MULTILINGUAL END-TO-END SPEECH TRANSLATION

Hirofumi Inaguma[1], Kevin Duh[2], Tatsuya Kawahara[1], Shinji Watanabe[2]

[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

In this paper, we propose a simple yet effective framework for multilingual end-to-end speech translation (ST), in which speech utterances in source languages are directly translated to the desired target languages with a universal sequence-to-sequence architecture. While multilingual models have shown to be useful for automatic speech recognition (ASR) and machine translation (MT), this is the first time they are applied to the end-to-end ST problem. We show the effectiveness of multilingual end-to-end ST in two scenarios: *one-to-many* and *many-to-many* translations with publicly available data. We experimentally confirm that multilingual end-to-end ST models significantly outperform bilingual ones in both scenarios. The generalization of multilingual training is also evaluated in a transfer learning scenario to a very low-resource language pair. All of our codes and the database are publicly available to encourage further research in this emergent multilingual ST topic[1].

Index Terms— Speech translation, multilingual end-to-end speech translation, attention-based sequence-to-sequence, transfer learning

1. INTRODUCTION

Breaking the language barrier for communication is one of the most attractive goals. For several decades, the speech translation (ST) task has been designed by processing speech with automatic speech recognition (ASR), text normalization (e.g. punctuation restoration, case normalization etc.), and machine translation (MT) components in a cascading manner [1, 2]. Recently, end-to-end speech translation (E2E-ST) with a sequence-to-sequence model has attracted attention for its extremely simplified architecture without complicated pipeline systems [3, 4, 5]. By directly translating speech signals in a source language to text in a target language, the model is able to avoid error propagation from the ASR module, and also leverages acoustic clues in the source language, which have shown to be useful for translation [6]. Moreover, it is more memory- and computationally efficient since complicated decoding for the ASR module and the latency occurring between ASR and MT modules can be bypassed.

Although end-to-end optimization demonstrates competitive results compared to traditional pipeline systems [5, 7] and even outperforms them in some corpora [4, 8], these models are usually trained with a single language pair only (i.e. bilingual translation). There is a realistic scenario in the applications of ST models when a speech utterance is translated to multiple target languages in a lecture, news reading, and conversation domains. For example, TED talks are mostly conducted in English and translated to more than
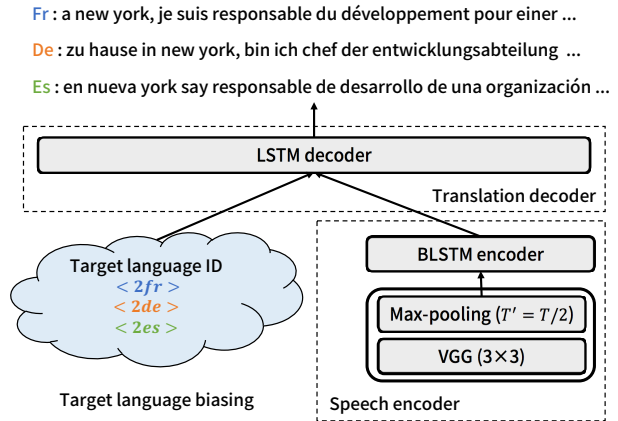

Fig. 1: System overview for the multilingual end-to-end speech translation model

70 languages in the official website [9]. In these cases, it is a natural choice to support translation of multiple language pairs from speech.

A practical approach for multilingual ST is to construct (mono- or multi-lingual) ASR and (bi- or multi-lingual) MT systems separately and combine them as in the conventional pipeline system [10]. Thanks to recent advances in sequence-to-sequence modeling, we can build strong multilingual ASR [11, 12, 13, 14, 15], and MT systems [16, 17, 18] even with a single model. However, when speech utterances come from multiple languages, mis-identification of the the source language by the ASR system disables the subsequent MT system from translating properly since it is trained to consume text in the correct source language[2]. In addition, text normalization, especially punctuation restoration, must be conducted for ASR outputs in each source language, from which additional errors could be propagated.

In this paper, we propose a simple and effective approach to perform multilingual E2E-ST by leveraging a universal sequence-to-sequence model (see Figure 1). Our framework is inspired by [16], where all parameters are shared among all language pairs, which also enables zero-shot translations. By building the multilingual E2E-ST system with a universal architecture, it is free from the source language identification and the complexities of training and decoding pipelines are drastically reduced. Furthermore, we do not have to care about which parameters to share among multiple language pairs, which can be learned automatically from training data. To the best of our knowledge, this is the first attempt to investigate multilingual training for the E2E-ST task.

---

[1]Available at `https://github.com/espnet/espnet`.

[2]In case of *one-to-many* situation, this does not occur since only the monolingual ASR is required. However, error propagation from the ASR module and latency between the ASR and MT modules is still problematic.

We conduct experimental evaluations with three publicly available corpora: Fisher-CallHome Spanish (Es→En) [19], Librispeech (En→Fr) [20], and Speech-Translation TED corpus (En→De) [21]. We evaluate *one-to-many* (O2M) and *many-to-many* (M2M) translations by combining these corpora and confirm significant improvements by multilingual training in both scenarios. Next, we evaluate the generalization of multilingual E2E-ST models by performing transfer learning to a very low-resource ST task: Mboshi (Bantu C25)→Fr corpus (4.4 hours) [22]. We show that multilingual pre-training of the seed E2E-ST models improves the performance in the low-resource language pair unseen during training, compared to bilingual pre-training. Our codes are put to the public project so that results can be reproducible and strictly compared in the same pre-processing (e.g., data split, text normalization, and feature extraction etc.), model implementation, and evaluation pipelines.

