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, Japan
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA

Summary

- We evaluated if training data from other language pairs are helpful for the end-to-end speech translation (ST) task
- Directly translate source speech to target languages with a single sequence-to-sequence (S2S) model
  - ➢ Many-to-many (M2M)
  - ➢ One-to-many (O2M)
- Outperformed the bilingual end-to-end/pipeline speech translation models
- Performed transfer learning to a very low-resource ST task: Mboshi->French (4.4h)
  - ➢ Shared representations obtained from multilingual E2E-ST were more effective than those from the bilingual one

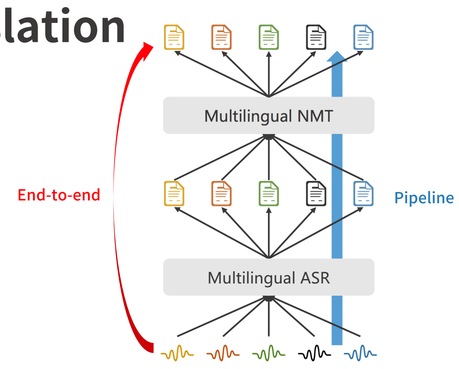
Background

◆ End-to-end speech-to-text translation (E2E-ST)

- Simplified architecture (no ASR decoder, LM, and MT encoder) -> reduced number of parameters
- Avoid error propagations from ASR

◆Multilingual speech-to-text translation

- Applications
  - ➢ One-to-many (O2M)
    e.g., lecture, news reading etc.
  - ➢ Many-to-many (M2M)
    e.g., dialogue, meeting etc.
- Conventional pipeline ST system
  - ➢ Language identification->ASR->text normalization->MT



Problems of multilingual pipeline ST systems

- ➢ Data sparseness issue for low-resource directions
- ➢ Mis-identification of source languages in ASR
- ➢ Increased number of parameters
- ➢ Need text normalization per source language

Proposed method

◆ Universal sequence-to-sequence (S2S) model

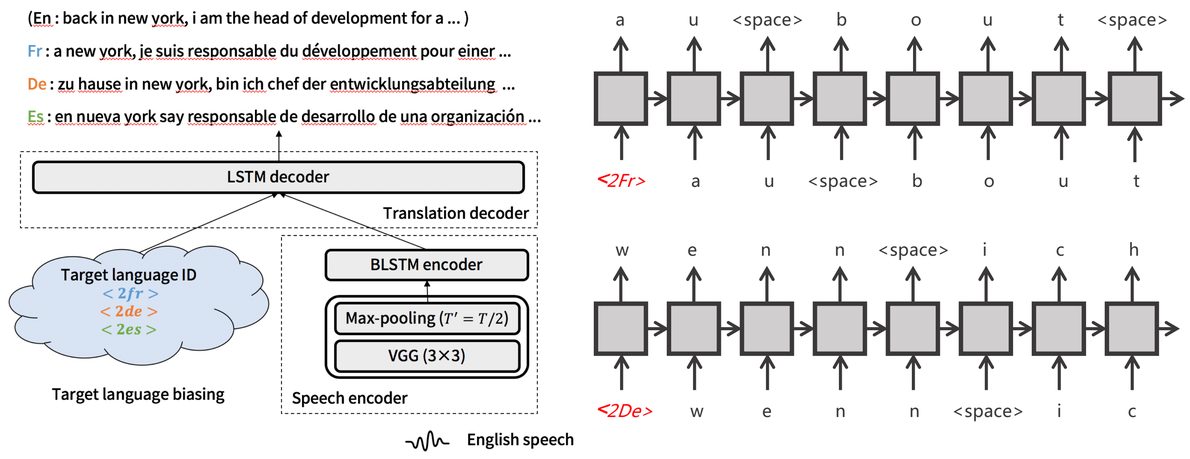
- A single S2S architecture (a shared encoder/decoder)
- Does not generate source sentence (no ASR)

Why end-to-end architecture for multilingual speech translation?

