she was). In Sample 2d, "تأريبالنامبل شتشويالأمور" exemplifies improper segmentation, where "شوي" (a little) and "الامور" (matters) were incorrectly merged as "شتشويالأمور".

Lexical errors were evident, particularly in Sample 2b, where "التشيكيين" (the Czechs) was incorrectly outputted as "تشكيم", missing both the prefix "ال" (the) and misinterpreting the main noun due to a phonetic mix-up of "م" and "ن", which are both nasal consonants. Phonetic confusion also occurred

Reference	تشغيل	مسابقة	اسمها	كان	هي
Transcription	[tʃaɣiːl]	[musaːbaʔat]	[ʔismeːha]	[kaːn]	[hiyye]
ASR Output	تشغيل	مثمة	اسما	هييكان	
Transcription	[tʃaɣiːl]	[maθmat]	[ʔisma]	[hiyyekaːn]	

(a)

Reference	التشيكيين	عن	احكي	راح	مش	
Transcription	[ittʃiːkijjiːn]	[ʕan]	[ʔaħki]	[raːħ]	[miʃ]	
ASR Output	تشكيم	عن	مرحك			
Transcription	[tʃakiːm]	[ʕan]	[marħak]			

(b)

Reference	عليها	فقدمت	
Transcription	[ʕalejha]	[faqaddamat]	
ASR Output	علي	سأدمت	
Transcription	[ʕali]	[saʔaddamat]	

(c)

Reference	سوريا	في	الأمور	شوي	بلشت	تقريبا	ما
Transcription	[suːrja]	[fiː]	[ilʔumuːr]	[ʃwayy]	[ballaʃat]	[taqriːban]	[ma]
ASR Output	سورية	بي	شتشويالأمور		تأريبالنمبل	ما	
Transcription	[suːrja]	[biː]	[taʔriːbaːlnmbal]	[ʃatʃwaːilʔumuːr]		[ma]	

(d)

Table 2: Reference transcription samples compared to the system output produced by our ASR system

in Sample 2d, where "في [fi:]" (in) was replaced with "بي [bi:]". In the case of "سوريا" (Syria), the ASR output was "سورية", only differing by the final character. These two characters have the same pronunciation in word-final position, so the difference is just orthographic.

The analysis revealed that the Character Error Rate (CER) was consistently better than the Word Error Rate (WER), highlighting that while individual characters are often recognized correctly, the system struggles to assemble these into correct word forms. This indicates foundational competence at the character level but significant challenges in managing the complexity of word formation, especially considering the morphological and contextual nuances of North Levantine Arabic.

Whilst some character-level errors seem to be due to similar phonetic characteristics of different characters, it is clear that multiple errors can be attributed to Levantine-specific dialectal differences, most prominently the "ق" (qaf) and "ء" (hamza) distinction. These character-level errors impact word-level recognition and subsequent performance on the downstream machine translation task.

The prevalence of errors due to dialectal differences underscores the need to integrate Levantine-specific training data and develop a dedicated language model to handle the nuances brought by dialectal variations.

4 Machine Translation

4.1 Data Sources

To train our translation models we make use of a variety of sources for parallel data including those provided for the shared task as well as others which we could find. The datasets used are summarized in Table 3.

To train our Maltese translation model, we used a combined dataset of Common Voice (CV) (Ardila et al., 2020) and MASRI project (Hernandez Mena et al., 2020) (henceforth referred to as CV+MASRI), both of which were the datasets provided officially for the shared task. In addition, we also used OPUS-100, which is a comprehensive English-focused dataset (Zhang et al., 2020; Tiedemann, 2012). The dataset consists of 100 languages and English is common in every 99 translated language pairs. We chose this dataset because of its vastness, especially considering it offered 1M parallel sentences for the English and Maltese pair. We preprocessed the data to drop any data points that were empty as well as duplicate instances.

For our Arabic translation systems, we utilized a range of datasets. Specifically, we used the North Levantine (APC)-MSA-English textual data provided for the task (Sellat et al., 2023), along with the IWSLT 2022 Tunisian Arabic (AEB) speech translation data (Anastasopoulos et al., 2022). Additionally, the MSA data, which was included with both the Tunisian and Levantine datasets, was also used. However, since the size of this data was miniscule, we also incorporated the Arab-Acquis MSA-English parallel data (Habash et al., 2017). To further augment the Arabic data, we also incorporated the CV+MASRI Maltese dataset, which was transliterated to match the script of our primary data as detailed in Section 4.2 (referred to as MLT_{ARA}). We merged datasets from the same dialect or language obtained from multiple sources and shuffled them to ensure diversity and randomness in our training process.

