



Overview of the IWSLT 2009 Evaluation Campaign

Michael Paul

National Institute of Information and Communications Technology
Kyoto, Japan

Outline of Talk

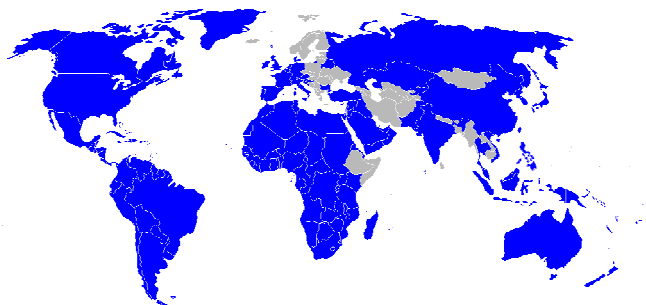
1. Evaluation Campaign:

- Participants
- What's New?
- Language Resources
- Challenge Task 2009
- Evaluation Specifications

2. Evaluation Results:

- Automatic Evaluation
- Subjective Evaluation
- Correlation between Evaluation Metrics
- Innovative Ideas explored by Participants

IWSLT 2009 Participants



ES: 2



SG: 2



FR: 3



US: 2



IE: 1



TR: 2



IT: 1



ZH: 2



JP: 3

Teams: 18
Engines: 35

Research Group		System
TR	AppTek, Inc.	apptek *
ES	Barcelona Media	bmrc *
IE	Dublin City University	dcu
IT	Fondazione Bruno Kessler	fbk
FR	University of Caen Basse-Normandie	greyc
SG	Institute for Infocomm Research	i2r
ZH	Chinese Academy of Science, ICT	ict
FR	University J. Fourier, LIG	lig
FR	University of Le Mans, LIUM	lium
US	MIT Lincoln Lab / Air Force Research Lab	mit
JP	NICT	nict
ZH	Chinese Academy of Science, NLPR	nlpr
SG	National University of Singapore	nus *
JP	University of Tokyo	tokyo *
JP	Tottori University	tottori
TR	TÜBİTAK-UEKAE	tubitak
ES	University Politecnica de Valencia	upv *
US	University of Washington	uw

* first-time participation

What's New?

- Challenge Task
 - translation of **cross-lingual human-mediated dialogs** in a travel situation (SLDB data, **Chinese↔English**)
 - **context annotations** (dialog, speaker-role)
 - ASR output (lattices, N/1-BEST lists)
- BTEC Task
 - **only TEXT input** for all classic BTEC tasks (**Arabic/Chinese→English**)
 - **new input languages: Turkish →English**
- Single Data Track
 - **usage of supplied language resources only**
- Extended Training/Run Submission Period
 - 2 month for training, 2 weeks for submitting runs
- Evaluation
 - investigate **effects of dialog information on MT quality**

