JHU IWSLT 2025 Low-resource System Description

Nathaniel R. Robinson*[1] Niyati Bafna* Xiluo He* Tom Lupicki*
Lavanya Shankar* Cihan Xiao Qi Sun Kenton Murray David Yarowsky
Johns Hopkins University Center for Language and Speech Processing
*Equal contribution, [1]Contact author
{nrobin38,nbafna1,xhe69,tlupick1,ls1,cxiao7,qsun29,kenton,yarowsky}@jhu.edu

Abstract

We present the Johns Hopkins University's submission to the 2025 IWSLT Low-Resource Task. We competed on all 10 language pairs. Our approach centers around ensembling methods – specifically Minimum Bayes Risk Decoding. We find that such ensembling often improves performance only slightly over the best performing stand-alone model, and that in some cases it can even hurt performance slightly.

1 Introduction and Background

Despite many recent advances in deep learning and artificial intelligence, challenges in low-resource and dialectal speech translation still preclude high-quality automated translation systems for many language communities. Cross-lingual transfer and multilingual models have allowed for recent progress in scarce data settings, but performance still lags significantly behind that of higher resource languages (Ziems et al., 2023; Joshi et al., 2024).

Due to a lack of training resources, low-resource languages systems tend to generate hypotheses with higher variance than is seen in higher-resourced conditions. In other words, different models might generate diverse outputs; hence a single system might not be optimal in all scenarios. This motivates attempting to select the best option from multiple potential systems—i.e., ensembling.

For this year's IWSLT low-resource speech translation campaign (Abdulmumin et al., 2025), we, the Johns Hopkins University (JHU) team decided to focus on Minimum Bayes Risk Decoding (MBR) with the interest in exploring in-depth how combining methods across a range of language pairs can improve performance in a low-resource setting (Bickel and Doksum, 1977; Kumar and Byrne, 2004).

Following our approach from last year (Robinson et al., 2024), we submitted systems for all language pairs, with a focus on seeing how robust our

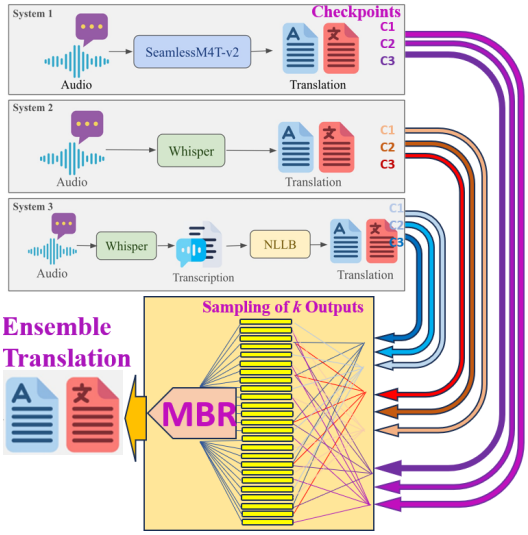

Figure 1: We apply Minimum Bayes Risk (MBR) ensembling to a variety of systems.

methods are across a wide range of data settings and typologically diverse languages. Rather than focusing on a specific language, our line of inquiry was geared towards broader exploration, with the interest of discovering language-agnostic trends in mind.

Our approach in 2024 focused on fine-tuning pre-trained models Whisper (Radford et al., 2023a), NLLB (NLLB Team et al., 2022), and SeamlessM4T v2 (Barrault et al., 2023) for both ASR+MT cascading and end-to-end speech translation. We also incorporated joint training for language pairs with common targets, as well as fine-tuning with a regularization technique known as intra-distillation (Xu et al., 2022, 2023; Robinson et al., 2024). In this year's submission, we similarly gather a variety of different systems for each translation language pair. We use these to obtain a diverse set of outputs for the language pairs sampled from various checkpoints of the different cascaded and end-to-end fine-tuned systems. To maximize variety of systems, and following our submission

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 315–323

from last year, we experimented with combining fine-tuning data for language pairs with a common target language, and with use of additional or supplementary training data. Rather than comparing these diverse systems directly, however, we ensembled them into a single inference method. We used MBR to select the best performing candidate translation from the resulting pool for each source audio.