2. BACKGROUND: SPEECH TRANSLATION

In this section, we describe the architecture of the pipeline and end-to-end speech translation (ST) system. Our ASR, MT, and ST systems are all based on attention-based RNN encoder-decoder models[3] [24, 25]. Let $\boldsymbol{x}^{\mathrm{src}}$ be the input speech features in a source language, $\boldsymbol{y}^{\mathrm{src}}$ and $\boldsymbol{y}^{\mathrm{tgt}}$ be the corresponding reference transcription and translation, respectively. In this work, we adopt a character-level unit both for source and target references[4].

2.1. Pipeline speech translation

The pipeline ST model is composed of three modules: automatic speech recognition (ASR), text normalization, and neural machine translation (NMT) models [2].

2.1.1. Automatic speech recognition (ASR)

We build the ASR module based on hybrid CTC/attention framework [27, 28], where the attention-based encoder-decoder is enforced to learn monotonic alignments by jointly optimizing with Connectionist Temporal Classification (CTC) objective function [29]. Our ASR model consists of three modules: the speech encoder, transcription decoder, and the softmax layer for calculating the CTC loss. The speech encoder transforms input speech features $\boldsymbol{x}^{\mathrm{src}}$ into a high-level continuous representation, and then the transcription decoder generates a probability distribution $P_{\mathrm{asr}}(\boldsymbol{y}^{\mathrm{src}}|\boldsymbol{x}^{\mathrm{src}}) = \prod_i P_{\mathrm{asr}}(y_i^{\mathrm{src}}|y_{<i}^{\mathrm{src}}, \boldsymbol{x}^{\mathrm{src}})$ conditioned over all previously generated tokens. We adopt a location-based scoring function [30]. During training, parameters are updated so as to minimize the linear interpolation of the negative log-likelihood $\mathcal{L}_{\mathrm{att}} = -\log P_{\mathrm{att}}(\boldsymbol{y}^{\mathrm{src}}|\boldsymbol{x}^{\mathrm{src}})$ and the CTC loss $\mathcal{L}_{\mathrm{ctc}} = -\log P_{\mathrm{ctc}}(\boldsymbol{y}^{\mathrm{src}}|\boldsymbol{x}^{\mathrm{src}})$ with a tunable parameter λ $(0 \leq \lambda \leq 1)$: $\mathcal{L}_{\mathrm{asr}} = (1 - \lambda)\mathcal{L}_{\mathrm{att}} + \lambda\mathcal{L}_{\mathrm{ctc}}$. During the inference, left-to-right beam search decoding is performed jointly with scores from both an external recurrent neural network language model (RNNLM) [31] (referred to as *shallow fusion*) and the CTC outputs. We refer the readers to [27, 28] for more details.

For multilingual ASR models, we prepend the corresponding language ID to reference labels so that the decoder can jointly identify the target language while recognizing speech explicitly, which

can be regarded as multi-task learning with ASR and language identification tasks [11].

2.1.2. Text normalization

In this work, we skip punctuation restoration for the simplicity[5]. Instead, we train the MT model so that it translates source references without punctuation marks to target references with them, where text normalization task is jointly conducted with the MT task and it can be seen as multi-task learning. During inference, the MT model consumes hypotheses from the ASR model.

2.1.3. Neural machine translation (NMT)

Our NMT model consists of the source embedding, text encoder, and translation decoder. The text encoder maps a sequence of source tokens $\boldsymbol{y}^{\mathrm{src}}$ into the distributed representation following the source embedding layer. The translation decoder generates a probability distribution $P(\boldsymbol{y}^{\mathrm{tgt}}|\boldsymbol{y}^{\mathrm{src}})$. The only differences between the transcription and translation decoders are the score function for the attention mechanism. We adopt an additive scoring function [24]. Optimization is performed so as to minimize the negative log-likelihood $-\log P(\boldsymbol{y}^{\mathrm{tgt}}|\boldsymbol{y}^{\mathrm{src}})$.

2.2. End-to-end speech translation (E2E-ST)

Our end-to-end speech translation (E2E-ST) model is composed of the speech encoder and translation decoder. To compare strictly, we use the same speech encoder and translation decoder as ASR and NMT tasks, respectively. Parameters are updated so as to minimize the negative log-likelihood $-\log P(\boldsymbol{y}^{\mathrm{tgt}}|\boldsymbol{x}^{\mathrm{src}})$.

3. MULTILINGUAL E2E SPEECH TRANSLATION

We now propose an efficient framework that extends the bilingual E2E-ST model described previously to a multilingual one.

3.1. Universal sequence-to-sequence model

We adopt a universal sequence-to-sequence architecture instead of preparing separate parameters per language pair for four reasons. First, E2E-ST can be generally considered as a more challenging task than MT due to its more complex encoder, which requires more parameters (e.g., VGG+BLSTM). In addition, training sentences in standard ST corpora are much smaller than MT tasks (<300k) although input speech frames are much longer than text. Therefore, by sharing all parts, the total number of parameters are also reduced considerably and the E2E-ST model can have more training samples for better translation performance. Furthermore, it is not necessary to change the existing architecture. Second, we do not have to carefully pre-define a mini-batch scheduler for the language cycle as in [33] (see Section 3.3). Third, translation performance in low-resource directions can be improved by the aid of high-resource language pairs. Fourth, we can realize zero-shot translation in a direction which has never been seen during training [16].