Language Resources

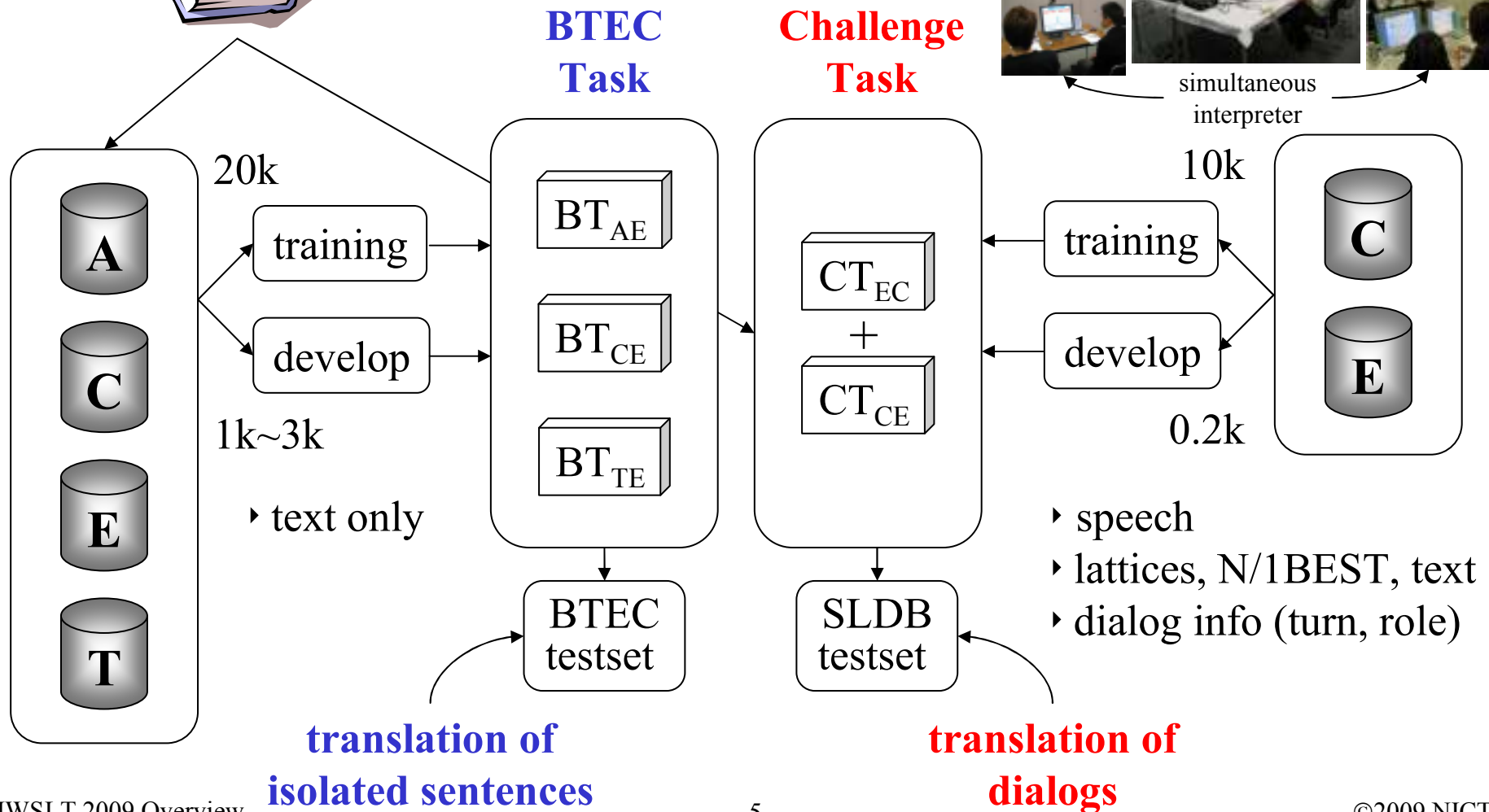
BTEC

(Basic Travel Expression Corpus)