Our decision to attempt an ensemble approach was inspired by other team submissions from last year. Our submissions performed best in last year's evaluation campaign (Ahmad et al., 2024) for Irish-to-English (gle-eng), Bemba-to-English (bem-eng), and Bhojpuri-to-Hindi (bho-hin) translation. However the teams that performed best for Levantine Arabic-to-English (apc-eng), Maltese-to-English (mlt-eng), and Quechua-to-Spanish (que-spa) all employed ensemble models. Ben Kheder et al. (2024) ensembled 26 model checkpoints for their apc-eng system, and E. Ortega et al. (2024) ensembled 10 checkpoints for their que-spa system. For their mlt-eng system, Li et al. (2024) ensembled cascade systems with end-to-end models (just as we attempt to do in this work).

2 Methodology and Experiments

Our methodology is illustrated in Figure 1. Given any language pair, we have a number of cascade and end-to-end systems (three in the figure). These systems employ either Whisper ST, Whisper ASR with NLLB MT, or SeamlessM4T v2 ST, and may have other more minor variations differentiating them. The number of systems varies between one and four, depending on the language pair. The systems we use for each language pair are listed in Table 2.

We keep three fine-tuning checkpoints from the final model of each of the systems (NLLB in the case of the bi-model cascade) and for each input audio, we sample five hypothesis translations from each system checkpoint, resulting in a total of 15 hypotheses per system for each input. In the case of language pairs like apc-eng, for which we ensembled four different systems, this amounted to $15 * 4 = 60$ hypotheses for each input audio, which are reduced to a single hypothesis via the MBR process.

2.1 Task description and data

This year's task focuses on speech translation for ten language pairs: Levantine Arabic to English (apc-eng), Tunisian Arabic to English (aeb-eng), Bemba to English (bem-eng), Fongbe to French (fon-fra), Irish to English (gle-eng), Bhojpuri to Hindi (bho-hin), Estonian to English (est-eng), Maltese to English (mlt-eng), Marathi to Hindi (mar-eng), and Quechua to Spanish (que-spa). Fongbe and Estonian are new as source languages in this year's task. Fongbe is a Gbe language of the Niger-Congo family spoken in Benin; while Estonian is a Uralic language spoken in Estonia.

In developing our systems, we utilize a combination of organizer-provided data as well as some external data. We summarize our data sources in Table 1.

Lang.	Type	Amount	Sources
apc-eng	ASR	28h	Makhoul et al. (2005)
	MT	120k lines	Sellat et al. (2023)
aeb-eng	E2E	167h	Anastasopoulos et al. (2022)
	ASR	324h	Anastasopoulos et al. (2022)
bem-eng	ST	180h	Sikasote et al. (2023)
	ASR	24h	Sikasote and Anastasopoulos (2022)
fon-fra	E2E	57h	Kponou et al. (2024)
gle-eng	E2E	11h	Agarwal et al. (2023)
bho-hin	E2E	25h	Agarwal et al. (2023)
est-eng	E2E	1262h	Sildam et al. (2024)
mlt-eng	ST	14h	CV; Hernandez Mena et al. (2020)
	MT	2.1M lines	Bañón et al. (2023, 2020)
mar-hin	E2E	30h	Agarwal et al. (2023)
	ASR	1300h	CV; He et al. (2020); Bhogale et al. (2022)
que-spa	ST	1.7h	Ortega et al. (2020)
	ASR	48h	Cardenas et al. (2018)
	MT	26k lines	Tiedemann (2012); Ortega et al. (2020)

Table 1: Data information for each language pair. "CV" refers to Common Voice (https://commonvoice.mozilla.org/).

2.2 Seamless E2E systems

We detail the different systems listed in Table 2.

The SeamlessM4T v2 model (Barrault et al., 2023) is a state-of-the-art multilingual model developed specifically for speech translation (ST). It supports both speech-to-text and speech-to-speech translation, enabling direct translation of spoken language in 143 languages. The model is trained on a large and diverse corpus that combines supervised and semi-supervised data, allowing it to perform well even on low-resource language pairs. Its architecture is optimized for end-to-end processing of speech inputs, without relying on intermediate

Lang.	Systems	Lang.	Systems
aeb-eng	Seamless Whisper+NLLB (2023)	gle-eng	Seamless Seamless comb.
apc-eng	Seamless Seamless comb. Whisper+NLLB Whisper+NLLB+ID	mar-hin	Seamless Seamless comb. Seamless Shrutilipi Whisper MTL
bem-eng	Seamless Whisper+NLLB Whisper+NLLB+ID Whisper MTL	mlt-eng	Seamless Seamless comb. Whisper MTL
bho-hin	Seamless Seamless comb. Whisper E2E	que-spa	Seamless Whisper+NLLB Whisper+NLLB+ID
est-eng	Seamless	fon-fra	Seamless

Table 2: Systems used for ensembling. For each system we sample five outputs from three model checkpoints and perform MBR on the total sets of sampled outputs (which vary in number from 15 to 60, since the number of systems varies from one to four).

text transcriptions. This design makes it suitable for real-time applications and improves translation accuracy by minimizing error propagation across stages (Barrault et al., 2023).

