JHU IWSLT 2023 Multilingual Speech Translation System Description

Henry Li Xinyuan[1*] **Neha Verma**[1*] **Bismarck Bamfo Odoom**[1] **Ujvala Pradeep**[1]
Matthew Wiesner[2] **Sanjeev Khudanpur**[1,2]
[1]Center for Language and Speech Processing, and
[2] Human Language Technology Center of Excellence,
Johns Hopkins University
{xli257, nverma7, bodoom1, upradee1, wiesner, khudanpur}@jhu.edu

Abstract

We describe the Johns Hopkins ACL 60-60 Speech Translation systems submitted to the IWSLT 2023 Multilingual track, where we were tasked to translate ACL presentations from English into 10 languages. We developed cascaded speech translation systems for both the constrained and unconstrained subtracks. Our systems make use of pre-trained models as well as domain-specific corpora for this highly technical evaluation-only task. We find that the specific technical domain which ACL presentations fall into presents a unique challenge for both ASR and MT, and we present an error analysis and an ACL-specific corpus we produced to enable further work in this area.

1 Introduction

In this work, we describe the 2023 JHU 60-60 Multilingual speech translation track submissions and their development (Agarwal et al., 2023; Salesky et al., 2023). This multilingual task involved the translation of ACL conference oral presentations, given in English, into 10 different target languages. High quality translation systems that can assist in translating highly technical and scientific information helps in the dissemination of knowledge to more people, which in turn can help make our field more inclusive and accessible.

We briefly describe the task in Section 2. In Section 3 we describe the collection and preparation of in-domain ACL data to improve ASR and MT performance by addressing the domain-specificity of the task. We then describe our systems in Section 4, including their motivation and design in context of this shared task. Technical details of our experiments are in 5. We present our results and a discussion of our contributions in Section 6.

* Authors contributed equally

2 The Speech Translation of Talks Task

In 2022, the ACL began the 60-60 initiative, a diversity and inclusion initiative to translate the ACL Anthology into 60 languages for its 60th anniversary. The initiative provided evaluation data for the IWSLT 2023 *multilingual track* on *speech translation of talks* from English into 10 major languages.

It was further split into *constrained* and *unconstrained* subtracks. The constrained subtrack allowed the use of only certain datasets and pre-trained models, whereas the unconstrained subtrack had no such restrictions. We submitted systems to both subtracks and describe them in Section 4.

2.1 Evaluation Data

The ACL 60-60 development data provided to participants is composed of the audio of 5 talks, their transcripts, and multi-parallel translations into 10 languages. Each talk is about 12 minutes in length – a total of about an hour of English speech for the entire set. Additionally, participants are provided with the text abstract of each talk taken from the corresponding paper.

The nature of these data presents a few major challenges for speech translation. The ACL is a global community of researchers from many different countries who speak in a variety of accents, which can pose a challenge to even modern day speech recognition systems. Additionally, the content of these talks is highly technical and contains terms and acronyms that are specific to the field. Sentence-level translations of the talks are provided along with unsegmented audio of the full $\sim$12 minute talk. An audio segmentation produced with the SHAS baseline segmentation method (Tsiamas et al., 2022) is also provided.

3 In-domain Data

Utilizing additional in-domain data has been shown to be helpful in improving the performance and

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 302–310
July 13-14, 2023

robustness of translation systems. In light of this, we scraped talks and papers from the proceedings and workshops of ACL 2021.

3.1 Data Collection

About 65% of the papers accepted in ACL 2021 have video presentations recorded and uploaded on the ACL website. We scraped 1847 papers and 1193 talks from the proceedings and workshops. The format of the papers and talks are pdf and mp4 respectively. We extract the text from the papers using `pypdf`.[1] The talks are split into 30-second chunks, converted into FLAC format, and resampled to 16KHz. This amounts to about 155 hours of speech and about 200K lines of text. We plan to release the data under a CC BY 4.0 license[2] (same as the license for the ACL talks).

3.2 Data Filtering

To make the corpora (including ACL papers before 2022) useful, we first denoised the data and made it similar to ASR text outputs. A comprehensive list of the filters we applied to the data includes:

- Removing any information past the References section.
- Removing links ("https..").
- Reforming broken words since the text was in a two column format.
- Removing any information before the Abstract section.
- Removing any non alpha-numeric or punctuation characters.
- Removing any lines that start with or that have too many numbers (to account for tables with data).
- Removing any lines with less that 10 characters (number obtained from averaging minimum character length of each sentence in dev data).
- Removing any lines larger than 297 characters (number obtained through a similar process as above).
- Reformatting the data such that it has one sentence per line.