SLDB

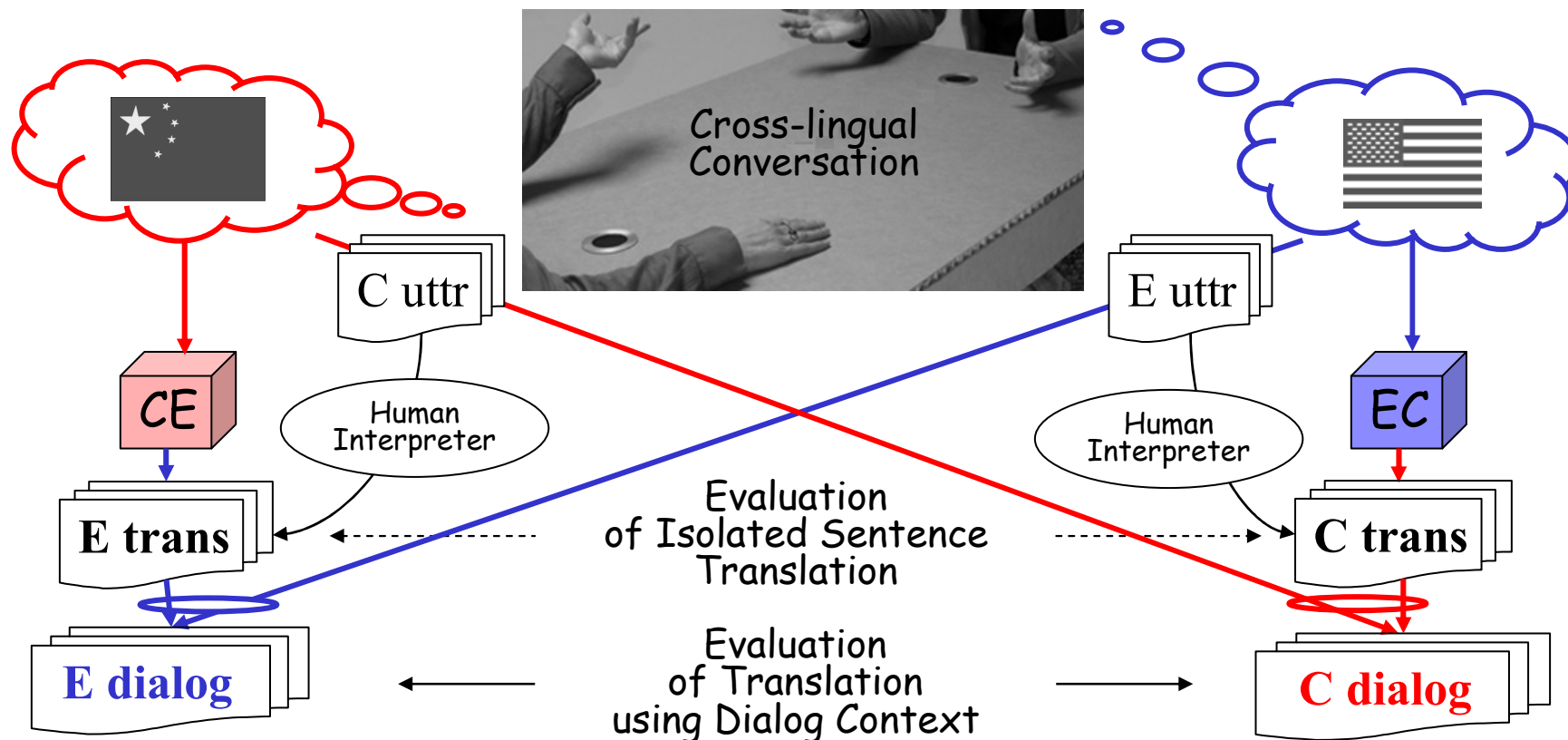
(Spoken Language Database)

Travel Domain



Challenge Task

- translation of human-mediated **cross-lingual conversations**
 - **task-oriented dialogs (role-play) in a travel situation**
 - *translation directions: C → E, E → C*



Challenge Task

- SLDB dialog data:

	English		Chinese	
Train	...	←	...	}
		→	...	
	...	←	...	
		→	...	
Dev	→	...	}
Test	←	...	

(text only)

(speech data)
(text data)

(train) 400 dialogs, ~10,000 sen
 (dev) 10 dialogs, ~400 sen
 (test) 27 dialogs, ~800 sen

Challenge Task

• **Dialog Example:**

(speaker) interjections uttered
(interpreter) *interjections skipped*

Agent:	Okay, no problem. And, will you be paying by cash or charge, sir? (interpreter) 好的。您用现金，还是用信用卡？
Customer:	嗯 我要用信用卡。 (interpreter) By <u>credit card</u> .
Agent:	Okay. Could I have your <u>number</u> in that case, please? (interpreter) 好的。那么，请告诉我 <u>信用卡号码</u> 。
Customer:	维萨卡。 (interpreter) It's a VISA card.
Customer:	号码，四九八零零四五九。 (interpreter) The number is four nine, eight, o, o four, five nine.
Customer:	九一九九五三一三。 (interpreter) Nine one, nine nine, five three, one three.
Agent:	Okay. Thank you. Uhmm and when does <u>it</u> expire? (interpreter) 知道了。 <u>信用卡</u> 什么时候到期？
Customer:	嗯 明年四月到期。 (interpreter) It <u>expires</u> in April, next year.