Similar to Robinson et al. (2024), we employ SEAMLESSM4T v2 for our multilingual translation experiments. We re-fine-tuned SEAMLESSM4T v2 models, training most language pairs for 10 epochs (contrasting the 4 epochs of Robinson et al. (2024)). We found that dev BLEU scores continued to increase with longer train times and hence selected 10 epochs instead of 4. We generally used the same learning rates detailed by Robinson et al. (2024): 1×10^{-6} for almost all language pairs.The following language pairs used the standard learning rate and number of epochs: gle-eng, mlt-eng, bho-hin, mar-hin, apc-eng, fon-fra, and aeb-eng.

As Robinson et al. (2024), for the bem-eng pair, we used a reduced learning rate of 1×10^{-7} while keeping the number of epochs fixed at 10. We also trained the (que-spa) system for 100 epochs due to the small dataset size. Our Estonian ASR system was trained for only 0.5 epochs with a learning rate of 1×10^{-6}, due to the massive dataset size and computational constraints.

A more comprehensive list of all hyperparameters used for these experiments is provided in Appendix A.

Mixed data Training We also experimented with combining data for different language pairs for mixed data training. In this configuration, we explored the effect of combining similar languages for joint training. Specifically, we trained the following groups together:

- bem, mlt, gle, est, aeb, apc → eng
- bho, mar → hin

The rationale behind this setup is that grouping related languages can lead to more robust representations, particularly in low-resource settings, by effectively increasing the size of the training data and enhancing cross-lingual generalization.

For the into-Hindi combined system, we fine-tuned SEAMLESSM4T v2 for 10 epochs in the standard way. For the into-English combined system, trained on a mixture of data that included the full standard fine-tuning sets for mlt-eng, gle-eng, and apc-eng; and we added 5.2% of the Estonian ASR files[1], 36.1% of the bem-eng files, and 38.8% of the Tunisian Arabic-to-English (aeb-eng) files. This was done for the language pairs with the largest datasets to prevent data imbalance, and the percentages were selected to keep the dataset size withing roughly 250 hours total (about 25 times the size of the gle-eng dataset). We fine-tuned SEAMLESSM4T v2 for $\sim$ 1.3 epochs on this combined dataset. Both of these combined-data systems are denoted "Seamless comb." in Table 2.

We also experimented with fine-tuning SEAMLESSM4T v2 on additional data. We attempted to augment the gle-eng dataset (11 hours of audio, as in the 2024 shared task (Ahmad et al., 2024)) with the synthetic data provided by the task organizers for 2025. However, we met this attempt with little success. After more than a full epoch of training, the dev BLEU score had not increased above 1.0 BLEU, so we decided to terminate fine-tuning to preserve our computational resources for other experiments. We did, however, include among our Marathi-to-Hindi (mar-hin) systems a SEAMLESSM4T v2 model that was fine-tuned using the massive Shrutilipi dataset (Bhogale et al., 2022), which contains 1280 hours of Marathi ASR data. In this approach we were inspired by **?**, who employed this dataset among others to develop a successful mod-hin submission in 2023. We used NLLB (NLLB Team et al., 2022) off-the-shelf for Marathi-to-Hindi translation to translate the transcription labels of the dataset to Hindi, and then we employed it as ST data for SEAMLESSM4T v2 fine-tuning (combined with the original mar-hin

[1]This was a mistake, as we mistakenly thought the Estonian ASR data was Estonian-to-English ST data, due to a miscommunication in our team.

data). We trained this model for approximately 2 epochs, and we denote it "Seamless Shrutilipi".

