Overview of the IWSLT 2008 Evaluation Campaign

Michael Paul

National Institute of Information and Communications Technology
ATR Spoken Language Communication Research Laboratories
Kyoto, Japan

Outline of Talk

1. Evaluation Campaign:

- schedule
- participants
- language resources
- evaluation specifications

2. Evaluation Results:

- automatic evaluation
- subjective evaluation
- correlation between evaluation metrics
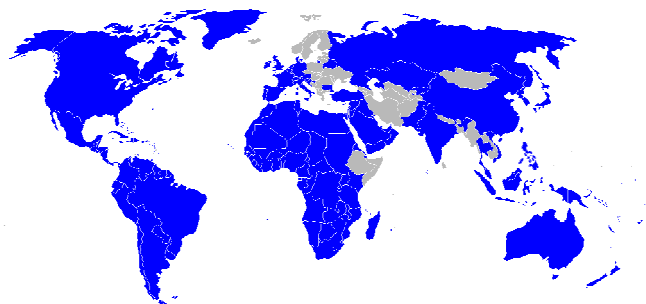- grader consistency

3. Discussions:

- Challenge Task 2008
- innovative idea's explored by participants

Evaluation Campaign Schedule

Event	Date
MT system registration	**Apr 04, 2008**
release of training/development sets	**Jun 02, 2008**
release of test sets	**Jun 30, 2008**
run submission	**Jul 04, 2008**
result feedback (automatic metric scores)	**Jul 11, 2008**
system descriptions	**Jul 18, 2008**
evaluation server online	**Jul 24, 2008**
result feedback (subjective evaluation scores)	**Aug 22, 2008**
camera-ready paper	**Sep 05, 2008**
IWSLT 2008 workshop	**Oct 20-21, 2008**

IWSLT 2008 Participants

	Research Group	System
US	Carnegie Mellon University, InterACT	cmu
IE	Dublin City University	dcu
IT	Fondazione Bruno Kessler	fbk
FR	University of Caen Basse-Normandie	greyc
SG	Insititute for Infocomm Research	i2r
ZH	Chinese Academy of Science, ICT	ict
FR	University J. Fourier, GETALP/LIG	lig
FR	University of Le Mans, LIUM	lium *
US	MIT Lincoln Lab / Air Force Research Lab	mitll
JP	NICT-ATR	nict
ZH	Chinese Academy of Science, NLPR	nlpr
JP	NTT Communication Science Laboratory	ntt
KR	Pohang University of Science & Technology	postech *
GB	Queen Mary University of London	qmul *
DE	Rheinisch Westfälische Hochschule	rwth
ES	TALP-UPC Research Center	talp
ZH	Toshiba China R&D Center	tch *
JP	Tottori University	tottori
TR	TÜBÌTAK-UEKAE	tubitak

DE: 1   JP: 3

ES: 1   KR: 1

FR: 3   SG: 1

GB: 1   US: 2

IE: 1   TR: 1

IT: 1   ZH: 3

Teams: 19
Engines: 58

* first-time participation

Data Track Participation

Task	Translation Direction		Team	Run	
				primary	contrastive
Challenge	English-Chinese	CT_{EC}	**7**	7	11
	Chinese-English	CT_{CE}	**11**	11	32
BTEC	Arabic-English	BT_{AE}	**10**	10	13
	Chinese-English	BT_{CE}	**14**	14	29
	Chinese-Spanish	BT_{CS}	**8**	8	5
Pivot	Chinese-(English)-Spanish	PV_{CS}	**8**	8	11
Total			**19**	**58**	**101**