Since the speech transcriptions do not produce casing and punctuation information, and the evaluation for the shared task also ignores these features, we preprocess all translation data as such. For both

Dataset	Train Size	Validation Size
CV	3,773	1,235
MASRI	4,811	648
CV+MASRI	8,584	1,883
OPUS-100	672,196	-

(a) Maltese Model

Language/Dialect	Train Size	Validation Size
APC	99,519	21,081
AEB	173,612	-
MSA	133,074	-
MLT_{ARA}	8,886	1,883

(b) Arabic Model

Table 3: Data used to train the MT models and size in number of sentences

languages, preprocessing included text normalization such as converting to lowercase and removing punctuation, while retaining hyphens and apostrophes for Maltese datasets, as these characters hold linguistic significance in Maltese. In addition, for Arabic we also remove diacritics.

4.2 Transliteration

Following Micallef et al. (2023, 2024), we explored integrating transliteration of Maltese into Arabic script, due to the close relationship between Maltese and Arabic as Semitic languages. We took inspiration from this approach to supplement the data used for training the Arabic Machine Translation system. Since Micallef et al. (2024) saw more promising results when using a mixed pipeline of transliteration, that involved transliterating Maltese words of Arabic origin and translating the other words, we continue with this mixed approach. We utilize the etymology model and mapping systems from Micallef et al. (2024). Specifically, we follow the $\text{X}_{\text{ara}}/\text{T}_{\text{ara}}$ pipeline, which transliterates tokens of Arabic-origin and symbols, translating everything else to Arabic.

However, we make certain modifications to this to better suit our approach. Firstly, we modify the translation component by swapping out the precomputed word translations from Google Translate with a pretrained NLLB model (NLLB Team et al., 2022), as extracting translations using Google Translate was too expensive, especially considering the different outputs produced by the ASR while experimenting. Translation is performed into Tunisian Arabic (AEB) instead of MSA, using English as a pivot language. The reason for doing this is that translating through English generally yields better results rather than going directly to Arabic, due to the larger availability of parallel data, and this is also observed empirically in Micallef et al. (2024).

Secondly, we merge tokens to more closely reflect the way in which Arabic is written, reducing the signals from Maltese tokenization. For example, "u il-kelma" (English 'and the word'), are written together in Arabic script as one word, "والكلمة", where "و" is the conjunction corresponding to "u" (and), "ال" is the definite article corresponding to "il-" (the), and the rest of the word corresponds to "kelma" (word). The annotation for such token mappings from Micallef et al. (2023), includes special markers indicating that such words would be merged in Arabic, so given the 3 tokens *u*, *il-*, and *kelma*, the system would initially output و+, ال+, and كلمة, which we merge into a single word. While Micallef et al. (2023, 2024) ignore this signal as they mostly deal with token tagging tasks, we use this signal to merge words. Note that using this method, punctuation symbols are still space separated, but since the data is preprocessed to remove such symbols, this is not an issue in our case.

We applied the transliteration pipeline to the Maltese datasets provided for the shared task (CV+MASRI). The training dataset provided additional data for training the Arabic MT model. In addition to the data augmentation benefit of integrating transliterated Maltese (henceforth referred to as MLT_{ARA}) for Arabic-English MT, this also increases the cross-lingual capacities of our Arabic MT model, allowing for the evaluation of MLT_{ARA} ASR outputs using the Arabic MT model.

4.3 Approach

We explored various machine translation (MT) systems for translating North Levantine Arabic and Maltese into English. Initially, we established a baseline by fine-tuning the NLLB 1.3B model[3] (NLLB Team et al., 2022) on the shared task data, specifically on the CV+MASRI dataset for Maltese and the AEB dataset for Arabic.

Subsequently, we experimented with different fine-tuning strategies. We first attempted a two-

[3] https://huggingface.co/facebook/nllb-200-1.3B

Fine-Tuning Data		Dev Set
Stage 1	Stage 2	BLEU ↑
-	CV+MASRI	60.3
OPUS-100	-	37.6
OPUS-100	CV+MASRI	**60.6**
MSA	APC+AEB+MLT_{ARA}	37.0

(a) Maltese

Fine-Tuning Data		Dev Set
Stage 1	Stage 2	BLEU ↑
-	APC	34.3
MSA	APC	**39.5**
MSA	APC+AEB	37.6
MSA	APC+AEB+MLT_{ARA}	37.4

(b) Levantine Arabic

Table 4: Machine Translations Results

stage fine-tuning process where we fine-tune with a large dataset from a different domain or dialect. For the first stage, we considered the OPUS-100 data for the Maltese model and the MSA data for the Arabic model, while the second stage included the same data used for the baseline for both languages. Additionally, for the Arabic we tested fine-tuning with a mix of Levantine (APC) and Tunisian (AEB) data, as well as a combination of Levantine (APC), Tunisian (AEB), and transliterated Maltese (MLT_{ARA}) data. The training on MLT_{ARA}, allows us to evaluate this system on both the Maltese and North Levantine development sets.