3.2. Target language biasing

To perform translations for multiple target languages with a single decoder, we have to specify a target language to translate to. In

---

[3]We leave to investigate Transformer architectures [23] for future work. However, our framework is model agnostic and can be applied to any sequence-to-sequence models.

[4]Although we also conducted experiments with byte-pair-encoding (BPE) [26], the character unit is better than BPE in all settings due to the data sparseness issue. Therefore, we only report results on the character-level unit.

[5]In this paper, we use lowercased references. Therefore, we do not consider truecasing as text normalization.

Translation	Corpus	#hours	#utterances	#words	#vocab	domain
Bilingual	(A) Fisher-CallHome Spanish (Es→En)	170	138 k	1.7 M	66	conversation
	(B) Librispeech (En→Fr)	99	45 k†	0.8 M	112	reading
	(C) ST-TED (En→De)	203	133 k	2.2 M	109	lecture
One-to-many (O2M)	(B) + (C) (En→{Fr, De})	302	178 k	3.3 M	153	mixed
Many-to-many (M2Ma)	(A) + (B) ({En, Es}→{Fr, En})	269	183 k	2.8 M	121	mixed
Many-to-many (M2Mb)	(A) + (C) ({En, Es}→{De, En})	373	272 k	4.0 M	119	mixed
Many-to-many (M2Mc)	(A) + (B) + (C) ({En, Es}→{Fr, De, En})	472	317 k	5.1 M	157	mixed

Table 1: Statistics in each corpus. Each value is calculated <u>after</u> normalizing references and removing short and long utterances. *Speed perturbation* based data augmentation [32] is not performed here. †Two translation references are prepared per source speech utterance.

[16, 17, 18], an artificial token to represent the target language (target language ID) is prepended in the source sentence. However, this is not suitable for the ST task since the ST encoder directly consumes speech features. Instead, we replace a start-of-sentence ($\langle sos \rangle$) token in the decoder with a target language ID $\langle 2lang \rangle$ (see Figure 1). For example, when English speech is translated to French text, $\langle sos \rangle$ is replaced with French ID token $\langle 2fr \rangle$.

3.3. Mixed data training

We train multilingual models with mixed training data from multiple languages. Thus, each mini-batch may contain utterances from different language pairs. We bucket all samples so that each mini-batch contains utterances of speech frames of the similar lengths regardless of language pairs. As a result, we can use the same training scheme as the conventional ASR and bilingual ST tasks.

4. DATA

We build our systems on three speech translation corpora: Fisher-CallHome Spanish, Librispeech, and Speech-Translation TED (ST-TED) corpus. To the best of our knowledge, these are the only public available corpora recorded with a reasonable size of real speech data[6]. The data statistics are summarized in Table 1.

4.1. Bilingual translation

(A) Fisher-CallHome Spanish: Es→En

This corpus contains about 170-hours of Spanish conversational telephone speech, the corresponding transcription, and the English translations[7] [19]. Following [4, 19, 35], we report results on the five evaluation sets: *dev*, *dev2*, and *test* in Fisher corpus (with four references), and *devtest* and *evltest* in CallHome corpus (with a single reference). We use the Fisher/*train* as the training set and Fisher/*dev* as the validation set. All punctuation marks except for apostrophe are removed during evaluation in ST and MT tasks to compare with previous works [4, 19].

(B) Librispeech: En→Fr

This corpus is a subset of the original Librispeech corpus [36] and contains 236-hours of English read speech, the corresponding transcription, and the French translations [20]. We use the clean 99-hours of speech data for the training set [5]. Translation references in the training set are augmented with Google Translate following [5], so we have two French references per utterance. We use the *dev* set as the validation set and report results on the *test* set.

(C) Speech-Translation TED (ST-TED): En→De

This data contains 271-hours of English lecture speech, the corresponding transcription, as well as the German translation[8]. Since the original training set includes a lot of noisy utterances due to low alignment quality, we take a data cleaning strategy. We first force-aligned all training utterances with a Gentle forced aligner[9] based on Kaldi [37], then excluded all utterances where all words in the transcription were not perfectly aligned with the corresponding audio signal [38]. This process reduced from 171,121 to 137,660 utterances. We sampled two sets of 2k utterances from the cleaned training data as the validation and test sets, respectively (totally 4k utterances). Note that all sets have no text overlap and are disjoint regarding speakers, and data splits are available in our codes. We report results on this test set and *tst2013*. *tst2013* is one of the test sets provided in IWSLT2018 evaluation campaign. Since there are no human-annotated time alignment provided in these test sets, we decided to sample the disjoint test set from the training data with alignment information.

4.2. Multilingual translation

We perform experiments in two scenarios: *one-to-many* (O2M) and *many-to-many* (M2M)[10].

One-to-many (O2M)

For one-to-many (O2M) translation, speech utterances in a source language are translated to multiple target languages. We concatenate Librispeech (En→Fr) and ST-TED (En→De), and build models for En→{Fr, De} translations (see Table 1).

Many-to-many (M2M)

For many-to-many (M2M) translation, speech utterances in multiple source languages are translated to all target languages given in training. We can regard this task as a more challenging optimization problem than O2M and M2O translations. We concatenate Librispeech (En→Fr) and Fisher-CallHome Spanish (Es→En), then build models for {En, Es}→{Fr, En} translations (M2Ma)[11]. Other combinations such as Fisher-CallHome Spanish and ST-TED ({En, Es}→{De, En}, M2Mb), and all three directions ({En, Es}→{Fr, De, En}, M2Mc) are also investigated.