(speaker) number
(interpreter) *credit card number*

(speaker) anaphoric expression
(interpreter) *nominal antecedent*

(speaker) "ends at"
(interpreter) *context-specific word selection*

Statistics of Evaluation Data Sets

Track	Lang	Sen	Length	Word	Voc	Ref
CT _{EC}	E	393	11.0	4,329	570	—
	C		10.5	16,558	872	4
CT _{CE}	C	405	11.3	4,562	653	—
	E		11.5	18,594	764	4
BT _{CE}	C	469	5.5	1,808	877	—
	E		7.1	23,149	1,526	7

- **BTEC sentences are shorter** than CHALLENGE utterances
- **CHALLENGE vocabulary is smaller** than the BTEC vocabulary

Translation Task Complexity

Set	Lang	Entropy	Words	Total Entropy	Track
testset	C	6.18	4,142	25,580	CT _{EC}
	E	5.43	4,501	24,446	CT _{CE}
		5.80	2,844	15,063	BT _{CE} BT _{AE} BT _{TE}

- larger total entropy for CHALLENGE references
 → **CHALLENGE task** is supposed to be **more difficult than BTEC task**

Recognition Accuracy

Set	Lang	Word (%)		Sentence(%)		Track
		Lattice	1BEST	Lattice	1BEST	
testset	C	91.82	75.81	57.64	29.32	CT_{CE}
	E	89.58	82.20	50.13	37.15	CT_{EC}

- **large difference** in word recognition accuracy **for lattice vs. 1BEST** for Chinese utterances, but smaller for English
- **even larger difference** in recognition accuracies **on the sentence-level** for both, Chinese and English

→ **decoding of lattices** (or at least NBEST) has potential to **produce translations of better quality**

Evaluation Specifications

Automatic Evaluation: → all primary run submissions

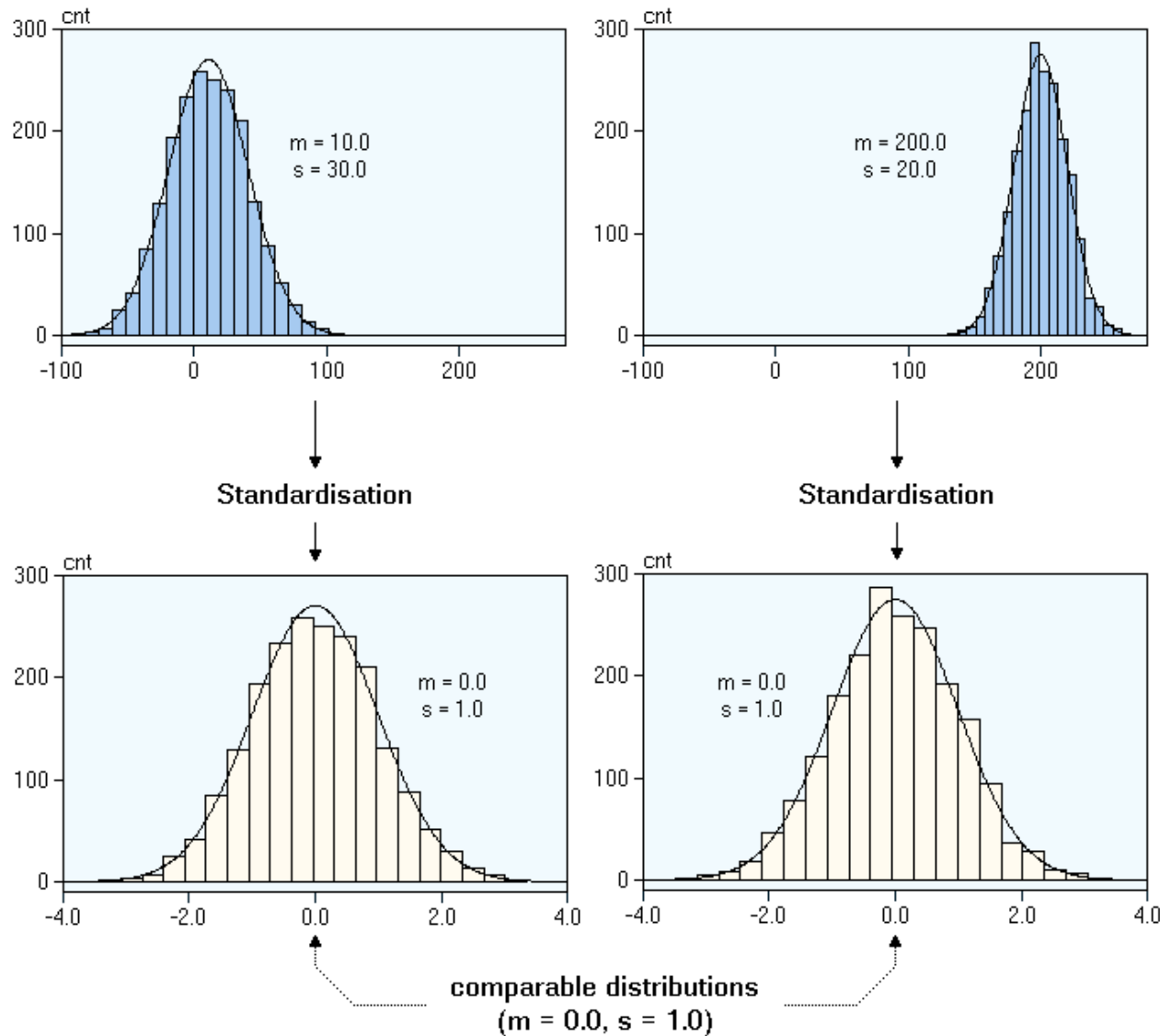
- **case-sensitive, with punctuation marks** (*case+punc*)
- case-insensitive, without punctuation marks (*no_case+no_punc*)
- 7 standard metrics:
 - + BLEU + NIST + WER +TER
 - + METEOR (f1) + GTM + PER