The same hyperparameters were applied across both MT systems: a learning rate of 2e-5, and a weight decay of 0.01. The training was conducted over three epochs.

4.4 Results and Discussion

Table 4 reports the BLEU scores of the Maltese and Arabic models on the transcriptions having reference translations from the respective development sets for Maltese and North Levantine Arabic.

The results for Maltese are reported in Table 4a. We see that fine-tuning using OPUS-100 only, is detrimental compared to the baseline system trained only on CV+MASRI. However, including both OPUS-100 and CV+MASRI yields the best performance. Furthermore, when evaluating the Arabic model trained on transliterated Maltese, in addition to other Arabic data, we observe that it is the worst-performing model. However, the performance is quite comparable to that obtained for the model fine-tuned only on OPUS-100.

Table 4b shows the performance of each of the experimented machine translation systems on the APC validation set. Among the experimented methods, the best performance on the North Levantine Arabic development set was achieved using the two-stage fine-tuning process that started with MSA data followed by Levantine data.

An important observation that arose from our experimentation was the impact of adding MLT_{ARA} data to the training of the Arabic MT system. We can see that the system fine-tuned firstly on MSA data and subsequently on APC, AEB, and MLT_{ARA} gave very competitive results for Arabic MT, with a difference in dev performance of just 0.2 BLEU compared to the same system without MLT_{ARA} data. By comparison, the same system performed comparably to Arabic on the Maltese dev set (after being transliterated) with a BLEU of 37.0. Whilst the machine translation systems fine-tuned specifically for Maltese still significantly outperformed the Arabic system fine-tuned with MLT_{ARA} data, we note that adding MLT_{ARA} data in the fine-tuning of Arabic MT systems can vastly improve the cross-lingual capacity of the model, with substantial benefits to the performance on Maltese, and very little impact on the MT performance for Arabic.

5 Speech Translation Pipeline

Following our evaluation on individual tasks in Sections 3 and 4, we now combine both systems by first getting the transcription using an ASR system and then passing this transcription through the MT system to get the translation. The best-performing ASR and MT systems on our validation sets were selected for the pipelines.

For Maltese, we only choose the MT system trained with the 2 stage training, OPUS-100+CV+MASRI (which we refer to as MLT_{LAT}) and combine it with the ASR systems with the 3-gram, 4-gram, and 6-gram models, to compose our Primary, Contrastive 1, and Contrastive 2 systems, respectively. For North Levantine Arabic, we use the only trained model for ASR, paired with all 3 MT systems which made use of 2 stage training, namely MSA+APC+AEB, MSA+APC+AEB+MLT_{ARA}, and MSA+APC, to compose our Primary, Contrastive 1, and Contrastive 2 systems, respectively. Table 5 summarises the results obtained with these pipelines on the development and testing sets, on ASR only and Speech Translation (ASR+MT). For the Arabic test

Pipeline	ASR System	MT System	Dataset	Dev Set		Test Set	
				WER ↓	BLEU ↑	WER ↓	BLEU ↑
Primary	3-gram	MLT_{LAT}	CV	0.098	**58.4**	0.094	**60.9**
			MASRI	0.239	**42.9**	**0.233**	**43.9**
			CV+MASRI	**0.10**	**52.1**	**0.143**	**52.4**
Contrastive 1	4-gram	MLT_{LAT}	CV	0.097	**58.4**	0.094	**60.9**
			MASRI	0.239	**42.9**	**0.233**	**43.9**
			CV+MASRI	**0.10**	**52.1**	**0.143**	**52.4**
Contrastive 2	6-gram	MLT_{LAT}	CV	**0.096**	58.3	**0.093**	**60.9**
			MASRI	**0.238**	42.7	0.234	43.7
			CV+MASRI	0.11	51.9	**0.143**	52.3

(a) Maltese

Pipeline	ASR System	MT System	Dev Set		Test Set		
			WER ↓	BLEU ↑	BLEU ↑	COMET ↑	ChrF ↑
Primary	Common Voice	MSA+APC+AEB	**1.08**	**5.0**	4.74	53.69	24.10
Contrastive 1	Common Voice	MSA+APC+AEB+MLT_{ARA}	**1.08**	4.8	**5.09**	**53.78**	**24.50**
Contrastive 2	Common Voice	MSA+APC	**1.08**	3.7	3.53	51.96	21.56

(b) North Levantine Arabic

Table 5: Speech Translation Pipeline Results

set, only Speech Translation results were provided.