Bi-∗: Bilingual, Mono-∗: Monolingual, Multi-∗: Multilingual

Table 2

	Model	Multilingual	Fisher			CallHome	
			dev	dev2	test	devtest	evltest
			BLEU (↑)				
MT	Bi-SMT [35]	–	–	65.4	62.9	–	–
	Bi-NMT [4]	–	58.7	59.9	57.9	28.2	27.9
	Bi-NMT [39]	–	61.9	62.8	60.4	–	–
	Bi-NMT	–	**60.6**	**62.0**	**59.6**	**29.4**	**28.9**
	Multi-NMT	M2Ma	50.2	50.6	49.5	22.8	22.8
	Multi-NMT	M2Mb	57.4	58.3	56.7	27.9	27.7
	Multi-NMT	M2Mc	56.7	57.5	56.2	27.8	27.7
E2E ST	Bi-ST [4]	–	46.5	47.3	47.3	16.4	16.6
	+ ASR task [4]	–	48.3	49.1	48.7	16.8	17.4
	(E-B-1) Bi-ST	–	40.4	41.4	41.5	14.1	14.2
	(E-Ma-1) Multi-ST	M2Ma	41.1	41.7	41.3	15.1	15.2
	(E-Mb-1) Multi-ST	M2Mb	43.5	44.5	44.2	15.3	15.8
	(E-Mc-1) Multi-ST	M2Mc	**44.1**	**45.4**	**45.2**	**16.4**	**16.2**
Pipe ST	Mono-ASR/Bi-SMT [35]	–	–	–	40.4	–	–
	Mono-ASR/Bi-NMT [4]	–	45.1	46.1	45.5	16.2	16.6
	(P-B) Mono-ASR/Bi-NMT	–	**37.3**	39.6	**38.6**	16.8	16.5
	(P-Ma) Multi-ASR/Bi-NMT	M2Ma	**37.9**	**40.3**	**39.2**	**17.6**	**17.2**
	(P-Mb) Multi-ASR/Bi-NMT	M2Mb	37.6	39.6	38.9	17.0	17.0
	(P-Mc) Multi-ASR/Bi-NMT	M2Mc	37.6	39.7	38.5	17.0	16.9
	Model		WER (↓)				
ASR	Mono-ASR [4]	–	25.7	25.1	23.2	44.5	45.3
	Mono-ASR (Es)	–	26.0	25.6	23.6	45.4	45.9
	Multi-ASR (Es, En)	M2Ma	**25.6**	**25.0**	**22.9**	**43.5**	44.5
	Multi-ASR (Es, En)	M2Mb	25.9	25.2	23.3	44.2	44.7
	Multi-ASR (Es, En)	M2Mc	26.0	25.4	23.6	44.5	**44.2**

Table 2: Results of MT, ST, and ASR systems on Fisher-CallHome Spanish (Es→En). (E-B-1): Bilingual E2E-ST. (E-Ma/Mb/Mc-1): Proposed *many-to-many* (M2M) E2E-ST. (P-B): Bilingual pipeline-ST. (P-Ma/Mb/Mc): M2M pipeline-ST.

Table 3

	Model	Multilingual	BLEU (↑)
MT	Bi-NMT [5]	–	19.2
	Google Translate [5]	–	22.2
	Bi-NMT	–	**18.3**
	Multi-NMT	O2M	16.2
	Multi-NMT	M2Ma	12.2
	Multi-NMT	M2Mc	14.8
E2E ST	Bi-ST [5]	–	12.9
	+ Pre-training + MTL [5]	–	13.4
	Bi-ST + KD [7]	–	17.0
	(E-B-1) Bi-ST	–	15.7
	(E-O-1) Multi-ST	O2M	17.2
	(E-Ma-1) Multi-ST	M2Ma	16.4
	(E-Mc-1) Multi-ST	M2Mc	**17.3**
Pipe ST	Mono-ASR/Bi-NMT [5]	–	14.6
	(P-B) Mono-ASR/Bi-NMT	–	15.8
	(P-O) Mono-ASR†/Bi-NMT	O2M	**16.7**
	(P-Ma) Muti-ASR/Bi-NMT	M2Ma	16.4
	(P-Mc) Muti-ASR/Bi-NMT	M2Mc	**16.7**
	Model		WER (↓)
ASR	Mono-ASR [5]	–	17.9
	Mono-ASR (En)	–	9.0
	Mono-ASR† (En)	O2M	**6.6**
	Multi-ASR (En, Es)	M2Ma	8.6
	Multi-ASR (En, Es)	M2Mc	6.8

Table 3: Results of MT, ST, and ASR systems on Librispeech (En→Fr). †Training data is augmented with ST-TED. (E-O-1): Proposed *one-to-many* (O2M) E2E-ST. (P-O): O2M pipeline-ST.

5. EXPERIMENTAL EVALUATIONS

5.1. Settings

For data pre-processing of references in all languages, we lowercased and normalized punctuation, followed by tokenization with the `tokenizer.perl` script in the Moses toolkit[12]. For source references, we further removed all punctuation marks except for apostrophe. We report case-insensitive BLEU [40] with the `multi-bleu.perl` script in Moses. The character vocabulary was created jointly with both source and target languages.

We used 80-channel log-mel filterbank coefficients with 3-dimensional pitch features, computed with a 25ms window size and shifted every 10 ms using Kaldi [37], resulting 83-dimensional features per frame. The features were normalized by the mean and the standard deviation for each training set. We augmented speech data by a factor of 3 by *speed perturbation* [32]. We removed utterances having more than 3000 frames or more than 400 characters due to the GPU memory efficiency.