Evaluation Specifications

Significance Test:

- (1) perform a **random sampling with replacement** from the evaluation testset
- (2) **calculate** respective evaluation metric scores for each MT engine and the **differences between the two MT engine scores**
- (3) **repeat sampling/scoring steps** iteratively (2000 iterations)
- (4) **apply Student's t-test at a significant level of 95%** to test whether score differences are significant

Metric Score Combination



Metric Score Combination

Z-Transform:

- standardize a distribution so that:
 - + it has a **zero mean** ($\mu = 0$)
 - + it has **unit variance** ($\sigma^2 = 1$)

$$z_i = \frac{(x_i - \mu)}{\sigma}$$

$\{x_i\}$: a set of n sample values from score distribution

μ : mean of sample values

σ : standard deviation

σ^2 : variance of the distribution

Evaluation Specifications

Automatic Evaluation: → all primary run submissions

- **case-sensitive, with punctuation marks** (*case+punc*)
- case-insensitive, without punctuation marks (*no_case+no_punc*)

◦ 7 standard metrics:

+ BLEU	+ NIST	+ WER	+ TER
+ METEOR (f1)	+ GTM	+ PER	

◦ combine multiple metric scores (z-avg):

+ normalize single-metric scores so that score distribution has a zero mean and unit variance → *z-score*

+ for each MT system, calculate **z-avg** as the average of all obtained metric *z-scores*

◦ for each translation task, **order MT systems according to z-avg**

Evaluation Specifications

Human Assessment:

- **Ranking** (grades 4 – 0) → all primary run submissions
 - + rank each whole sentence translation from Best to Worst relative to the other choices (ties are allowed)

- **Fluency/Adequacy** (grades 4 – 0) → top-ranked MT engine
 - + *Fluency* indicates how the translation sounds to a native speaker
 - + *Adequacy* judges how much reference information is expressed in the translation

- **Dialog Adequacy** (grades 4 – 0) → top-ranked MT engine
 - + an adequacy evaluation that takes into account the context of the respective dialog
 - + omitted information in translation that is understood in the dialog context should not result in a lower *dialog adequacy* grade

Outline of Talk

1. Evaluation Campaign:

- Participants
- What's New?
- Language Resources
- Challenge Task 2009
- Evaluation Specifications

2. Evaluation Results:

- Automatic Evaluation
- Subjective Evaluation
- Correlation between Evaluation Metrics
- Innovative Ideas explored by Participants

Data Track Participation

Task	Translation Direction	Team	Run		
			primary	contrastive	
Challenge	English-Chinese	CT _{EC}	7	6	14
	Chinese-English	CT _{CE}	7	6	12
BTEC	Arabic-English	BT _{AE}	9	9	9
	Chinese-English	BT _{CE}	12	12	19
	Turkish-English	BT _{TE}	7	7	15
Total			18	40	69