[1] https://github.com/py-pdf/pypdf

[2] https://github.com/IWSLT-23/60_60_data/tree/main/acl_data

These constraints were applied in order to mimic the text-normalization of the dev data so that these scraped ACL data could be incorporated into our model's source language side.

4 Systems

In this section, we separately describe our unconstrained and constrained submissions. Since we built cascaded models, we describe the automatic speech recognition (ASR) and machine translation (MT) components of each system.

4.1 Unconstrained Subtrack

4.1.1 Automatic Speech Recognition

An important characteristic of ACL presentations is the wide array of accents represented, which reflects the diverse background of NLP researchers. Accent-robust speech recognition continues to present a challenge to the community (Tadimeti et al., 2022; Riviere et al., 2021; Radford et al., 2022).

One model that demonstrated a degree of robustness to accented speech, is Whisper (Radford et al., 2022), an ASR model trained on 680,000 hours of web-crawled data. Its performance on the accented splits of the VoxPopuli (Wang et al., 2021), while significantly worse than non-accented English, was comparable (without an external language model) to methods designed for accent robustness (with a strong language model) (Riviere et al., 2021). This robustness to accented speech, as well as its overall strong performance on English ASR makes it well-suited for the accent-diverse ACL presentations.

The domain specificity and technical terms of ACL presentations may still prove difficult for a strong ASR model like Whisper. We therefore condition the decoder towards key technical vocabulary and named entities by prompting Whisper with the corresponding abstracts when decoding each presentation.

Additionally, we test the effect of using the pre-segmented audio files (with oracle segmentation provided by the IWSLT 60-60 challenge organizers) versus using longer speech segments for Whisper decoding. We find that decoding the full talk at once results in a lower WER than decoding segment-by-segment. For Whisper-large, the best performing model, this difference is 0.6 WER. Longer form inputs more closely match the training segments of Whisper, which were in 30 second segments (Radford et al., 2022).

4.1.2 Audio Segmentation

Since we found that decoding using unsegmented audio outperformed decoding using the predefined segments, we segment our ASR text output in order to perform sentence-level machine translation. We choose to perform sentence-level machine translation rather than incorporating more document context because our final systems make use of many large pre-trained multilingual models that are trained at a sentence level rather than a document level.

Because we require sentence-level segments from our ASR outputs, we use the state-of-the-art `ersatz` neural sentence segmenter. `ersatz` has been shown to be more robust to technical terms including acronyms and irregular punctuation, which is particularly helpful in the ACL domain (Wicks and Post, 2021).

4.1.3 Machine Translation

We test several pre-trained MT systems on our data. Specifically, we test NLLB-200 (NLLB Team et al., 2022), mBART50 (Tang et al., 2020), and M2M100 (Fan et al., 2021). All 10 of our target languages are supported by these models.

The original NLLB-200 model is a 54 billion parameter Mixture-of-Experts model that translates to and from 200 languages. It is trained on a large amount of mined parallel, back-translated, and monolingual data. We use the 3.3B parameter version of NLLB-200, which is a dense Transformer model that is trained via online distillation of the original model, but still supports all of the original 200 languages.

mBART50 is the second iteration of the multilingual BART model, which is a dense transformer architecture trained on multilingual text using a denoising task. The authors of mBART50 also release a checkpoint of mBART50 that is fine-tuned on the one-to-many translation task, which we will refer to as mBART50-1toN. In this case, English is the source, and all 50 covered languages are the targets.

Finally, M2M100 is another transformer-based model that is trained directly on the MT task. It translates to and from 100 languages, and is a previous iteration of the initiative that produced NLLB-200. However, we still test both models because sometimes adding additional language pairs to a model can lead to the reduced performance of some language pairs (Aharoni et al., 2019; Arivazhagan et al., 2019). We use the 1.2B parameter version of M2M100 in our experiments.

4.1.4 Domain-Specific Data

Using the 2021 ACL data described in Section 3, we attempted to perform sequence knowledge distillation (SeqKD) (Kim and Rush, 2016). Because we only had additional source-side monolingual data, SeqKD could give us pseudo-target labels in order to retrain our best model on these outputs.

