Findings of the Third Workshop on Automatic Simultaneous Translation

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Haifeng Wang
Liang Huang, Qun Liu, Julia Ive, Wolfgang Macherey

Shared Task @ The 3rd Workshop of AutoSimTrans

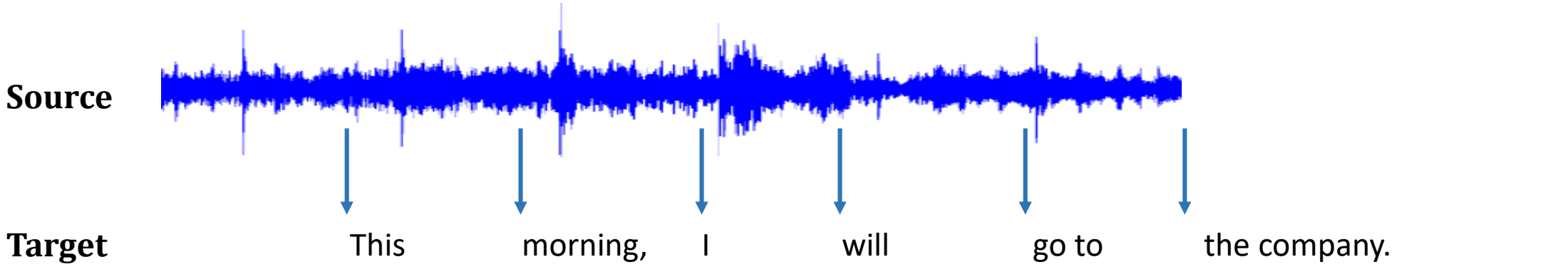
Three tracks:

- Chinese-English Text-to-text ST track
- Chinese-English Speech-to-text track
- English-Spanish Text-to-text track

1. Text-to-text Track

Source	今天	上午	我	要	去趟	公司。	
Target		This	morning,	I	will	go to	the company.

2. Speech-to-text Track



Participants & Submissions

Three tracks:

	#Submissions of 2021	#Submissions of 2022
• Chinese-English Text-to-text ST track	4	13
• Chinese-English Speech-to-text track	2	4
• English-Spanish Text-to-text track	0	7
Sum	6	24

14 participants:

Team	Organization
BIT-Xiaomi	Beijing Institute of Technology & Xiaomi Inc., Beijing, China
Huawei	Huawei Noah's Ark Lab, Guangdong, China
HAU	Huazhong Agricultural University, Hubei, China
USST-ECUST	Univ. of Shanghai for Science and Technology & East China Univ. of Science
HZLHZ	Anonymous
ZXN	Zhejiang Univ. & Xiamen Univ. & North China Institute of Aerospace Engineering
TMU	Tianjin Medical University, Tianjin, China
CITC	Changchun Information Technology College, Jilin, China
NCIAE	North China Institute of Aerospace Engineering, Hebei, China
XJTU	Xi'an Jiaotong University, Shanxi, China
HIT	Harbin Institute of Technology, Heilongjiang, China
ZJU	Zhejiang University, Zhejiang, China
Nuctech	Nuctech Company, Beijing, China
A23	Anonymous

Table 1: List of participants.

Introduction of the three tracks

	Features
Chinese-English Text-to-text ST track	Input: transcriptions of TED-like lectures, contain speech disfluencies but no ASR errors
Chinese-English Speech-to-text track	Input: speech
English-Spanish Text-to-text track	Input: official records, with no disfluencies and no ASR errors

Corpora:

	Corpus	Subset	Talks	Utterances	Transcription (words)	Translation (words)	Audio (hours)
Zh-En	BSTC (ST)	Train	215	37,901	1,004,128	620,263	64.57
		Dev	16	956	24,711	15,794	1.58
		Test	20	2,305	72,695	42,836	4.26
	CWMT19 (MT)	Train	/	9,023,456	264,652,945	182,840,035	/
En-Es	UN (MT)	Train	/	21,911,121	517,327,737	608,514,316	/
		Dev	/	500	12,400	14,701	/
		Test	/	500	13,421	15,935	/

Evaluation and Ranking Methods

Evaluation:

