



NAIST Simultaneous Speech-to-Text Translation System for IWSLT 2022

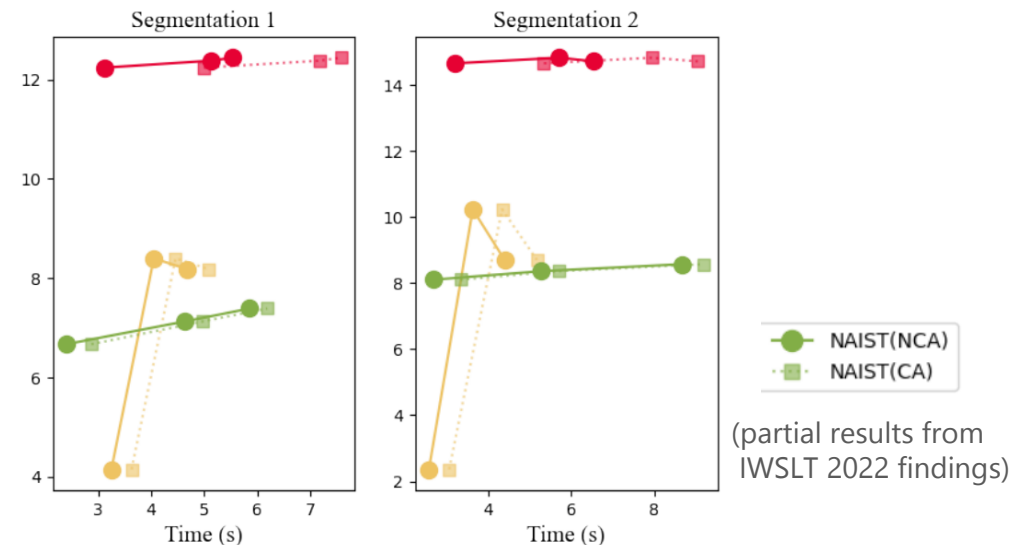
Ryo Fukuda¹, Yuka Ko¹, Yasumasa Kano¹, Kosuke Doi¹, Hiroataka Tokuyama¹,
Sakriani Sakti^{1,2}, Katsuhito Sudoh¹, and Satoshi Nakamura¹

¹Nara Institute of Science and Technology (NAIST), Japan

²Japan Advanced Institute of Science and Technology (JAIST), Japan

Brief Summary

- IWSLT 2022 Simultaneous Speech Translation task
 - Track: Text-to-Text, Speech-to-Text
 - Language: English \rightarrow {German, Japanese, Mandarin Chinese}
- We applied *Bilingual Prefix Alignment* (Kano et al., 2022) to speech-to-text
 - Boundary prediction model with adaptive segmentation policies
 - Prefix-to-prefix translation model fine-tuned on bilingual prefix pairs
- Our system did not perform as well as the other teams but did demonstrate robustness to low latency



Background

Simultaneous Machine Translation (SimulMT) approaches

- Fixed policy
 - **Wait- k** (Ma et al., 2019; Elbayad et al., 2020), ...
- Adaptive policy
 - **RL-based** (Gu et al., 2017), **Monotonic attention** (Arvazhagan et al., 2019; Ma et al., 2020), **Prefix Alignment** (Zhang et al., 2020, Kano et al., 2022), ...
- For simultaneous speech translation (SimulST), adaptive policies can be more effective than fixed policies
 - e.g. fixed policies provide output even when speech is paused

Bilingual Prefix Alignment (BPA) (Kano et al., 2022)

Extract **bilingual prefix pairs** and use them to

- (1) train a boundary prediction model and
- (2) fine-tune a offline translation model

	Read source words	Boundary Prediction	translation
Step 1	I	$\Rightarrow 0.9 > \mathbf{0.5} \Rightarrow$	私は
Step 2	I bought	$\Rightarrow 0.2 < \mathbf{0.5} \Rightarrow$	
Step 3	I bought a	$\Rightarrow 0.3 < \mathbf{0.5} \Rightarrow$	
Step 4	I bought a pen	$\Rightarrow 0.7 > \mathbf{0.5} \Rightarrow$	<u>私は</u> ペンを買った
Step 5	I bought a pen .	$\Rightarrow 0.7 > \mathbf{0.5} \Rightarrow$	<u>私はペンを買った。</u>

The translation process of SimulMT based on Bilingual Prefix Alignment.