Automatic Evaluation

z-avg (BLEU, METEOR, 1-WER, 1-PER, 1-TER, GTM, NIST)

CT_{EC}

nlpr_ASR.5	1.364
nict_ASR.1	1.017
fbk_ASR.1	0.881
dcu_ASR.1	0.659
ict_ASR.20	0.194
tottori_ASR.1	-1.544

nlpr_CRR	1.249
fbk_CRR	1.026
ict_CRR	0.748
nict_CRR	0.613
dcu_CRR	0.570
tottori_CRR	-1.634

input

ASR



lattice, N/1BEST

CRR



correct recognition
result

CT_{CE}

nlpr_ASR.5	2.063
dcu_ASR.1	0.704
fbk_ASR.1	0.530
ict_ASR.20	0.405
nict_ASR.1	-0.378
tottori_ASR.1	-0.751

nlpr_CRR	1.851
dcu_CRR	1.214
fbk_CRR	0.398
ict_CRR	0.326
nict_CRR	-0.424
tottori_CRR	-0.793

Automatic Evaluation

z-avg (BLEU, METEOR, 1-WER, 1-PER, 1-TER, GTM, NIST)

BT _{AE}	
mit+tub	1.504
mit	1.432
fbk	0.940
tubitak	0.504
lium	0.465
bmrc	0.456
lig	0.325
uw	0.137
greyc	-1.941

BT _{CE}	
nlpr	2.178
nus	1.344
i2r	1.250
uw	0.786
dcu	0.545
bmrc	0.489
lium	0.323
upv	0.068
tokyo	0.022
ict	-0.011
tottori	-0.405
greyc	-1.444

BT _{TE}	
mit+tub	1.304
mit	1.216
tubitak	1.043
fbk	0.912
dcu	0.502
apptek	-0.536
greyc	-1.441

Ranking

CT_{EC}

nlpr_ASR.5	3.48
nict_ASR.1	3.02
dcu_ASR.1	2.80
fbk_ASR.1	2.79
ict_ASR.20	2.63
tottori_ASR.1	2.18

nlpr_CRR	3.84
ict_CRR	3.67
nict_CRR	3.42
fbk_CRR	3.32
dcu_CRR	3.31
tottori_CRR	2.58

NormRank



normalized ranks
on a per-judge basis
[Blatz et.al. 2003]


MT systems marked
in **blue** were ranked
differently by
automatic metrics

CT_{CE}

nlpr_ASR.5	3.52
ict_ASR.1	2.90
dcu_ASR.1	2.84
nict_ASR.20	2.80
fbk_ASR.1	2.75
tottori_ASR.1	2.60

nlpr_CRR	3.67
dcu_CRR	3.32
fbk_CRR	3.26
ict_CRR	3.20
nict_CRR	3.11
tottori_CRR	2.83

Ranking

NormRank 0  4

BT _{AE}	
mit	3.29
mit+tub	3.28
fbk	3.03
tubitak	3.03
lium	3.01
bmrc	2.95
lig	2.87
uw	2.86
greyc	2.38

BT _{CE}	
nlpr	3.55
nus	3.24
i2r	3.17
uw	3.12
dcu	3.01
bmrc	2.99
lium	2.95
upv	2.91
tokyo	2.87
ict	2.84
tottori	2.78
greyc	2.63

BT _{TE}	
mit+tub	3.26
mit	3.25
tubitak	3.23
fbk	3.13
dcu	2.92
apptek	2.74
greyc	2.39

Best Rank Difference

- use the MT system with highest ranking score as a point-of-reference
- rank systems according to difference in rank against the best system

◦ *metric*: gain $\left(\frac{\text{better} - \text{worse}}{\text{graded}}\right)$ of the top MT towards any other system in %



CT_{EC}		<i>better</i>	<i>same</i>	<i>worse</i>
nlpr_AS.R.5		—	—	—
nict_AS.R.1	29.85	50.24	29.37	20.39
fbk_AS.R.1	36.72	53.09	30.54	16.37
dcu_AS.R.1	36.96	51.70	33.56	14.74
ict_AS.R.20	48.64	61.26	26.12	12.62
tottori_AS.R.1	61.65	70.42	20.81	8.77