- ➢ Total number of parameters are drastically reduced
- ➢ No need to pre-define a mini-batch scheduler for the language cycle
- ➢ No need to change the existing architecture
- ➢ Potential zero-shot translation (not investigated in this work)

◆ Target language biasing

- Feed a target language ID token (e.g., $< 2Fr >$, $< 2De >$) to the translation decoder at the first step instead of $< sos >$



Data & Task

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

◆ One-to-many (O2M)

- Librispeech + ST-TED

◆ Many-to-many (M2M)

- M2Ma: Fisher + Librispeech
- M2Mb: Fisher + ST-TED
- M2Mc: Fisher + Librispeech + ST-TED

Experimental evaluations

◆ Main results

Fisher-CallHome (Es->En)

Model		BLEU	
		Fisher-test	CH-evltest
MT	Bi-NMT	**59.6**	**28.9**
	(M2Ma) Multi-NMT	49.5	22.8
	(M2Mb) Multi-NMT	56.7	27.7
	(M2Mc) Multi-NMT	56.2	27.7
E2E-ST	Bi-ST	41.5	14.2
	+ ASR pre-training	45.2	15.4
	(M2Ma) Multi-ST	41.3	15.2
	(M2Mb) Multi-ST	44.2	15.8
	(M2Mc) Multi-ST	**45.2**	**16.2**
	+ ASR pre-training	**46.3**	**17.2**
Pipeline-ST	Mono ASR -> Bi NMT	38.6	16.5
	(M2Ma) Multi-ASR -> Bi-NMT	**39.2**	**17.2**
	(M2Mb) Multi-ASR -> Bi-NMT	38.9	17.0
	(M2Mc) Multi-ASR -> Bi-NMT	38.5	16.9

Librispeech (En->Fr)

Model		BLEU
		test
MT	Bi-NMT	**18.3**
	(O2M) Multi-NMT	16.2
	(M2Ma) Multi-NMT	12.2
	(M2Mc) Multi-NMT	14.8
E2E-ST	Bi-ST	15.7
	+ ASR pre-training	16.3
	(O2M) Multi-ST	17.2
	(M2Ma) Multi-ST	16.4
	(M2Mc) Multi-ST	**17.3**
	+ ASR pre-training	**17.6**
Pipeline-ST	Mono ASR -> Bi NMT	15.8
	(O2M) Multi-ASR -> Bi-NMT	**16.7**
	(M2Ma) Multi-ASR -> Bi-NMT	16.4
	(M2Mc) Multi-ASR -> Bi-NMT	**16.7**

ST-TED (En->De)

Model		BLEU
		test
MT	Bi-NMT	**23.0**
	(O2M) Multi-NMT	18.9
	(M2Mb) Multi-NMT	17.5
	(M2Mc) Multi-NMT	17.2
E2E-ST	Bi-ST	16.0
	+ ASR pre-training	17.1
	(O2M) Multi-ST	17.6
	(M2Mb) Multi-ST	16.7
	(M2Mc) Multi-ST	**17.7**
	+ ASR pre-training	**18.6**
Pipeline-ST	Mono ASR -> Bi NMT	18.1
	(O2M) Multi-ASR -> Bi-NMT	**18.5**
	(M2Mb) Multi-ASR -> Bi-NMT	17.7
	(M2Mc) Multi-ASR -> Bi-NMT	18.1

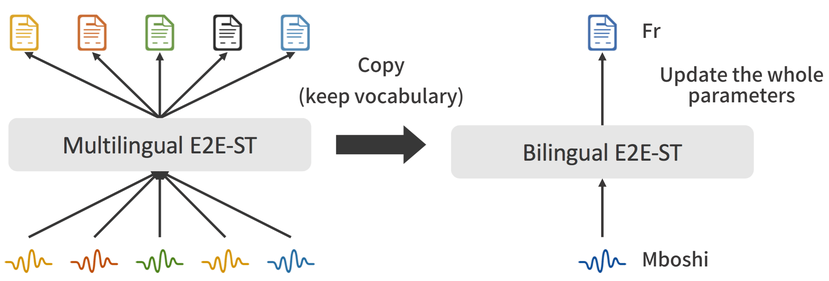
Take-home-message (many-to-many scenario)

- Additional training data from other language pairs was effective
- Multilingual E2E-ST models were not affected by the domain mismatch
- E2E-ST models got more gains from multilingual training than the pipeline systems

Take-home-message (one-to-many scenario)

- O2M training was more effective than M2M training from the perspective of data efficiency
- However, using all training data (M2Mc) got a further small gain
- O2M multilingual training benefits from not only additional English speech data but also the direct optimization

◆ Transfer learning to a very low-resource



Mboshi->Fr: 4.4h

Model	BLEU
	dev
Bi-ST (Librispeech)	4.55
O2M-ST	**6.92**
M2Ma-ST	5.50
M2Mc-ST	6.52

Take-home-message

- Multilingual E2E-ST seed was more effective than the bilingual one
- O2M seed showed the best performance among all models