Bilingual Prefix Alignment (BPA) (Kano et al., 2022)

Extract **bilingual prefix pairs** and use them to

- (1) train a boundary prediction model and
- (2) fine-tune a offline translation model

	Read source words	Boundary Prediction	translation
Step 1	I	$\Rightarrow 0.9 > \mathbf{0.5} \Rightarrow$	私は
Step 2	I bought	$\Rightarrow 0.2 < \mathbf{0.5} \Rightarrow$	
Step 3	I bought a	$\Rightarrow 0.3 < \mathbf{0.5} \Rightarrow$	
Step 4	I bought a pen	$\Rightarrow 0.7 > \mathbf{0.5} \Rightarrow$	<u>私は</u> ペンを買った
Step 5	I bought a pen .	$\Rightarrow 0.7 > \mathbf{0.5} \Rightarrow$	<u>私はペンを買った。</u>

The translation process of SimulMT based on Bilingual Prefix Alignment.

Bilingual Prefix Alignment (BPA) (Kano et al., 2022)

Extract **bilingual prefix pairs** and use them to

- (1) train a boundary prediction model and
- (2) fine-tune a offline translation model

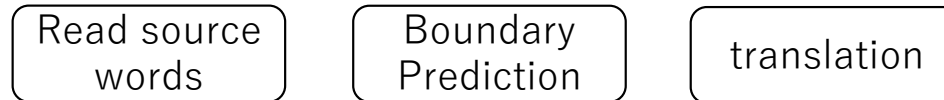
	Read source words	Boundary Prediction	translation
Step 1	I	$\Rightarrow 0.9 > 0.5 \Rightarrow$	私は
Step 2	I bought	$\Rightarrow 0.2 < 0.5 \Rightarrow$	
Step 3	I bought a	$\Rightarrow 0.3 < 0.5 \Rightarrow$	
Step 4	I bought a pen	$\Rightarrow 0.7 > 0.5 \Rightarrow$	<u>私は</u> ペンを買った
Step 5	I bought a pen .	$\Rightarrow 0.7 > 0.5 \Rightarrow$	<u>私はペンを買った。</u>

The translation process of SimulMT based on Bilingual Prefix Alignment.

Bilingual Prefix Alignment (BPA) (Kano et al., 2022)

Extract **bilingual prefix pairs** and use them to

- (1) train a boundary prediction model and
- (2) fine-tune a offline translation model