Translation Quality: BLEU

Latency: AL for text-to-text ST

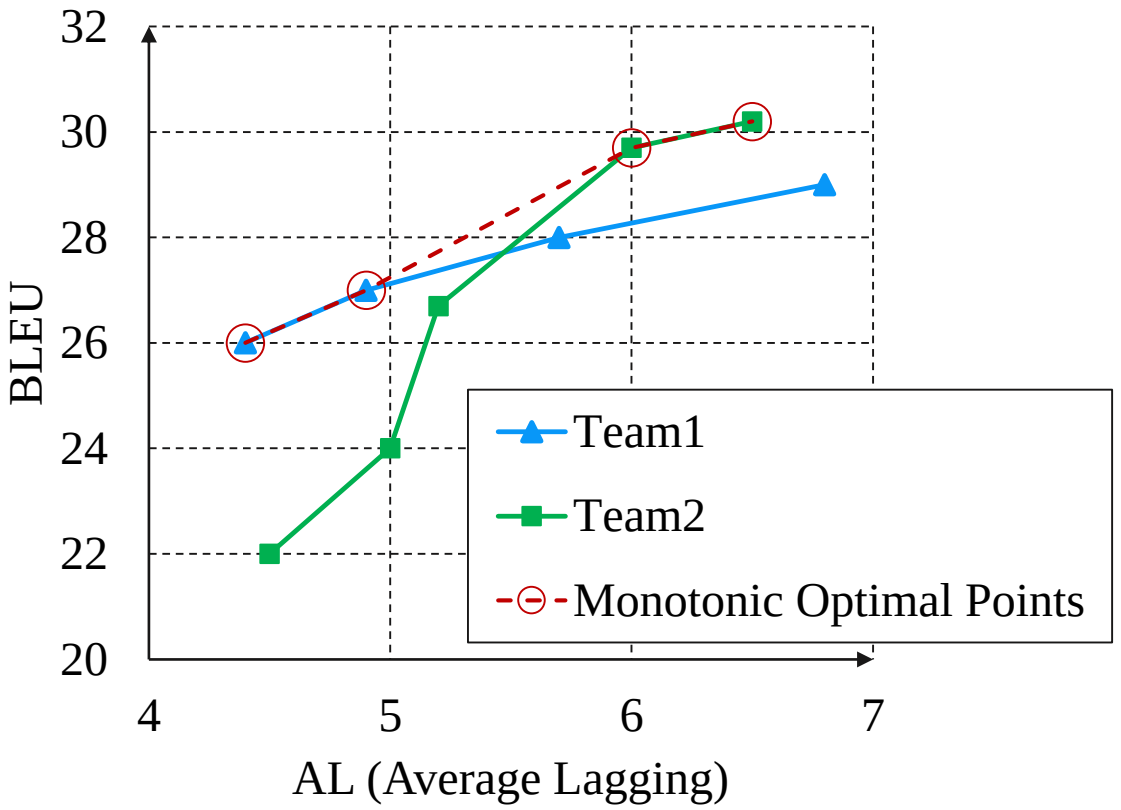
CW for speech-to-text ST

Ranking:

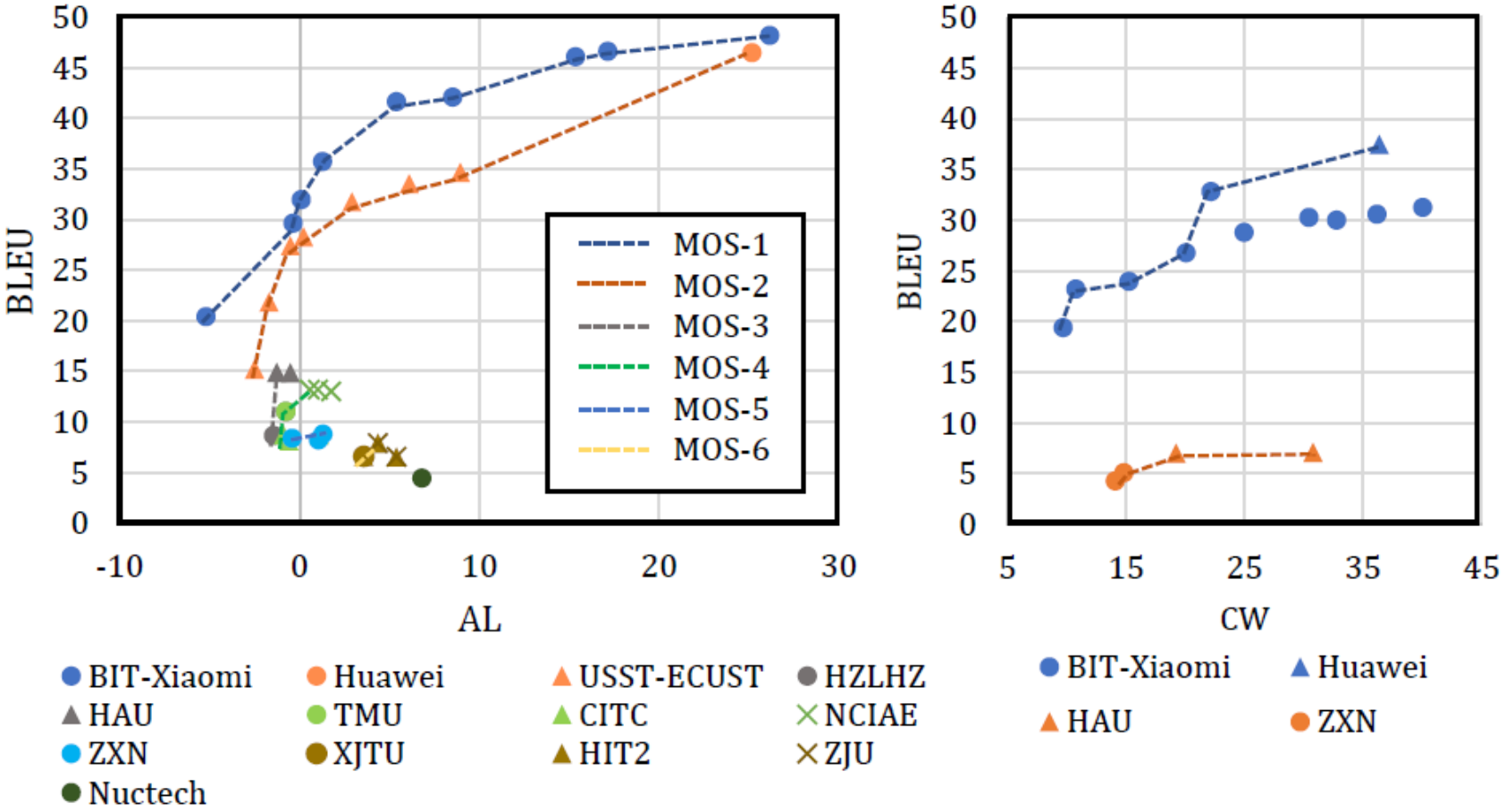
I-MOS algorithm: iteratively builds a monotonic optimal sequence (MOS) and considers the proportion of optimal points as the ranking basis.

Optimal Point:

One result is considered optimal if **there is no other point or line above it at an identical latency**. In this case, the result is of the highest translation quality at that latency and we define it as an Optimal Point.



Results of submissions



(a). Zh-En Text-to-text ST track

(b). Zh-En Speech-to-text ST track

	Text-to-text	Speech-to-text
BIT-Xiaomi	48.17	31.26
Huawei	46.49	37.46

Table 5: The highest BLEU scores achieved by BIT-Xiaomi and Huawei for the same testset with different input modalities. The Speech-to-text track inputs audios while the Text-to-text track inputs golden transcription.

BLEU gap between the two input modalities: 16.91 & 9.03, respectively.

Discussion on the evaluation metrics

Quality Estimation:

Whether BLEURT/BERTScore is more suitable than BLEU in simultaneous translation scenario?

	Metrics	$r(\uparrow)$	$\rho(\uparrow)$	$\tau(\uparrow)$
SYS1	SentBLEU	0.546	0.484	0.390
	BERTScore	0.553	0.484	0.388
	BLEURT	**0.708**	**0.655**	**0.537**
SYS2	SentBLEU	0.584	0.516	0.415
	BERTScore	0.587	0.540	0.433
	BLEURT	**0.729**	**0.693**	**0.568**
SYS3	SentBLEU	0.525	0.468	0.374
	BERTScore	0.529	0.498	0.396
	BLEURT	**0.670**	**0.654**	**0.532**
SYS4	SentBLEU	0.467	0.408	0.322
	BERTScore	0.135[6]	0.467	0.368
	BLEURT	**0.637**	**0.629**	**0.507**
SYS5	SentBLEU	0.451	0.422	0.332
	BERTScore	0.518	0.522	0.414
	BLEURT	**0.656**	**0.672**	**0.539**
SYS6	SentBLEU	0.370	0.350	0.274
	BERTScore	0.475	0.480	0.376
	BLEURT	**0.559**	**0.578**	**0.459**

Table 7: Sentence-level agreement with human ratings on 6 ST systems. Given 6 source documents, each system (SYSi) performs ST, and the translation results are evaluated by sentenceBLEU (sentBLEU), BertScore, and BLEURT with 4 references. We calculate the Pearson correlation (r), the Spearman correlation (ρ), and the Kendall Tau (τ) score between the automatic metrics and human ratings. BLEURT has obvious advantages over the other two metrics in all the 6 systems.