Although NLLB-200-3.3B is our best model for many of our language pairs, we fine-tune NLLB-200-1.3B instead due to computational constraints. While benchmarking these models, however, there is only a marginal improvement in using the larger model over the smaller (average +0.6 chrF). For en-ja, however, we continue to use mBART50-1toN.

Despite the large amount of in-domain source language data we made available, we did not see much benefit from it ourselves, specifically for data augmentation via SeqKD. We speculate that the data may be too noisy in spite of filtering, and that its best use may be as *source context* during inference, rather than for *training data augmentation*.

4.2 Constrained Subtrack

4.2.1 Automatic Speech Recognition

We leveraged the pre-trained wav2vec 2.0 model (Baevski et al., 2020) for the constrained ST task. Wav2vec 2.0 was trained in a self-supervised fashion and requires fine-tuning on an annotated corpus in order to be used for the ASR task, with the domain-similarity between the choice of the fine-tuning corpus and the evaluation data being crucial for ASR performance. The most commonly used wav2vec 2.0 model is fine-tuned with a CTC objective on Librispeech, a corpus made of audiobooks that is considered to have a considerable domain mismatch compared to the ACL 60-60 data. Since the development split of the ACL 60-60 data alone is insufficient for wav2vec 2.0 fine-tuning, we instead performed a two-stage fine tuning with TED-LIUM 3 (Hernandez et al., 2018) being used in the first stage and the ACL 60-60 development data used in the second.

Our approach to tackling the content domain mismatch between the training data and ACL presentations is to perform ASR decoding with the help of an content-domain matching language model. What it means in practice is that we rescore the per-frame output trellis with a content-domain matching language model, which in turn was created by interpolating a general language model (trained from all the available English corpora in the constrained challenge) and a domain-specific language model (trained with transcripts from the ACL 60-60 development data). In order to bias our model towards named entities mentioned in each specific presentation, we train a separate language model for each presentation by re-interpolating the above-mentioned language model with one trained with the corresponding paper abstract.

4.2.2 Machine Translation

In the constrained setting, we use mBART50-1toN and M2M100 as our base models. We additionally test fine-tuning these models on MuST-C data, which we hypothesized to be closely related to the ACL talk data, domain-wise (Di Gangi et al., 2019). This data is comprised of professionally translated English TED talks, which matches the presentation domain as well as some of the technical nature of the ACL talks, although to a lesser degree.

We fine-tune both mBART and M2M100 using the MuST-C transcripts and translations available in all 10 language pairs. We use data from both v1.2 (v1.0 is contained in v1.2) and v2.0 depending on language pair availability. A summary of this data is provided in Table 1. For mBART, we additionally test multilingual fine-tuning where we fine-tune on all the language pairs simultaneously, rather than fine-tuning on a single language pair bitext (Tang et al., 2020).

lang. pair	MuST-C release	# lines
en-ar	v1.2	212085
en-de	v1.0	229703
en-fa	v1.2	181772
en-fr	v1.0	275085
en-ja	v2.0	328639
en-nl	v1.0	248328
en-pt	v1.0	206155
en-ru	v1.0	265477
en-tr	v1.2	236338
en-zh	v1.2	184801

Table 1: Dataset statistics and source of MuST-C bitext across the 10 task language pairs.

5 Experimental Setup

In this section, we provide technical details of our experiments and our evaluation practices.

5.1 ASR Experiments

5.1.1 Prompting Whisper

In the unconstrained setting, we evaluate Whisper on both the segmented and unsegmented audio files. We simulate LM biasing by using the "prompt" interface provided by Whisper.

5.1.2 Decoding with an Interpolated Language Model

In the constrained setting, we build a domain-adapted language model as follows: first we combine transcripts from a number of ASR corpora that are available in the constrained challenge, namely Librispeech, VoxPopuli, Common Voice (Ardila et al., 2020), and TED-LIUM 3, to train a flexible 6-gram general bpe-level language model for English. We proceed to interpolate the general English language model with one trained on the development split transcripts from the ACL 60-60 challenge, allowing the model to gain exposure to technical terms within the NLP field. Finally, during decoding, we further interpolate the previously obtained language model with a low-order language model trained from the paper abstract corresponding to the current presentation, biasing our model towards technical terms and named entities that are likely to appear in the presentation.