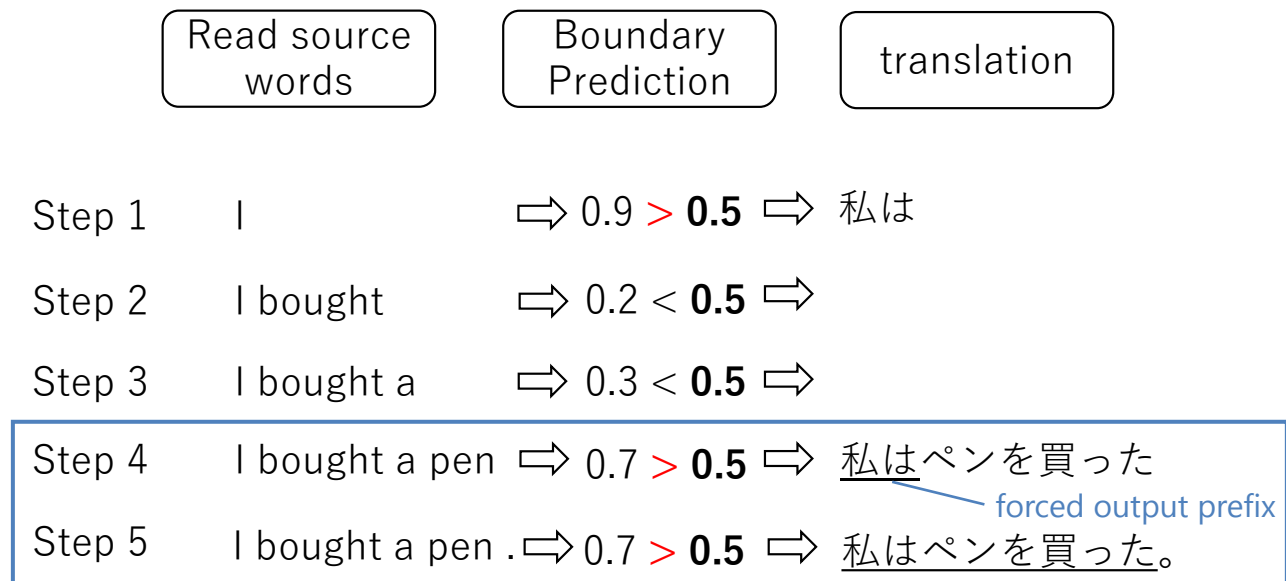
Step 1	I	$\Rightarrow 0.9 > \mathbf{0.5} \Rightarrow$	私は
Step 2	I bought	$\Rightarrow 0.2 < \mathbf{0.5} \Rightarrow$	
Step 3	I bought a	$\Rightarrow 0.3 < \mathbf{0.5} \Rightarrow$	
Step 4	I bought a pen	$\Rightarrow 0.7 > \mathbf{0.5} \Rightarrow$	<u>私は</u> ペンを買った
Step 5	I bought a pen .	$\Rightarrow 0.7 > \mathbf{0.5} \Rightarrow$	<u>私はペンを買った。</u>

The translation process of SimulMT based on Bilingual Prefix Alignment.

Bilingual Prefix Alignment (BPA) (Kano et al., 2022)

Extract **bilingual prefix pairs** and use them to

- (1) train a boundary prediction model and
- (2) fine-tune a offline translation model



The translation process of SimulMT based on Bilingual Prefix Alignment.

BPA - Stage 1. Extracting Prefix Pairs

Prefix
pairs

Full-sentence translation

私はペンを買った。

Output

私は。



NMT



Input

I

$x_{\leq 1}$

BPA - Stage 1. Extracting Prefix Pairs

Prefix pairs

[I, 私は]

Full-sentence translation

私はペンを買った。

Output

私は。

NMT

Input

I

$x_{\leq 1}$

common prefix

BPA - Stage 1. Extracting Prefix Pairs

Prefix pairs

[I, 私は]

Full-sentence translation

私はペンを買った。

ignore the same

Output

私は。



NMT



Input

I

$x_{\leq 1}$

私は買った。



NMT



I bought

$x_{\leq 2}$

BPA - Stage 1. Extracting Prefix Pairs

Prefix
pairs

[I, 私は]

Full-sentence translation

私はペンを買った。

Output

私は。



NMT



Input

I

$x_{\leq 1}$

私は買った。



NMT



I bought

$x_{\leq 2}$

私は買った。



NMT



I bought a

$x_{\leq 3}$

BPA - Stage 1. Extracting Prefix Pairs

Prefix pairs

[I, 私は]

[I bought a pen, 私はペンを買っ]

Full-sentence translation

私はペンを買った。

Output

私は。

私は買った。

私は買った。

私はペンを買って。

NMT

NMT

NMT

NMT

Input

I

I bought

I bought a

I bought a pen

$x_{\leq 1}$

$x_{\leq 2}$

$x_{\leq 3}$

$x_{\leq 4}$

BPA - Stage 1. Extracting Prefix Pairs

Prefix pairs

[I, 私は]