Our apc-eng SEAMLESSM4T v2 model was also trained on synthetic labels. Robinson et al. (2024) did not fine-tune a SEAMLESSM4T v2 model for apc-eng because of the nature of the apc-eng data (only ASR and MT datasets separately, with no ST labels). This year we bypassed this challenge by using Robinson et al.'s (2024) NLLB model fine-tuned for apc-eng MT to translate all the transcriptions in the provided apc ASR dataset into English. Then we fine-tuned SEAMLESSM4T v2 on the resulting dataset.

2.3 Whisper and NLLB

Whisper (Radford et al., 2023b) is a speech recognition model created by OpenAI. It is trained on a large amount of audio data—around 680,000 hours—from the internet. This includes many languages and different types of speech, such as conversations, lectures, and translations. Whisper works well in many languages without needing extra training for each new language. In our work, we use Whisper for both ASR and ST. Whisper's strength is its robustness—it can understand different accents, background noise, and even low-quality recordings. During pretraining, the model was already trained on data from over 90 languages such as English, Marathi, Hindi, Maltese, and Modern Standard Arabic. However, it lacks exposure to several low-resource languages like Bemba, Bhojpuri, and Quechua.

In this work we employed the same Whisper models used by Robinson et al. (2024). Whisper used in tandem with NLLB is used only as an ASR module to convert from the speech domain for text translation. "Whisper E2E" in Table 2 refers to using Whisper as an end-to-end translator via "psuedo-translation" (Robinson et al., 2024). This is the practice of fine-tuning Whisper on bho-hin data with a Hindi ASR objective. "Whisper MTL" also refers to using Whisper as an end-to-end ST system but with a mixed ASR and ST fine-tuning objective. This approach is typically most suitable for into-English ST, since English is the only ST target language officially supported by Whisper.

We also employ NLLB (NLLB Team et al., 2022), an extensive multilingual machine translation system, just as Robinson et al. (2024). NLLB covers more than 200 languages, including Arabic, Quechua, and Bemba. (We use the 600M-parameter release of the model, fine-tuned by Robinson et al. (2024).)

While "NLLB" refers to use of these fine-tuned NLLB model checkpoints in Table 2, "NLLB+ID" refers to the use of NLLB fine-tuned on the same data, but with intra-distillation (Xu et al., 2022; **?**), a regularlization technique designed to ensure that all network parameters contribute equally to successful inference. Intra-distillation was effective in enhancing MT performance in the 2024 shared task campaign.

2.4 System ensembling via MBR

Minimum Bayes Risk Decoding (MBR) (Bickel and Doksum, 1977) is a method of ensembling that aims to choose candidates that have the lowest risk – i.e., high probability but also consistent with other candidates. In other words, if multiple candidates are similar, they are more likely to be correct and it is not too risky to select one of them (Bertsch et al., 2023). It was originally used in machine translation in the early days of phrase-based, statistical methods (Kumar and Byrne, 2004), but has been shown to be very robust to common errors in neural methods (Müller and Sennrich, 2021), explored in-depth theoretically (Ohashi et al., 2024), as well as correlated well with human judgments (Freitag et al., 2022).

See a depiction of traditional MBR in Algorithm 1. $p(c_i)$ is usually set as the posterior probability of the translation candidate when ensembling candidates from a single system; in our case, since we are ensembling different systems, we simply set $p(c_i) = 1$; i.e. we apply a uniform weighting scheme to our candidates. We experiment with using both BLEU (Papineni et al., 2002) and chrF (Popović, 2015) as our similarity metric.

Algorithm 1 Minimum Bayes Risk (MBR) Decoding for Ensembling

Require: Candidate translations $C = \{c_1, c_2, \ldots, c_n\}$
Require: A similarity metric $\text{sim}(\cdot, \cdot)$ (e.g., BLEU, chrF)
Ensure: MBR-selected translation c^*
1: **for all** $c_i \in C$ **do**
2: Compute risk for c_i:
3: $R(c_i) = \sum_{c_j \in C} (1 - \text{sim}(c_i, c_j)) \cdot p(c_j)$
4: **end for**
5: $c^* \leftarrow \arg\min_{c_i \in C} R(c_i)$
6: **return** c^*

3 Results and Conclusions

The results of our different language systems are in Table 3. We used MBR ensembling with a BLEU objective as our primary system for each language pair, and MBR with chrF as our "contrastive 1" submission. In cases where the best performing of the newly fine-tuned SEAMLESSM4T v2 models outperformed all of the systems using Whisper and NLLB from Robinson et al. (2024) (using dev BLEU of the final checkpoint to compare), we selected that system's final checkpoint as our "contrastive 2" submission.