The speech encoders in ASR and ST models were composed of two VGG blocks [41] followed by 5-layers of 1024-dimensional (per direction) bidirectional long short-term memory (LSTM) [42]. Each VGG-like block composed of 2-layers of CNN having a 3×3 filter followed by a max-pooling layer with a stride of 2×2, which resulted in 4-fold time reduction. The text encoders in MT models were composed of 2-layers of 1024-dimensional (per direction) BLSTM. Both transcription and translation decoders were two layers of unidirectional LSTM with 1024-dimensional memory cells. The dimensions of the attention layer and embeddings for decoders were set to 1024. We used 2-layers of LSTM LM with 1024 memory cells for shallow fusion as discussed in Section 2.1.1.

---

[12] https://github.com/moses-smt/mosesdecoder

Training was performed using Adadelta [43] for sequence-to-sequence models and Adam [44] for RNNLM. For regularization, we adopted dropout [45], label smoothing [46], scheduled sampling [47], and weight decay. Beam search decoding was performed with a beam width of 20 with CTC and LM scores in the ASR task as shown in Section 2.1.1, and a beam width of 10 with a length penalty in ST and MT tasks. Detailed hyperparameter settings during training and decoding are available in our codes.

5.2. Baseline results: Bilingual systems

First, we evaluate baseline bilingual MT and ST systems. Bilingual E2E-ST and pipeline-ST models are labeled (E-B-1) and (P-B) in each table, respectively.

(A) Fisher-CallHome Spanish: Es→En

We present our results on Fisher-CallHome Spanish (hereafter, Fisher-CallHome) in Table 2. ASR and NMT results were competitive to the previous work [4] while the E2E-ST and pipeline-ST models underperformed it. Note that our translation decoders in E2E-ST and NMT models were trained so as to predict lowercased references with punctuation marks to compare with multilingual models, unlike previous works [4, 19, 35], where all punctuation marks except for apostrophe are removed. For the comparison of our E2E-ST and pipeline-ST models, the baseline bilingual E2E-ST model (E-B-1) outperformed the pipeline-ST model (P-B) in the Fisher sets but underperformed it in the CallHome sets. To investigate this discrepancy, we evaluated them with a single reference in the Fisher tests, which results in 26.4/28.2/27.7 (Pipe-ST) vs. 23.5/25.2/24.8 (E2E-ST) and the pipeline system was shown to be better. This is intuitive since the E2E-ST model skipped the ASR decoder, RNNLM in the source language, and MT encoder parts.

Model		Multi-lingual	test	tst2013
			BLEU (↑)	
MT	Bi-NMT	–	**23.0**	**24.9**
	Multi-NMT	O2M	18.9	20.3
	Multi-NMT	M2Mb	17.5	18.7
	Multi-NMT	M2Mc	17.2	18.0
E2E ST	(E-B-1) Bi-ST	–	16.0	12.5
	(E-O-1) Multi-ST	O2M	17.6	14.4
	(E-Mb-1) Multi-ST	M2Mb	16.7	12.9
	(E-Mc-1) Multi-ST	M2Mc	**17.7**	**14.8**
Pipe ST	(P-B) Mono-ASR/Bi-NMT	–	18.1	13.1
	(P-O) Mono-ASR†/Bi-NMT	O2M	**18.5**	**14.0**
	(P-Mb) Multi-ASR†/Bi-NMT	M2Mb	17.7	12.6
	(P-Mc) Multi-ASR†/Bi-NMT	M2Mc	18.1	13.3
Model			WER (↓)	
ASR	Mono-ASR (En)	–	20.3	36.6
	Mono-ASR† (En)	O2M	**19.0**	**33.9**
	Multi-ASR (En, Es)	M2Mb	20.5	38.7
	Multi-ASR (En, Es)	M2Mc	20.1	36.5

Table 4: Results of MT, ST, and ASR systems on ST-TED (En→De). †Training data is augmented with Librispeech.

In our preliminary experiments, we confirmed the E2E-ST model can outperform the pipeline system by stacking more BLSTM layers on top of the speech encoder to match the number of parameters between them. Moreover, pre-training the speech encoder and translation decoder with the corresponding ASR encoder and NMT decoder also drastically improved the performances (see Table 5 in Section 5.3). However, it is worth noting that our goal in this paper is to show the effectiveness of multilingual training for E2E-ST models and therefore we will not seek these directions here.

(B) Librispeech: En→Fr

Next, results on Librispeech are shown in Table 3. Monolingual ASR, bilingual E2E-ST (E-B-1), and pipeline-ST (P-B) models outperformed the previous work [5]. The baseline bilingual E2E-ST model (E-B-1) showed the competitive performance compared to the pipeline-ST model (P-B).

(C) ST-TED: En→De

Results on ST-TED are shown in Table 4. Contrary to the above results, there is a large gap between the bilingual E2E-ST (E-B-1) and pipeline-ST (P-B) models in this corpus.

5.3. Main results: Multilingual systems

We now test multilingual models trained in two scenarios: *many-to-many* (M2M) and *one-to-many* (O2M) translations.

Many-to-many (M2M)

Results of M2M models on Fisher-CallHome, Librispeech, ST-TED are shown at the (*-Ma/Mb/Mc-1) lines in Table 2, Table 3, and Table 4, respectively. Ma, Mb, and Mc represent M2Ma, M2Mb, and M2Mc, respectively (see Table 1).