NICT
**Language
Translation**

ATR
**Spoken Language
Communication**

BTEC → useful sentences together with the translation into other languages usually found in phrasebooks for tourists traveling abroad
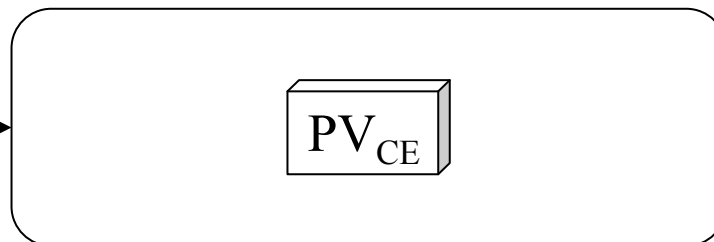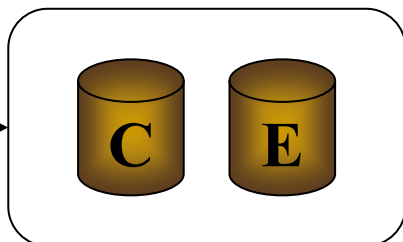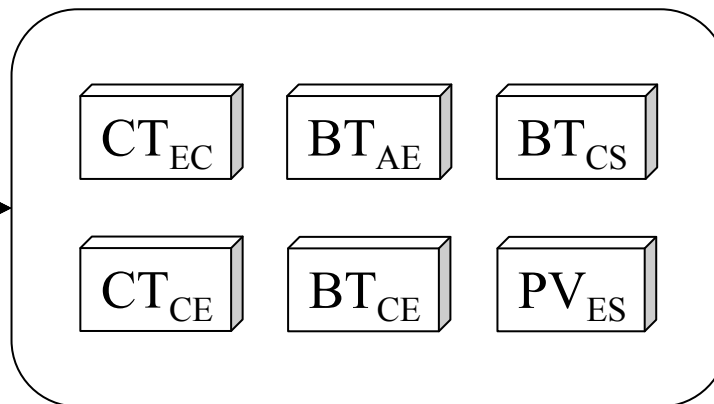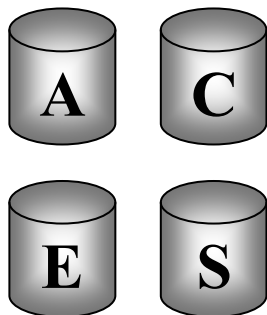
Challenge Task

Shared Resources

devset$_{1-7}$

devset

Corpus

A | C

E | S

20k

MT Engines

| CT_{EC} | BT_{AE} | BT_{CS} |

| CT_{CE} | BT_{CE} | PV_{ES} |

C | E

20k

PV_{CE}

Challenge Task 2008

- **machine-mediated conversational speech in a real situation**
  - ° **traveler speech with S2S system**
  - ° *translation directions:* English ↔ Japanese
    Chinese ↔ Japanese

- **field experiments in Kyoto/Japan**
  - ° foreign travelers carried out specific tourism-related tasks
    (*buying souvenirs* or *entrance tickets*, *asking for directions*, etc.)
  - ° communication with local staff
    using hand-held translation device
  - ° speech data of 50 English and
    50 Chinese travelers
  - ° 3-4 tasks per traveler
  - ° recordings at 5 different locations