We used KenLM (Heafield, 2011) to train and integrate our language models. The interpolation weights for each step were estimated using a leave-one-out strategy on the development split, minimising the perplexity on the held-out transcript and averaging the interpolation weights.

5.1.3 Decoding with a Language Model Trained on Additional ACL Anthology data

We use the text scraped from the proceedings and workshops of ACL 2021 to train a 6-gram domain-matching language model for decoding. Without interpolation or additional data, this gives a WER of 18.9 and a technical term recall of 0.47 using Wav2Vec2-TED-LIUM 3 as the acoustic model. We observe that using data from a similar domain improves performance even though the data are relatively noisy.

5.1.4 Evaluation

We compare ASR performance, as measured by Word Error Rate (WER), across the different systems that we built. Specifically, we compute WER on depunctuated lowercase transcripts. Since we

Acoustic Model	Language Model	WER	Tech. Term Recall
Whisper-medium.en	-	8.1	0.861
Whisper-medium.en	abstract prompting	8.7	0.865
Whisper-large	-	6.8	0.854
Whisper-large	abstract prompting	6.9	0.852
Whisper-large	abstract and conclusion prompting	6.7	0.863
Whisper-large	abstract, conclusion and intro prompting	6.6	0.851
Whisper-large	abstract, conclusion, intro & author name prompting	6.4	0.854
Wav2Vec2-960h librispeech	librispeech-4gram	25.1	0.306
Wav2Vec2-960h librispeech	interpolated LM	24.3	0.370
Wav2Vec2-960h librispeech	inter. LM + dev transcripts	24.1	0.382
Wav2Vec2-960h librispeech	inter. LM + dev + abstract	23.7	0.392
Wav2Vec2-960h librispeech	inter. LM + dev + abstract + ACL anthology	20.7	0.462
HUBERT-960h librispeech	librispeech-4gram	22.0	0.390
HUBERT-960h librispeech	interpolated LM	21.7	0.386
HUBERT-960h librispeech	inter. LM + dev transcripts	20.4	0.421
HUBERT-960h librispeech	inter. LM + dev + abstract	20.4	0.498
HUBERT-960h librispeech	inter. LM + dev + abstract + ACL anthology	18.5	0.473
Wav2Vec2-TED-LIUM 3	librispeech-4gram	20.9	0.383
Wav2Vec2-TED-LIUM 3	interpolated LM	19.5	0.422
Wav2Vec2-TED-LIUM 3	inter. LM + dev transcripts	18.9	0.436
Wav2Vec2-TED-LIUM 3	inter. LM + dev + abstract	14.2	0.626
Wav2Vec2-TED-LIUM 3	inter. LM + dev + abstract + ACL anthology	16.7	0.505
Wav2Vec2-TED-LIUM 3	ACL anthology only	18.9	0.470

Table 2: ASR results. WER is measured against depunctuated, all lower-case reference text.

either perform ASR on unsegmented talks (unconstrainted), or on the SHAS-segmented audio (constrained), we use `mwerSegmenter` to align our outputs to the gold transcripts (Matusov et al., 2005).

Because we are interested in the effect of using domain-specific text to improve ASR on technical terms, we compute the recall of NLP-specific technical words in our output. We obtain these technical terms by asking domain experts to flag all technical terms in the development set reference transcript.

5.2 MT Experiments

5.2.1 MuST-C fine-tuning

For bilingual fine-tuning on mBART50 and M2M100, we train for 40K updates, and use loss to select the best checkpoint. For multilingual fine-tuning on mBART50-1toN, we train for 100K updates, and use temperature sampling of the mixed datset using $T = 1.5$. We use loss to select the best checkpoint. For all experiments, we use an effective batch size of 2048 tokens.

5.2.2 Evaluation

For all experiments, we report BLEU and chrF scores as reported by `sacrebleu` (Post, 2018). For Japanese and Chinese, we use the appropriate tokenizers provided by `sacrebleu` (`ja-mecab` and `zh`, respectively).

For evaluating translations of ASR outputs, either segmented using `ersatz` or pre-segmented using the provided SHAS-segmented wav files, we use the `mwerSegmenter` to resegment the translations based on the references. For all languages except Japanese and Chinese, we use detokenized text as input to resegmentation. However, for Japanese and Chinese, we first use whitespace tokenization as input to `mwerSegmenter`, and then detokenize for scoring, which is retokenized according to the `sacrebleu` package.