In Fisher-CallHome (Table 2), our M2M multilingual E2E-ST models (E-Mb/Mc-1) significantly outperformed the bilingual one (E-B-1) while (E-Ma-1) slightly outperformed (E-B-1) except for Fisher/*test*. Among three M2M E2E-ST models, (E-Mc-1) showed the best performance, from which we can confirm that additional training data from other language pairs is effective. Multilingual

ASR models slightly outperformed the monolingual ASR model. Performances of the MT models were degraded by multilingual training due to the domain mismatch especially for punctuation marks (see Table 1). In contrast, multilingual E2E-ST models were not affected by the domain mismatch issue since they are not conditioned on the source language text, which is one for the advantages of the end-to-end models.

In all pipeline systems in Fisher-CallHome, we used the bilingual MT model since it showed the best performance. Pipeline systems with the multilingual ASR (P-M∗) were consistently improved even though WER improvements were very small. Our multilingual E2E-ST models significantly outperformed all the pipeline models in the Fisher sets.

In Librispeech (Table 3), all M2M E2E-ST models (E-Ma/Mc-1) outperformed the bilingual one (E-B-1). Multilingual ASR models also outperformed the monolingual one. Pipeline systems (P-Ma/Mc) are improved in proportion to the WER improvements. However, E2E-ST models got more gains from multilingual training.

In ST-TED (Table 4), we also confirmed the consistent BLEU improvements by the proposed multilingual framework. The similar trends can be seen as in Fisher-CallHome and Librispeech.

One-to-many (O2M)

Results of O2M models on Librispeech and ST-TED are shown in Table 3 and Table 4, respectively. We also obtained significant improvements of the E2E-ST models from multilingual training as well as in the M2M scenario on both corpora. Since the amount of additional training data for O2M and M2Mb from ST-TED is 99-hours (+Librispeech) and 170-hours (+Fisher-CallHome), respectively, and the O2M E2E-ST model is better than the M2Mb E2E-ST model in ST-TED (see Table 4), we can conclude that *O2M training is more effective than M2M training* in terms of data efficiency. However, the combination of all training data (M2Mc) got a further small gain. We can confirm the effectiveness of O2M training from WER improvements in the ASR task (6.6 vs. 8.6 at the second and third lines from bottom in Table 3). Thus, further additional multilingual training data could lead to the improvement. Gains from multilingual training were larger in the E2E-ST model (E-O-1) than in the best pipeline model (P-O)[13]. Considering the fact that the O2M NMT model underperformed the bilingual one, O2M multilingual training benefits from not only additional English speech data but also the direct optimization, which is one of our motivations in this work.

Pre-training with the ASR encoder

Finally, we show results of pre-training with the ASR encoder in Table 5. We observed improvements by pre-training both in bilingual and multilingual cases, similar to [5, 8, 48]. Pre-training with the NMT decoder was not necessarily effective. The best multilingual E2E-ST with pre-training (E-Mc-2) outperformed the corresponding best pipeline system in all test sets.

In summary, the proposed multilingual framework has shown to be effective regardless of the language combination, corpus domain, and data size. Although it is possible to improve the pipeline systems by carefully designing the source representations between ASR and MT modules (e.g., adding punctuation restoration module), it can be overcome by simply optimizing the direct mapping from source speech to target text with punctuation marks as we have shown.

---

[13]The best monolingual ASR → the best bilingual NMT in Table 4

Model		Multi-lingual	BLEU (↑)							
			Fisher-CallHome					Librispeech	ST-TED	
			dev	dev2	test	devtest	evltest	test	test	tst2013
E2E-ST	(E-B-1) Bi-ST	–	40.4	41.4	41.5	14.1	14.2	15.7	16.0	12.5
	(E-B-2) + ASR-PT	–	43.5	45.1	44.7	15.6	16.4	16.3	17.1	13.1
	(E-B-3) + MT-PT	–	44.4	45.1	45.2	15.6	15.4	16.8	17.4	13.5
	(E-Mc-1) Multi-ST	M2Mc	44.1	45.4	45.2	16.4	16.2	17.3	17.7	**14.8**
	(E-Mc-2) + ASR-PT	M2Mc	**46.3**	**47.1**	**46.3**	**17.3**	**17.2**	**17.6**	**18.6**	14.6
Pipe-ST	Best system	–	37.9	40.3	39.2	17.6	17.2	16.7	18.5	14.0

Table 5: Results of the end-to-end ST systems with pre-training

6. TRANSFER LEARNING FOR A VERY LOW-RESOURCE LANGUAGE SPEECH TRANSLATION

In this section, we evaluate generalization of multilingual ST models by performing transfer learning to a very low-resource ST task. We used Mboshi-French corpus[14] [22], which contains 4.4-hours of spoken utterances and the corresponding Mboshi transcriptions and French translations. Mboshi [49] is a Bantu C25 language spoken in Congo-Brazzaville and does not have standard orthography. We sampled 100 utterances from the training set as the validation set, and report results on the dev set (514 utterances) as in [50, 48].

We tried four different ways to transfer a non-Mboshi E2E-ST model to this task. In the bilingual case, we used the bilingual ST model in Librispeech ((E-B-1) in Table 3) as seed, then fine-tuned on the Mboshi-French data. In the multilingual case, we tried seeding with multilingual ST models in M2Ma (E-Ma-1), M2Mc (E-Mc-1), and O2M (E-O-1) settings. All parameters including the output layer are transferred from pre-trained ST models and we do not include any characters in Mboshi transcriptions in the vocabularies. Note that French references appear in the target side of all seed models during the pre-training stage.

Results are shown in Table 6. Multilingual E2E-ST models are more effective than the bilingual one, and O2M showed the best performance among three models. Although our transferred models underperformed [48], it is worth mentioning that they used other English ASR data (Switchboard corpus) and initialized the decoder with the French ASR decoder. Further improvements could be possible by leveraging Mboshi transcriptions, but we did not use any prior knowledge about Mboshi characters. This is a desired scenario for endangered language documentation and quite useful for automatic word discovery [50, 51, 52].