Challenge Task 2008

S2S-mediated conversation

Participants

run submissions

speech, transcripts — ASR → lattice N/1BEST — release →

CE → E trans

EC → C trans

human translation → reference translation — evaluation →
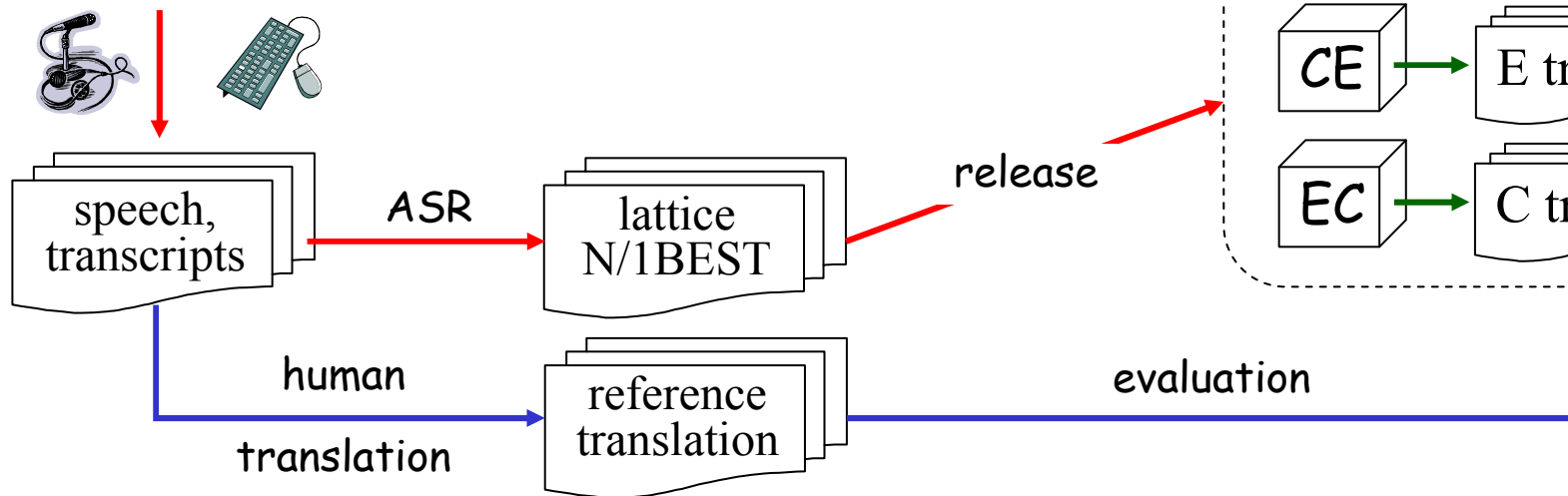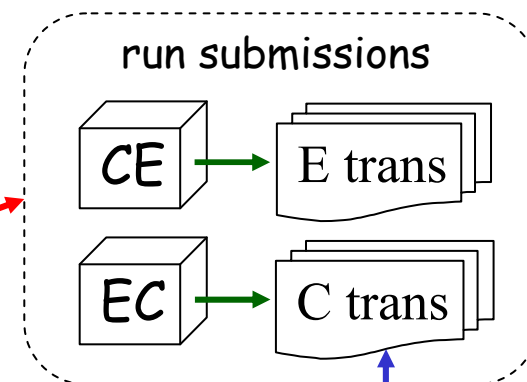
(dev) 250 x ASR output
(test) 500 x ASR output

Challenge Task Evaluation Data

Track	Lang	Sen	Length	Word	Voc	Ref
CT_{EC}	**E**	498	**5.8**	**2,867**	**312**	—
	C		5.7	20,016	708	7
CT_{CE}	**C**	504	**5.0**	**2,513**	**385**	—
	E		6.2	21,751	810	7

- comparison of Chinese CT_{CE} (English CT_{EC}) *user utterances* and Chinese CT_{EC} (English CT_{CE}) *human reference translations*
  - user utterances are **shorter**
  - user utterances have **limited vocabulary**
- Chinese sentences are shorter than English

BTEC/Pivot Task Evaluation Data

Track	Lang	Sen	Length	Word	Voc	Ref
BT$_{CE}$	C	507	5.5	2,808	885	—
	E		6.8	55,082	2,146	16
BT$_{AE}$	**A**	507	**5.1**	**2,585**	**1,205**	**—**
	E		6.8	55,082	2,146	16
BT$_{CS}$ PV$_{CS}$	C	507	5.5	2,808	885	—
	S		**6.2**	**50,169**	**2,569**	**16**

- shortest utterances with large vocabulary for Arabic ($\rightarrow$ word segmentation)

- Spanish references shorter than English ones with larger vocabulary ($\rightarrow$ word inflections)

Translation Task Complexity

Set	Lang	Entropy	Words	Total Entropy	Track
testset	C	9.51	3,962	35,111	CT$_{EC}$
	E	10.1	3,662	**36,986**	**CT$_{CE}$**
		9.83	4,057	**39,880**	**BT$_{CE}$** BT$_{AE}$
	S	10.25	3,885	39,821	BT$_{CS}$ PV$_{CS}$

- lower total entropy for English Challenge Task references
  - → **Challenge Task is supposed to be easier than BTEC Task**