[I bought a pen, 私はペンを買っ]

Full-sentence translation

私はペンを買った。

Output

私は。



NMT



Input

I

$x_{\leq 1}$

私は買った。



NMT



I bought

$x_{\leq 2}$

私は買った。



NMT



I bought a

$x_{\leq 3}$

私はペンを買って。



NMT



I bought a pen

$x_{\leq 4}$

私はペンを買った。



NMT



I bought a pen .

$x_{\leq 5}$

BPA - Stage 2. Training of Models

(1) Train a boundary prediction model with prefix- $\{0,1\}$ label pairs

- [I, 1]
- [I bought, 0]
- [I bought a , 0]
- [I bought a pen , 1]
- [I bought a pen. , 1]

(2) Fine-Tune a offline translation model with prefix pairs

- [I, 私は]
- [I bought a pen, 私はペンを買っ]

SimulST based on Bilingual Prefix Alignment

Applying Bilingual Prefix Alignment to SimulST

	Read source words	Boundary Prediction	translation
Step 1	I	$\Rightarrow 0.9 > 0.5 \Rightarrow$	私は
Step 2	I bought	$\Rightarrow 0.2 < 0.5 \Rightarrow$	
Step 3	I bought a	$\Rightarrow 0.3 < 0.5 \Rightarrow$	
Step 4	I bought a pen	$\Rightarrow 0.7 > 0.5 \Rightarrow$	<u>私は</u> ペンを買った
Step 5	I bought a pen .	$\Rightarrow 0.7 > 0.5 \Rightarrow$	<u>私はペンを買った。</u>

- SimulST takes the source language speech as input
 - no word unit
 - extremely long sequences

Extracting Speech Prefix Pairs

Prefix pairs

[I, 私は]

[I bought a pen, 私はペンを買っ]

Full-sentence translation

私はペンを買った。

Output

私は。



NMT



Input

I

$x_{\leq 1}$

私は買った。



NMT



I bought

$x_{\leq 2}$

私は買った。



NMT



I bought a

$x_{\leq 3}$

私はペンを買って。



NMT



I bought a pen

$x_{\leq 4}$

私はペンを買った。



NMT



I bought a pen .

$x_{\leq 5}$

Extracting Speech Prefix Pairs

Prefix pairs

[, 私は]

[, 私はペンを買った]

Full-sentence translation

私はペンを買った。

Output

私は。

私は買った。

私は買った。

私はペンを買って。

私はペンを買った。



NMT

NMT

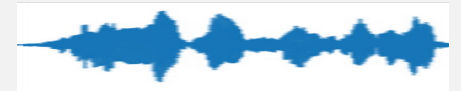
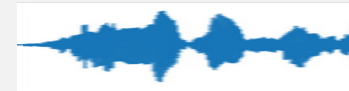
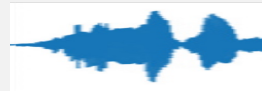
NMT

NMT

NMT



Input



$x_{\leq \tau}$

$x_{\leq 2\tau}$

$x_{\leq 3\tau}$

$x_{\leq 4\tau}$

$x_{\leq 5\tau}$

Background

Method

Experiments

Conclusions

τ : step size (number of frames)