Correlation between Automatic Evaluation and Ranking

◦ Spearman's rank correlation coefficient $\rho \in \{-1.0, 1.0\}$

task	metric	z-avg
------	--------	-------

CT_{EC} (ASR:6)	NormRank	0.9429
	BestRankDiff	0.8857

CT_{EC} (CRR:6)	NormRank	0.8286
	BestRankDiff	0.7143

CT_{CE} (ASR:6)	NormRank	0.7143
	BestRankDiff	0.6000

CT_{CE} (CRR:6)	NormRank	0.7143
	BestRankDiff	0.6000

task	metric	z-avg
------	--------	-------

BT_{CE} (12)	NormRank	-0.3846
	BestRankDiff	0.2098

BT_{AE} (9)	NormRank	0.0333
	BestRankDiff	0.1667

BT_{TE} (7)	NormRank	0.8571
	BestRankDiff	-0.6071

Correlation between Automatic Evaluation and Ranking

- combination of all investigated automatic metrics optimal?

task	NormRank	BestRankDiff
$CT_{EC}^{ASR} (6)$	(all)	TER
$CT_{EC}^{CRR} (6)$	f1+TER	TER
$CT_{CE}^{ASR} (6)$	METEOR	METEOR
$CT_{CE}^{CRR} (6)$	(all)	GTM
$BT_{CE} (12)$	BLEU	TER
$BT_{AE} (9)$	METEOR	PER
$BT_{TE} (7)$	NIST	TER

Correlation between Automatic Evaluation and Ranking

- **effects of combination of multiple metrics:**
 - better correlation for CT using *NormRank*
 - single metrics perform best for *BestRankDiff*
 - METEOR and TER work best for most translation tasks
 - BLEU best for BT_{CE} , but low correlation for all other tasks
 - **correlation depends on:**
 - selected evaluation metrics (subjective, automatic)
 - number of MT systems to be ranked
 - translation quality of respective MT system outputs
- **simply averaging metric scores might not be the best solution**
to combine multiple automatic evaluation metrics

Fluency/Adequacy/Dialog

median grade of 3 human grades

fluency

4	Flawless English
3	Good English
2	Non-native English
1	Disfluent English
0	Incomprehensible



dialog / adequacy

4	All Information
3	Most Information
2	Much Information
1	Little Information
0	None

CT_{EC}	CT_{CE}	BT_{CE}	BT_{AE}	BT_{TE}
------------------------	------------------------	------------------------	------------------------	------------------------

fluency	ASR: 2.35 CRR: 2.60	ASR: 2.37 CRR: 2.53	2.78	2.70	2.90
----------------	--------------------------------------	--------------------------------------	-------------	-------------	-------------

- **translation quality of translation tasks:**
 - fluency : $BT_{TE} > BT_{CE} > BT_{AE} > CT_{EC} > CT_{CE}$
 - adequacy: $BT_{TE} > BT_{CE} > CT_{CE} > CT_{EC} > BT_{AE}$
 - dialog adequacy: $CT_{CE} > CT_{EC}$
- **effects of dialog information on translation quality:**
 - CT_{CE} / CT_{EC} : dialog adequacy > adequacy
 - larger difference for CT_{CE}
- **dialog context helps** humans to understand MT outputs
- **sentence-by-sentence evaluation not sufficient** for spoken language translation technologies
- **develop new MT algorithm and evaluation metrics** capable of taking into account information beyond the current sentence

Innovative Ideas Explored by Participants

- morphological **preprocessing** techniques
- **statistical modeling techniques** integrating syntactic and source language information
- cross-domain **model adaptation**
- new **parameter optimization** techniques
- **lattice decoding**
- semi-supervised **reranking methods** of NBEST lists
- improved **system combinations** using hybrid MT engines

Acknowledgements

- *data preparation*
 - NICT team
 - TUBITAK team
- *human assessment*
 - FBK (English)
 - LIG (English)
 - AppTek (English)
 - UW (Chinese)
 - NICT (English, Chinese)
- *automatic evaluation software*
 - JHU: Chris Callison-Burch
 - NICT: Tatsufumi Shimizu
- *technical paper*
 - FBK team
- *local organization*
 - NICT team

Thank you!