In Table 3, "Test 1" denotes our internal test set, while "Eval" denotes the official evaluation set for the shared task, given by Abdulmumin et al. (2025). We had no internal test set for `apc-eng` since our only ST data for this language pair was synthetically labeled. We also exclude the standard SEAMLESSM4T v2 system for `mar-hin` from our internal test set evaluation since this model was trained on the internal test set.[2]

We see that MBR generally improves performance by a small amount over the best stand-alone system (as can be seen for `bem-eng`, `bho-hin`, `fon-fra`, and `mar-hin`). However, we also see that MBR can also hurt performance (usually slightly), as seen in the remaining language pairs. Disappointingly, we do not see any dramatically higher results on our internal test sets due to MBR, indicating that its benefit in these settings may be smaller than we had originally hypothesized. We point out that ensembling still provides a clear advantage to practitioners, in that they do not need to know which individual system performs best, and can still reach performance on par with whichever the best-performing model is, via this method. However, it does not itself appear to increase scores dramatically.

We also note that while there are significant score differences between different systems (such as SEAMLESSM4T v2 vs. Whisper or cascaded vs. end-to-end), training with combined language data or supplementary data (i.e. Shrutilipi) also did not cause any drastic score increases. Given the scantness of these results, we conclude that methods such as ensembling and multilingual training either have limited use in some low-resource speech translation settings, or that they require more creative and effective applications than those we explored here in order to be optimally useful. We encourage researchers to explore such creative applications of these techniques, as well as other techniques to improve low-resource systems, in the future.

[2]This was another mistake due to a file path misunderstanding.

Acknowledgments

We thank Neha Verma, Henry Li, Philipp Koehn, Yaohan Guan, Sanjeev Khudanpur, and Amir Hussein for their contributions to this work.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 30 others. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austia (in-person and online). Association for Computational Linguistics. To appear.

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, and 43 others. 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.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, and 25 others. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

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, and 24 others. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on*

Language Pair	System	Submission	Test1 BLEU	Eval BLEU
aeb-eng	JHU-cascade-2023		-	-
	Seamless	contr. 2	11.47	6.70
	MBR-BLEU	primary	10.73	8.20
	MBR-CHRF	contr. 1	10.76	8.90
apc-eng	Seamless		-	-
	Seamless comb.		-	-
	Whisper ASR + NLLB		-	-
	Whisper ASR + NLLB + ID		-	-
	MBR-BLEU	primary	-	14.64
	MBR-CHRF	contr. 1	-	15.39
bem-eng	Seamless		14.67	-
	Whisper MTL		17.76	-
	Whisper ASR + NLLB		27.58	-
	Whisper ASR + NLLB + ID		29.39	-
	MBR-BLEU	primary	28.80	26.8
	MBR-CHRF	contr. 1	27.84	28.1
bho-hin	Seamless	contr. 2	37.39	7.8
	Seamless comb.		39.08	-
	Whisper MTL		24.19	-
	MBR-BLEU	primary	39.38	10.5
	MBR-CHRF	contr. 1	39.39	10.7
fon-fra	Seamless	contr. 2	5.34	5.60
	MBR-BLEU	primary	4.76	5.96
	MBR-CHRF	contr. 1	5.57	6.26
gle-eng	Seamless	contr. 2	51.80	12.3
	Seamless comb.		47.65	-
	MBR-BLEU	primary	50.37	11.6
	MBR-CHRF	contr. 1	51.13	12.0
mar-hin	Seamless		-	-
	Seamless comb.		44.98	-
	Seamless Shrutilipi	contr. 2	43.17	40.0
	Whisper MTL		28.06	-
	MBR-BLEU	primary	45.64	41.4
	MBR-CHRF	contr. 1	45.27	40.7
mlt-eng	Seamless	contr. 2	40.62	56.10
	Seamless comb.		38.57	-
	Whisper MTL		21.37	-
	MBR-BLEU	primary	40.02	56.80
	MBR-CHRF	contr. 1	38.98	55.98
que-spa	Seamless		1.05	-
	Whisper ASR + NLLB		6.08	-
	Whisper ASR + NLLB + ID		10.69	-
	MBR-BLEU	primary	7.87	9.0
	MBR-CHRF	contr. 1	10.29	11.0