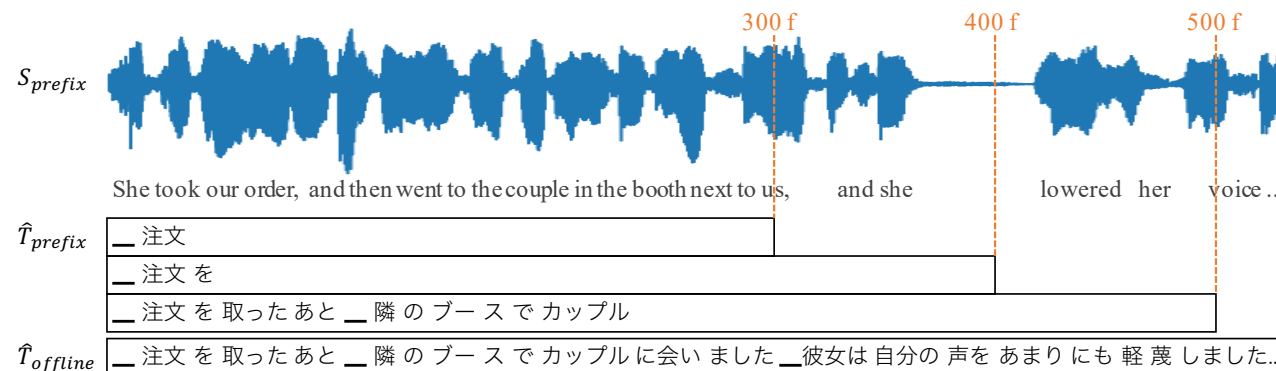
Data filtering of Prefix Pairs

Unbalanced prefix pairs were sometimes extracted

- pairs of a long source speech prefix and a short target text prefix

.. frequently appear between distant language pairs such as English and Japanese

- e.g. {English prefix, Japanese prefix} would consist of {S, S}, {SV, S}, {SVO, SOV}




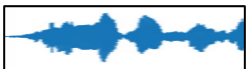



we removed such unbalanced pairs by the length ratio

- exclude if source prefix length / target prefix length $>$ *maxratio*

Boundary Predictor

- Training with pairs of a speech prefix and a corresponding binary label sequence

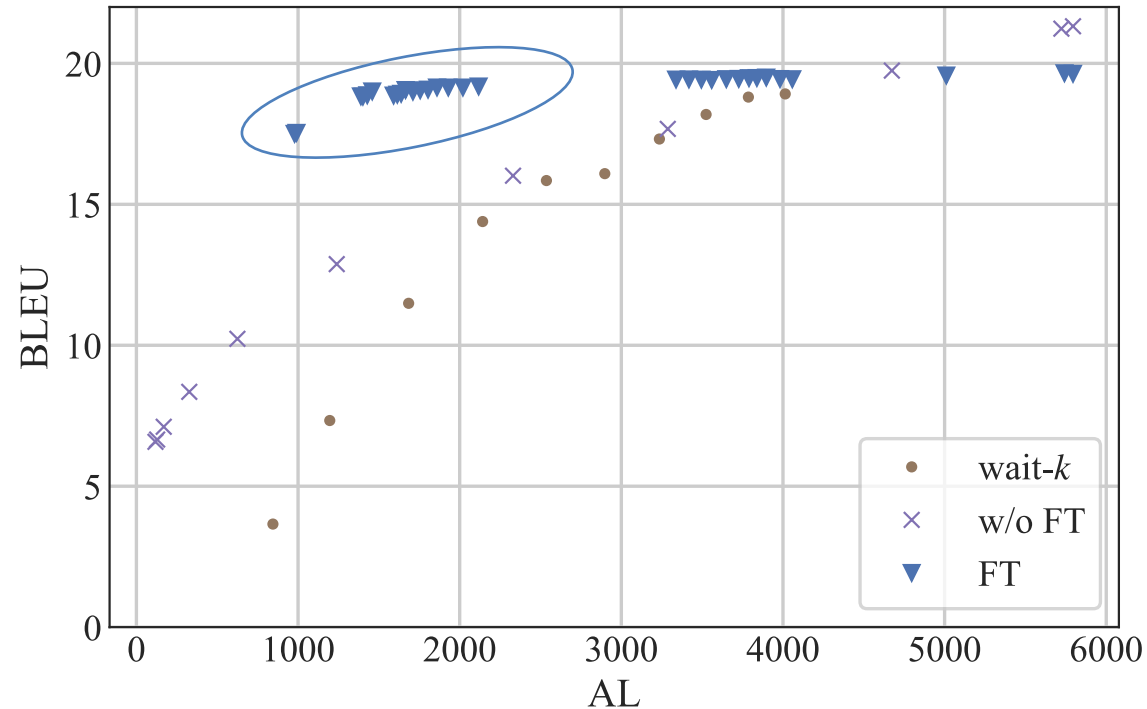
- { , [1..1] }
- { , [1..10..0] }
- { , [1..10..00..0] }
- { , [1..10..00..01..1] }
- { , [1..10..00..01..11..1] }