Seed		Multi-lingual	BLEU
Encoder	Decoder		
En300h-ASR	French20h-ASR [48]	–	**7.1**
Libri-ST	Libri-ST	–	4.55
O2M-ST	O2M-ST	✓	**6.92**
M2Ma-ST	M2Ma-ST	✓	5.50
M2Mc-ST	M2Mc-ST	✓	6.52

Table 6: Results of E2E-ST systems transferred from pre-trained E2E-ST models on a very low-resource corpus (Mboshi→Fr, 4.4 hours). The former and latter part of hyphen represents *data* and *task* for pre-training, respectively (*data-task*). Note that all models do not use any transcriptions in Mboshi during pre-training nor adaptation stage.

---

7. RELATED WORK

End-to-end speech translation

In [4], the E2E-ST model is simultaneously optimized with an auxiliary ASR task by sharing the whole encoder parameters. Pre-training approaches from the ASR encoder [48] and MT decoder are also investigated in [5, 8]. [8] proposed a data augmentation strategy, where weakly-supervised paired data is generated from monolingual source text data with text-to-speech (TTS) and MT systems (similar to back translation [53]) and speech data with a pipeline ST system (similar to knowledge distillation [53]). [50] proposed an efficient framework to better leverage higher-level intermediate representations by jointly attending to speech encoder and transcription decoder states. The most relevant work to ours is [48], where well-trained ASR parameters from the other language are used to initialize ST models and improve the ST performance in low-resource scenarios. Our work is distinct in that we focus on exploiting corpora in the multilingual setting and show that it outperforms the bilingual setting.

Multilingual ASR

In the multilingual ASR study, the language-independent acoustic representations can be obtained by sharing parameters, and then adapted to low-resource languages [54, 55, 56]. Recently, this approach is extended to end-to-end ASR paradigms: Connectionist Temporal Classification (CTC) [13], and attention-based encoder-decoder [11, 12, 14, 15]. Our work adopts this multilingual ASR in the pipeline system.

Multilingual NMT

Crosslingual parameter sharing approaches are investigated by tying a part of parameters [33, 57, 58], and even all parameters with a shared vocabulary [16, 17, 18] among multiple languages. Since the main drawback of the shared vocabulary is that the size of the vocabulary grows rapidly in proportion to the number of language pairs or the capacity per language shrinks when using BPE units [26], fully character-level multilingual framework is proposed to overcome the issue to some extent [59]. Our work is along with this trend of utilizing a universal translation model in one-to-many and many-to-many ST scenarios.

8. CONCLUSION AND FUTURE WORK

We performed multilingual training and end-to-end speech translation jointly, which has not yet been investigated before. We proposed a universal sequence-to-sequence framework and it outperformed the bilingual end-to-end, and the gap between strong pipeline systems became smaller. Its effectiveness was also confirmed by performing transfer learning to a very low-resource speech translation task. To encourage further research in this topic, we will place our codes to the public project. In future work, we will support more languages [7, 34] on our codebase and investigate multilingual training with non-related languages such as Chinese and Japanese.

9. REFERENCES

[1] Hermann Ney, "Speech translation: Coupling of recognition and translation," in *Proceedings of ICASSP*. IEEE, 1999, pp. 517–520.

[2] Eunah Cho, Jan Niehues, and Alex Waibel, "NMT-based segmentation and punctuation insertion for real-time spoken language translation.," in *Proceedings of Interspeech*, 2017, pp. 2645–2649.

[3] Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[4] Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Proceedings of Interspeech*, 2017, pp. 2625–2629.

[5] Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin, "End-to-end automatic speech translation of audiobooks," in *Proceedings of ICASSP*. IEEE, 2018, pp. 6224–6228.

[6] Salil Deena, Raymond WM Ng, Pranava Madhyastha, Lucia Specia, and Thomas Hain, "Exploring the use of acoustic embeddings in neural machine translation," in *Proceedings of ASRU*. IEEE, 2017, pp. 450–457.

[7] Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong, "End-to-end speech translation with knowledge distillation," *arXiv preprint arXiv:1904.08075*, 2019.

[8] Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *Proceedings of ICASSP*. IEEE, 2019.

[9] Mauro Cettolo, Christian Girardi, and Marcello Federico, "Wit3: Web inventory of transcribed and translated talks," in *Conference of European Association for Machine Translation*, 2012, pp. 261–268.

[10] Florian Dessloch, Thanh-Le Ha, Markus Müller, Jan Niehues, Thai Son Nguyen, Ngoc-Quan Pham, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Thomas Zenkel, et al., "Kit lecture translator: Multilingual speech translation with one-shot learning," in *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 2018, pp. 89–93.

[11] Shinji Watanabe, Takaaki Hori, and John R Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *Proceedings of ASRU*. IEEE, 2017, pp. 265–271.

[12] Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao, "Multilingual speech recognition with a single end-to-end model," in *Proceedings of ICASSP*. IEEE, 2018, pp. 4904–4908.

[13] Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black, "Sequence-based multi-lingual low resource speech recognition," in *Proceedings of ICASSP*. IEEE, 2018, pp. 4909–4913.

[14] Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, Shinji Watanabe, and Takaaki Hori, "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," in *Proceedings of SLT*. IEEE, 2018, pp. 512–527.