6 Results

6.1 ASR Results

For the Whisper-based systems, we focus on the effects of prompting; for the constrained systems, we contrast different families of pre-trained ASR models fine-tuned on different ASR corpora; finally, we assess the efficacy of incorporating an in-domain language model during decoding. The full list of results is shown in Table 2.

Contrary to what we expected, prompting Whisper with the corresponding paper abstracts not only had little impact on the ASR WER, but also failed

	mBART50-1toN		M2M100		NLLB-200	
language pair	BLEU	chrF	BLEU	chrF	BLEU	chrF
en-ar	22.6	52.9	16.2	46.3	37.6	**65.4**
en-de	37.4	66.0	39.7	66.8	42.9	**69.6**
en-fa	17.2	49.6	20.4	49.5	27.4	**57.3**
en-fr	46.4	70.4	54.5	74.6	55.9	**76.2**
en-ja	37.5	**45.9**	35.2	43.8	25.7	36.3
en-nl	41.0	69.0	50.9	75.3	51.5	**76.1**
en-pt	44.3	69.7	57.6	77.4	61.6	**79.0**
en-ru	22.2	52.0	24.3	54.3	27.4	**57.2**
en-tr	15.5	50.7	22.3	56.5	28.6	**62.8**
en-zh	43.8	38.8	45.7	**40.7**	42.2	38.5

Table 3: Unconstrained MT results on the development set using oracle transcripts as input. Both chrF and BLEU scores are computed using the mWER Segmenter and `sacrebleu`. BLEU scores for ja and zh are computed using the ja-mecab and zh tokenizers in `sacrebleu`, respectively. We bold our best chrF scores as it is the main metric of the task.

	mBART50-1toN		+MuST-C (indiv)		+MuST-C (multi)		M2M-100		+MuST-C (indiv)	
lang pair	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
en-ar	22.6	52.9	24.7	**55.9**	19.6	51.0	16.2	46.3	24.0	55.7
en-de	37.4	66.0	35.6	63.7	36.8	64.5	39.7	**66.8**	34.7	62.8
en-fa	17.2	49.6	28.9	**56.0**	26.3	52.4	20.4	49.5	17.9	54.4
en-fr	46.4	70.4	48.0	70.9	46.7	70.1	54.5	**74.6**	49.0	71.1
en-ja	37.5	**45.9**	24.0	35.7	24.9	37.0	35.2	43.8	21.0	32.3
en-nl	41.0	69.0	43.3	70.1	38.5	67.1	50.9	**75.3**	42.1	69.0
en-pt	44.3	69.7	48.2	71.4	42.8	68.5	57.6	**77.4**	50.0	72.3
en-ru	22.2	52.0	21.0	50.4	19.5	47.9	24.3	**54.3**	22.1	50.7
en-tr	15.5	50.7	18.9	53.3	15.6	50.8	22.3	**56.5**	21.4	56.0
en-zh	43.8	38.8	45.3	40.6	31.5	39.2	45.7	**40.7**	42.8	37.5

Table 4: Constrained MT results on the development set using oracle transcripts as input. Both chrF and BLEU scores are computed using the mWER Segmenter and `sacrebleu`. BLEU scores for ja and zh are computed using the ja-mecab and zh tokenizers in `sacrebleu`, respectively. We bold our best chrF scores as it is the main metric of the task.

to improve the recall of technical terms of the ASR system. Further increasing the length and relevance of the prompts provided to whisper, such as adding the conclusion and part of the introduction section of each paper corresponding to the ACL presentation in question, had marginal impact on both of the above-mentioned metrics. A more detailed look at the mechanism and behaviour of Whisper prompting could help to understand this observation.

On the constrained side, the incorporation of the interpolated LM during ASR decoding had a significant impact on the performance of our ASR systems, regardless of the upstream acoustic model. As expected, increasing the quality of the out-of-domain language model (from Librispeech-4gram to Interpolated LM) resulted in WER improvements while not necessarily helping technical term recall; by contrast, while LMs that better fit the domain may not necessarily help WER, they bring substantial gains in technical term recall.