- The boundary predictor is trained with weighted cross-entropy loss
- The boundary predictor predicts a boundary in every τ frames
 - WRITE if the proportion of label 1 is larger than or equals to λ_{thre} , otherwise READ

Experimental setting

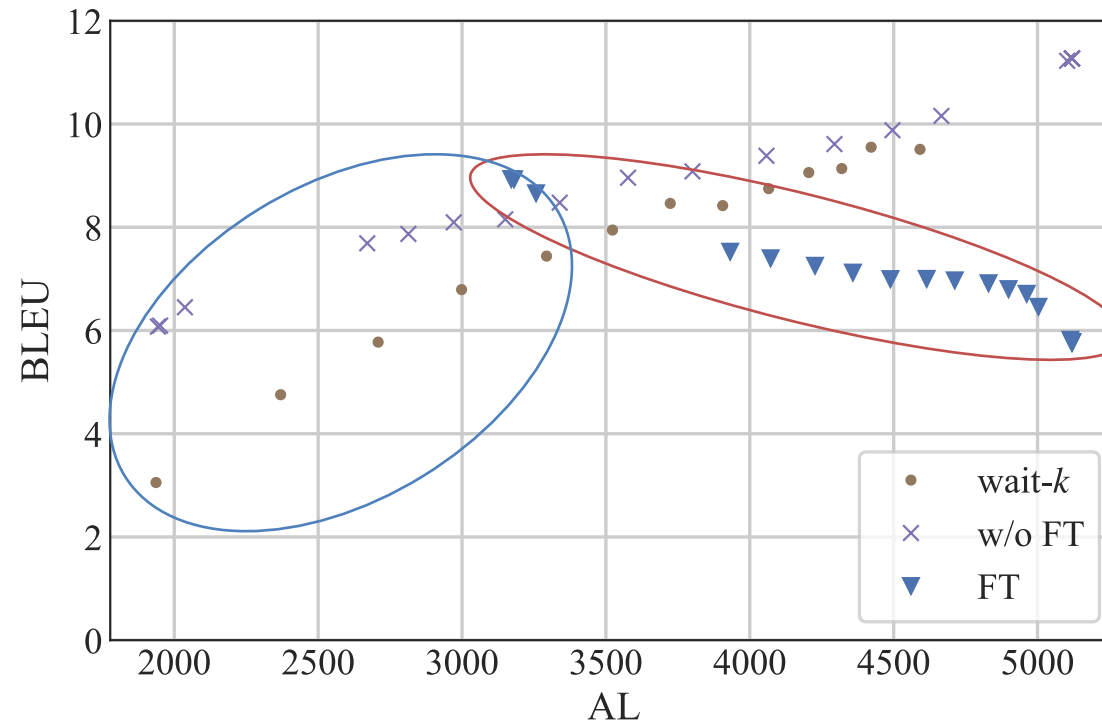
- Data: MuST-C v2
 - contained about 250k segments for En-De and 330k for En-Ja
- Models
 - **Speech Translation:** 12 encoder + 6 decoder Transformer layers
 - **Boundary prediction:** a 2D-convolution layer, a unidirectional LSTM layer, and an output linear layer
 - decision size: $\tau = 100$ frames
- Evaluated on MuST-C v2 tst-COMMON using SimulEval