Table 3: BLEU score results. "Test BLEU" refers to our internal tests, while "Eval BLEU" refers to the evaluation run by Abdulmumin et al. (2025)

Spoken Language Translation (IWSLT 2022), pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Marta Bañón, Malina Chichirau, Miquel Esplà-Gomis, Mikel L. Forcada, Aarón Galiano-Jiménez, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, and Jaume Zaragoza-Bernabeu. 2023. Maltese-english parallel corpus MaCoCu-mt-en 2.0. Slovenian language resource repository CLARIN.SI.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Waad Ben Kheder, Josef Jon, André Beyer, Abdel Messaoudi, Rabea Affan, Claude Barras, Maxim Tychonov, and Jean-Luc Gauvain. 2024. ALADAN at IWSLT24 low-resource Arabic dialectal speech translation task. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 192–202, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew R Gormley. 2023. It's mbr all the way down: Modern generation techniques through the lens of minimum bayes risk. In *The Big Picture Workshop*, page 108.

Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages. *arXiv preprint*.

Peter J Bickel and Kjell A Doksum. 1977. *Mathematical statistics: basic ideas and selected topics*. Holden-Day Inc.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP 2*, page 21.

John E. Ortega, Rodolfo Joel Zevallos, Ibrahim Said Ahmad, and William Chen. 2024. QUESPA submission for the IWSLT 2024 dialectal and low-resource speech translation task. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 125–133, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source multi-speaker speech corpora for building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu speech synthesis systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6494–6503, Marseille, France. European Language Resources Association.

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.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

D. Fortuné Kponou, Fréjus A. A. Laleye, and Eugène Cokou Ezin. 2024. FFSTC: Fongbe to French speech translation corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7270–7276, Torino, Italia. ELRA and ICCL.

Shankar Kumar and Bill Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176.

Zhaolin Li, Enes Yavuz Ugan, Danni Liu, Carlos Mullov, Tu Anh Dinh, Sai Koneru, Alexander Waibel, and Jan Niehues. 2024. The KIT speech translation systems for IWSLT 2024 dialectal and low-resource track. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 221–228, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

John Makhoul, Bushra Zawaydeh, Frederick Choi, and David Stallard. 2005. Bbn/aub darpa babylon levantine arabic speech and transcripts. *Linguistic Data Consortium (LDC), LDC Catalog No.: LDC2005S08*.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. 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 259–272, Online. 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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Atsumoto Ohashi, Ukyo Honda, Tetsuro Morimura, and Yuu Jinnai. 2024. On the true distribution approximation of minimum bayes-risk decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 459–468.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023a. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023b. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nathaniel Romney Robinson, Kaiser Sun, Cihan Xiao, Niyati Bafna, Weiting Tan, Haoran Xu, Henry Li Xinyuan, Ankur Kejriwal, Sanjeev Khudanpur, Kenton Murray, and 1 others. 2024. Jhu iwslt 2024 dialectal and low-resource system description. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 140–153.

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.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Claytone Sikasote, Eunice Mukonde, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. BIG-C: a multimodal multi-purpose dataset for Bemba. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2062–2078, Toronto, Canada. Association for Computational Linguistics.

Tiia Sildam, Andra Velve, and Tanel Alumäe. 2024. Finetuning end-to-end models for Estonian conversational spoken language translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 166–174, Bangkok, Thailand.

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).

Haoran Xu, Philipp Koehn, and Kenton Murray. 2022. The importance of being parameters: An intra-distillation method for serious gains. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 170–183, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haoran Xu, Jean Maillard, and Vedanuj Goswami. 2023. Language-aware multilingual machine translation with self-supervised learning. *arXiv preprint arXiv:2302.05008*.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A SEAMLESSM4T v2 hyperparameters

In our experiments, we use the SEAMLESSM4T v2 model. To keep things consistent, we cut off audio that is longer than 30 seconds. We also use a fixed random seed of 42 so that the results can be repeated.

We try different learning rates from the set $\{10^{-6}, 10^{-7}\}$ to see which works best. For most language pairs, we fine-tune the `SEAMLESSM4T v2-large` model for 10 epochs using a learning rate of 1×10^{-6} and a batch size of 32.

During training, we use a constant learning rate schedule and set 50 warm-up steps. When generating translations, we use greedy decoding and limit the output to 256 tokens.