[15] Hirofumi Inaguma, Jaejin Cho, Murali Karthick Baskar, Tatsuya Kawahara, and Shinji Watanabe, "Transfer learning of language-independent end-to-end asr with language model fusion," in *Proceedings of ICASSP*. IEEE, 2019, pp. 6096–6100.

[16] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al., "Google's multilingual neural machine translation system: enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, 2016.

[17] Thanh-Le Ha, Jan Niehues, and Alexander Waibel, "Toward multilingual neural machine translation with universal encoder and decoder," in *Proceedings of IWSLT*, 2016.

[18] Thanh-Le Ha, Jan Niehues, and Alexander Waibel, "Effective strategies in zero-shot neural machine translation," in *Proceedings of IWSLT*, 2017, pp. 105–112.

[19] Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur, "Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus," in *Proceedings of IWSLT*, 2013.

[20] Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif, "Augmenting Librispeech with French translations: A multimodal corpus for direct speech translation evaluation," in *Proceedings of LREC*, 2018.

[21] Niehues Jan, Roldano Cattoni, Stüker Sebastian, Mauro Cettolo, Marco Turchi, and Marcello Federico, "The iwslt 2018 evaluation campaign," in *Proceedings of IWSLT*, 2018, pp. 2–6.

[22] Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélène Maynard, Markus Müller, et al., "A very low resource language speech corpus for computational language documentation experiments," *arXiv preprint arXiv:1710.03501*, 2017.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Processings of NIPS*, 2017.

[24] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of ICLR*, 2015.

[25] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings of ICASSP*. IEEE, 2016, pp. 4960–4964.

[26] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *Proceedings of ACL*, 2016, pp. 1715–1725.

[27] Takaaki Hori, Shinji Watanabe, and John Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proceedings of ACL*, 2017, pp. 518–529.

[28] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[29] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of ICML*, 2006, pp. 369–376.

[30] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Proceedings of NIPS*, 2015, pp. 577–585.

[31] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur, "Recurrent neural network based language model," in *Proceedings of Interspeech*, 2010.

[32] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of Interspeech*, 2015, pp. 3586–3589.

[33] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," in *Proceedings of NAACL-HLT*, 2016, pp. 866–875.

[34] Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "MuST-C: a multilingual speech translation corpus," in *Proceedings of the NAACL-HLT*, 2019, pp. 2012–2017.

[35] Gaurav Kumar, Matt Post, Daniel Povey, and Sanjeev Khudanpur, "Some insights from translating conversational telephone speech," in *Proceedings of ICASSP*. IEEE, 2014, pp. 3231–3235.

[36] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of ICASSP*. IEEE, 2015, pp. 5206–5210.

[37] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proceedings of ASRU*. IEEE, 2011.

[38] Mattia Antonino Di Gangi, Roberto Dessì, Roldano Cattoni, Matteo Negri, and Marco Turchi, "Fine-tuning on clean data for end-to-end speech translation: FBK@IWSLT 2018," in *Proceedings of IWSLT*, 2018, pp. 147–152.

[39] Elizabeth Salesky, Susanne Burger, Jan Niehues, and Alex Waibel, "Towards fluent translations from disfluent speech," in *Proceedings of SLT*. IEEE, 2018, pp. 921–926.

[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002, pp. 311–318.

[41] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of ICLR*, 2015.

[42] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[43] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[44] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[45] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of CVPR*, 2016, pp. 2818–2826.

[47] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proceedings of NIPS*, 2015, pp. 1171–1179.

[48] Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proceedings of NAACL-HLT*, 2019, pp. 58–68.

[49] Annie Rialland, Martine Adda-Decker, Guy-Noël Kouarata, Gilles Adda, Laurent Besacier, Lori Lamel, Elodie Gauthier, Pierre Godard, and Jamison Cooper-Leavitt, "Parallel corpora in Mboshi (Bantu C25, Congo-Brazzaville)," in *Proceedings of LREC*, 2018, pp. 4272–4276.

[50] Antonios Anastasopoulos and David Chiang, "Tied multitask learning for neural speech translation," in *Proceedings of NAACL-HLT*, 2018, pp. 82–91.

[51] Marcely Zanon Boito, Alexandre Bérard, Aline Villavicencio, and Laurent Besacier, "Unwritten languages demand attention too! word discovery with encoder-decoder models," in *Proceedings of ASRU*. IEEE, 2017, pp. 458–465.

[52] Pierre Godard, Marcely Zanon-Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, and Laurent Besacier, "Unsupervised word segmentation from speech with attention," in *Proceedings of Interspeech*, 2018, pp. 2678–2782.

[53] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of ACL*, 2016, pp. 86–96.

[54] Karel Veselỳ, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Proceedings of SLT*. IEEE, 2012, pp. 336–341.

[55] Ngoc Thang Vu, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, and Hervé Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proceeding of ICASSP*. IEEE, 2014, pp. 7639–7643.

[56] Martin Karafiát, Murali Karthick Baskar, Karel Veselỳ, František Grézl, Lukáš Burget, et al., "Analysis of multilingual blstm acoustic model on low and high resource languages," in *Proceeding of ICASSP*. IEEE, 2018, pp. 5789–5793.

[57] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang, "Multi-task learning for multiple language translation," in *Proceedings of ACL*, 2015, pp. 1723–1732.

[58] Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun, "A neural interlingua for multilingual machine translation," in *Proceedings of the Third Conference on Machine Translation (WMT)*, 2018, pp. 84–92.

[59] Jason Lee, Kyunghyun Cho, and Thomas Hofmann, "Fully character-level neural machine translation without explicit segmentation," *Transactions of the Association for Computational Linguistics*, 2017.