As seen in Table 5a, all Maltese systems perform competitively with each other. Similar to the findings for the ASR system reported in Section 3, the Primary and Contrastive 1 systems get the best results with the 3-gram and 4-gram models, and the Contrastive 2 system is slightly behind with the 6-gram model. The best systems obtain 52.4 BLEU on the test set.

The results on the North Levantine Arabic data are shown in Table 5b. The systems all achieve low overall BLEU scores, due to the poor performance on ASR as outlined in Section 3. With a pipeline, we observe that using APC data only in addition to MSA performs the worst (Contrastive 2), and that adding data from other languages and dialects we achieve better BLEU scores with the Primary and Contrastive 1 systems.

6 Conclusion

Overall, this paper presented our findings for Maltese and North Levantine Arabic spoken language translation into English with a pipeline system in the unconstrained setting for the 2024 IWSLT low-resource track shared task. For our approach we fine-tune a Wav2Vec 2.0 XLS-R model for ASR, and an NLLB model for MT. We enhance the ASR model by correcting the outputs with a language model. Moreover, we augment the MT data from additional sources and employ a two-stage fine-tuning process to improve performance. Additionally, we exploit the cross-lingual similarities between Maltese and Arabic by transliterating Maltese to Arabic script, observing interesting performance boosts.

In terms of limitations, the lack of training data for North Levantine Arabic impeded the progress of our ASR system. By using MSA to train our Arabic ASR models, the resulting system struggled with non-standard pronunciation and dialect-specific variation. Furthermore, the absence of testing data for Tunisian Arabic hindered our models considering its close similarity with Maltese.

More general improvements could be undertaken in future work such as hyper-parameter tuning and supplementing currently available data with back-translation. Rather than relying solely on parallel data, implementing backtranslation with larger monolingual corpora holds promise for improving the MT systems discussed in this paper.

Acknowledgments

We acknowledge the assistance of the LT-Bridge Project (GA 952194) and DFKI for the use of their Virtual Laboratory.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej 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 Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, 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, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Diego Alves, Askars Salimbajevs, and Mārcis Pinnis. 2020. *Data Augmentation for Pipeline-Based Speech Translation*.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Vaino Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2: end-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 173–182. JMLR.org.

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.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Albert Borg and Marie Azzopardi-Alexander. 1997. *Maltese: Descriptive Grammars*. Routledge, London and New York.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

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.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

Albert Gatt and Slavomír Čéplö. 2013. Digital Corpora and Other Electronic Resources for Maltese. In *Proceedings of the International Conference on Corpus Linguistics*, pages 96–97. UCREL, Lancaster, UK.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Nizar Habash, Nasser Zalmout, Dima Taji, Hieu Hoang, and Maverick Alzate. 2017. A parallel corpus for evaluating machine translation between Arabic and European languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 235–241, Valencia, Spain. Association for Computational Linguistics.

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.

Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. 2020. MASRI-HEADSET: A Maltese corpus for speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France. European Language Resources Association.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A lexical distance study of Arabic dialects. *Procedia Computer Science*, 142:2–13. Arabic Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Kurt Micallef, Fadhl Eryani, Nizar Habash, Houda Bouamor, and Claudia Borg. 2023. Exploring the impact of transliteration on NLP performance: Treating Maltese as an Arabic dialect. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 22–32, Toronto, Canada. Association for Computational Linguistics.

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.

Kurt Micallef, Nizar Habash, Claudia Borg, Fadhl Eryani, and Houda Bouamor. 2024. Cross-lingual transfer from related languages: Treating low-resource Maltese as multilingual code-switching. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1025, St. Julian's, Malta. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Daniel Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Cubuk, and Quoc Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020*. ISCA.

Hashem Sellat, Shadi Saleh, Mateusz Krubiński, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. UFAL parallel corpus of north levantine 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Nivedita Sethiya and Chandresh Kumar Maurya. 2023. End-to-end speech-to-text translation: A survey. *Preprint*, arXiv:2312.01053.

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

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).

Abdul Waheed, Bashar Talafha, Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. Voxarabica: A robust dialect-aware arabic speech recognition system. *Preprint*, arXiv:2310.11069.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

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.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Slavomír Čéplö, Ján Bátora, Adam Benkato, Jiří Milička, Christophe Pereira, and Petr Zemánek. 2016. Mutual intelligibility of spoken Maltese, Libyan Arabic, and Tunisian Arabic functionally tested: A pilot study. *Folia Linguistica*, 50(2):583–628.