The language model that best fits our domain, namely the model that interpolates the LMs trained from every ASR corpus in addition to the development transcripts, from the current paper abstract, and from the crawled ACL anthology, provided substantial improvement on both WER and technical term recall for the weaker acoustic models (Wav2Vec2 fine-tuned on Librispeech) but not on

	Constrained			Unconstrained		
language	MT system	BLEU	chrF	MT system	BLEU	chrF
en-ar	mBART50-1toN+MuST-C	15.3	45.6	NLLB-200-3.3B	33.7	62.5
en-de	M2M100	24.3	55.2	NLLB-200-3.3B	39.6	67.8
en-fa	mBART50-1toN+MuST-C	14.8	42.0	NLLB-200-3.3B	24.5	54.3
en-fr	M2M100	33.3	61.9	NLLB-200-3.3B	49.3	72.5
en-ja	mBART50-1toN	21.9	29.9	mBART50-1toN	34.8	43.1
en-nl	M2M100	30.6	62.5	NLLB-200-3.3B	45.7	72.4
en-pt	M2M100	34.9	63.4	NLLB-200-3.3B	54.7	75.6
en-ru	M2M100	15.0	45.1	NLLB-200-3.3B	24.8	54.4
en-tr	M2M100	11.9	43.5	NLLB-200-3.3B	24.7	58.8
en-zh	M2M100	32.2	26.6	M2M100	37.7	33.5

Table 5: Final speech translation results for both our constrained and unconstrained systems on the development set. Both chrF and BLEU scores are computed using the mWER Segmenter and `sacrebleu`. BLEU scores for ja and zh are computed using the ja-mecab and zh tokenizers in `sacrebleu`, respectively. We used output from our strongest ASR system, Whisper-large with abstract prompting, as the input to our translation system.

the stronger acoustic models.

6.2 MT results

We detail the results of testing pre-trained MT models as described in Section 4 on the oracle transcripts in Table 3. This table reflects experiments we performed for the unconstrained setting. We find that for almost all language pairs, NLLB-200-3.3B has the best performance, except for en-ja and en-zh, which perform best with mBART and M2M100, respectively.

We summarize our fine-tuning results in Table 4. This table reflects experiments we performed for the constrained setting. We find that in general, the additional data can provide a boost over mBART50-1toN, but not for M2M100. Additionally, we find that despite positive results in Tang et al. (2020), multilingual fine-tuning does not outperform bilingual fine-tuning in this setting. For a majority of pairs, M2M100 without fine-tuning is the best system, but for en-ar and en-fa, mBART50-1toN with fine-tuning is the best system, and similar to the unconstrained system, mBART50-1toN without fine-tuning is the best system for en-ja.

6.3 ST Results

Final results for both our constrained and unconstrained systems are summarized in Table 5. We translate the transcripts from our best ASR systems using the best language-pair specific MT systems. In the unconstrained case, the average reduction in chrF from using ASR outputs versus oracle transcripts is -5.7 chrF. In the constrained case, this value is -12.8 chrF. The small reduction in the unconstrained system indicates that our cascaded approach of two strong components is a viable option for ST in this setting. However, our constrained system could likely benefit from techniques that help reduce the error propagation from ASR, like mixing ASR outputs with gold source sentences during MT training, or joint training of ASR and MT components.

7 Conclusion

We present a constrained and unconstrained system for the IWSLT 2023 Multilingual speech translation task. We address some of the major challenges of this dataset with our design choices: ASR robust to speaker accents, adaptation to match the domain specificity, and ASR prompting to incorporate context in this academic talk-level translation task. We additionally release a supplemental ACL audio and text corpus to encourage further work in high quality speech translation of ACL content.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, 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 Ojha Kr., 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)*. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. 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 3874–3884, Minneapolis, Minnesota. 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.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

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.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. 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 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

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.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. *CoRR*, abs/1805.04699.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. 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. *arXiv preprint arXiv:2212.04356*.

Morgane Riviere, Jade Copet, and Gabriel Synnaeve. 2021. Asr4real: An extended benchmark for speech models. *arXiv preprint arXiv:2110.08583*.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating Multilingual Speech Translation Under Realistic Conditions with Resegmentation and Terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Divya Tadimeti, Kallirroi Georgila, and David Traum. 2022. Evaluation of off-the-shelf speech recognizers on different accents in a dialogue domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6001–6008, Marseille, France. European Language Resources Association.

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. *arXiv preprint arXiv:2008.00401*.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. In *Proc. Interspeech 2022*, pages 106–110.

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.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. 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 3995–4007, Online. Association for Computational Linguistics.