En-De Results



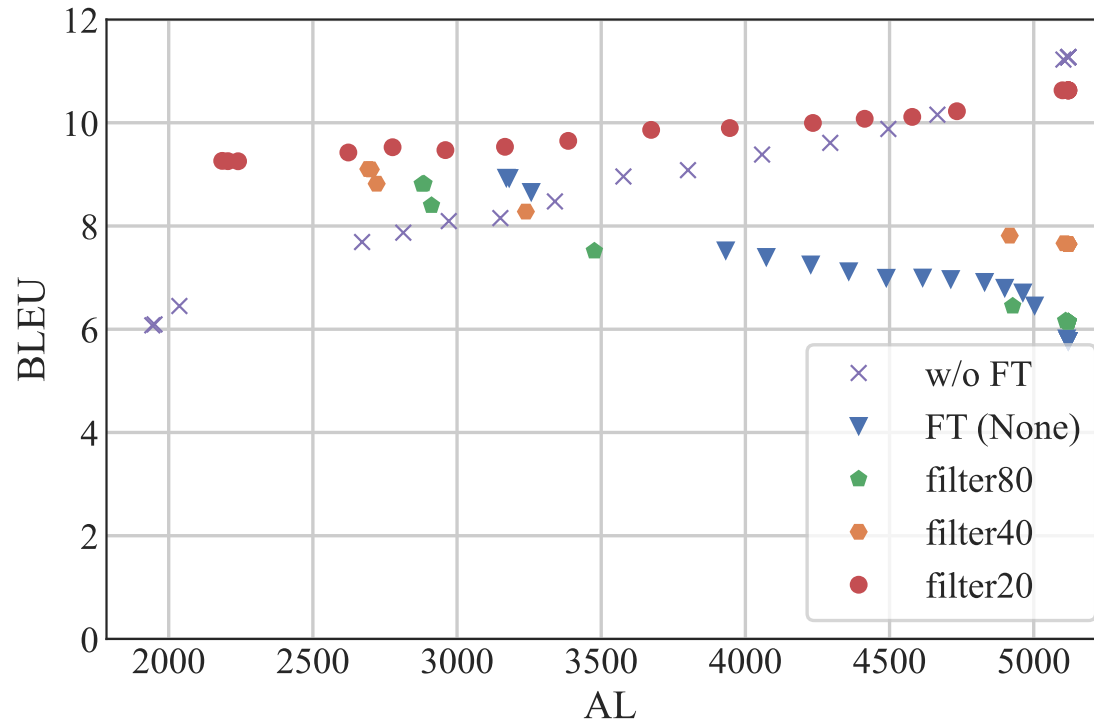
- a model fine-tuned with prefix pairs (FT) worked better than the non fine-tuned model (w/o FT) in the range of $AL \leq 4,000$
 - robust to lower latency
- non fine-tuned model worked better than the fine-tuned model in high latency

En-Ja Results



- non fine-tuned model worked better than wait-*k* baselines
 - large improvements in the low latency ranges
- fine-tuned model were worse than those of wait-*k* and non fine-tuned model
 - unbalanced prefix pairs degraded the performance

Data filtering for En-Ja



- the model fine-tuned with all prefix pairs (None) preferred too short outputs
- the model fine-tuned with filtered data on maxratio = 20 (filter20) significantly improved the performance and outperformed non fine-tuned model

Conclusions

- We described our SimulST systems in English-to-German and English-to-Japanese.
- We used *Bilingual Prefix Alignment* to
 - train boundary predictor that judges when to READ and WRITE and
 - fine-tune the offline speech translation model.
- Our system achieved some improvements compared to the wait- k baselines in every latency regime
 - especially, our system was robust to lower latency
 - data filtering was important for En-Ja because unbalanced prefix pairs frequently appeared due to differences in sentence structures