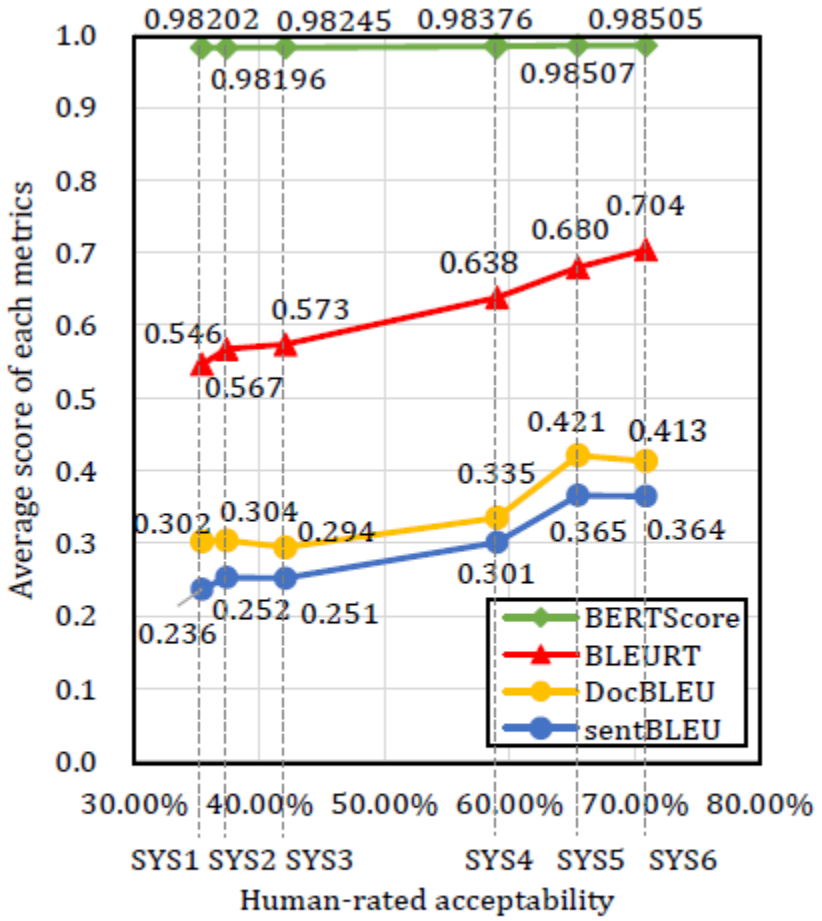


Figure 3: Human-rated acceptability vs. automatic metrics for the translation of 6 systems.

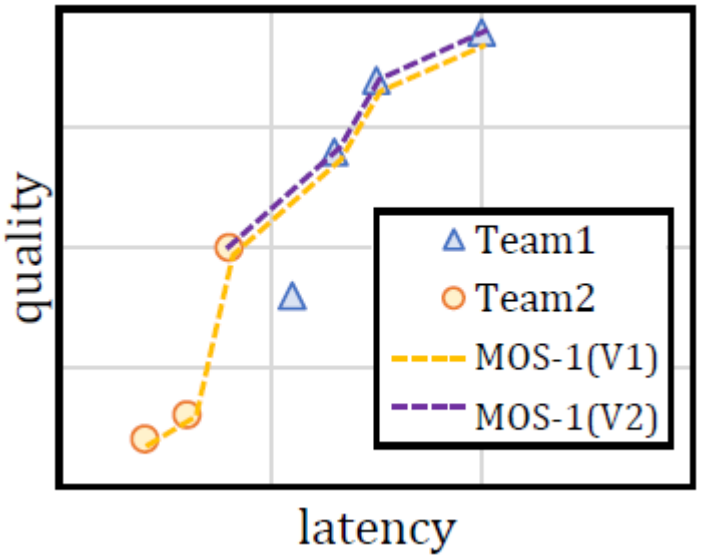
Discussion on the ranking algorithm

Ranking

The ranking problem of I-MOS algorithm:

The MOS curve is bound to select the leftmost point regardless of its translation quality, because the leftmost point is definitely an optimal point.

Therefore, I-MOS somehow encourages participants to submit only one point with extremely low latency, making the team ranked first place by I-MOS.



Modifications

1. We require each team to submit at least two points with different delays to make a latency-quality trade-off.
2. Before running the I-MOS algorithm, we first scan to remove the leftmost points whose quality is worse than others' submissions. If all submission points of a team are removed, the team will be ranked last.

Thank you!