Recognition Accuracy

Set	Lang	Word (%)		Sentence(%)		Track
		Lattice	1BEST	Lattice	1BEST	
testset (spont)	C	95.07	85.79	79.56	**53.77**	**CT$_{CE}$**
	E	87.27	79.77	65.06	53.01	CT$_{EC}$
testset (read)	A	—	72.80	—	36.10	BT$_{AE}$
	C	94.20	83.61	80.47	**63.31**	**BT$_{CE}$** BT$_{CS}$ PV$_{CS}$

- similar word accuracies for Challenge Task and BTEC Tasks

- 10% difference of 1BEST recognition accuracies on sentence-level

NICT
Language
Translation

ATR
Spoken Language
Communication

Evaluation Specifications

Automatic Evaluation:

° *metric*: $\underline{\textbf{(BLEU+METEOR)}}$ → all primary run submissions
  $\textbf{2}$

° **case-sensitive, with punctuation marks** (*official*)

° case-insensitive, without punctuation marks (*additional*)

Human Assessment:

° *metric*: **Fluency/Adequacy** → 4 MT engines per data track

° *metric*: **Ranking** → all primary run submissions

Human Evaluators:

° *method*: **Intra-Grader-Consistency**

° *method*: **Inter-Grader-Consistency**

target language	human evaluator
Chinese	7
English	6
Spanish	11

Outline of Talk

1. Evaluation Campaign:

- schedule
- participants
- language resources
- evaluation specifications

2. Evaluation Results:

- automatic evaluation
- subjective evaluation
- correlation between evaluation metrics
- grader consistency

3. Discussions:

- Challenge Task 2008
- innovative idea's explored by participants

Automatic Evaluation

metrics: $\dfrac{\text{(BLEU+METEOR)}}{2}$

BLEU 0 bad good 1

METEOR 0 bad good 1

° *official evaluation specifications* (case-sensitive, with punctuations)

BT$_{AE}$	
mitll	**0.4292**
rwth	0.4020
lium	0.3952
talp	0.3915
tubitak	0.3904
lig	0.3881
dcu	0.3798
postech	0.3360
qmul	0.3126
greyc	0.2467

BT$_{CS}$	
tch	**0.3109**
fbk	0.2619
tubitak	0.2594
dcu	0.2521
nict	0.2500
talp	0.2359
postech	0.2223
greyc	0.2021

PV$_{CS}$	
tch	**0.3452**
fbk	0.3212
talp	0.3144
nict	0.3103
tubitak	0.2863
dcu	0.2844
greyc	0.1755
qmul	0.1469

Automatic Evaluation

CT_{EC}	
tch	**0.6173**
ict	0.5749
nlpr	0.5708
dcu	0.5599
nict	0.5366
tottori	0.5166
mitll	0.4930

CT_{CE}	
nlpr	**0.5061**
tch	**0.5060**
i2r	0.4882
ict	0.4681
rwth	0.4502
mitll	0.4396
ntt	0.4341
dcu	0.4173
nict	0.3748
fbk	0.3580
tottori	0.3406

BT_{CE}	
tch	**0.5347**
ict	0.5226
cmu	0.5219
nlpr	0.5148
i2r	0.5130
rwth	0.4873
fbk	0.4639
dcu	0.4408
tubitak	0.4295
nict	0.4180
tottori	0.3901
postech	0.3656
greyc	0.3160
qmul	0.2638

Fluency/Adequacy

° *metrics*: **median grade of 3 human grades**

fluency, adequacy

0 ⟨ bad   good ⟩ **4**
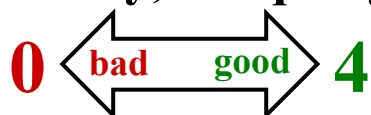
fluency

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

adequacy

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

	CT_{EC}	CT_{CE}	BT_{CE}	BT_{AE}	BT_{CS}	PV_{CS}
fluency	2.75 ~ 2.05	3.27 ~ 2.55	3.21 ~ 2.35	3.15 ~ 1.90	2.50 ~ 1.23	2.48 ~ 2.03
adequacy	2.39 ~ 1.66	2.36 ~ 1.70	2.46 ~ 1.76	2.18 ~ 1.45	1.87 ~ 1.22	1.81 ~ 1.41

- translation quality:
  - ° fluency : $CT_{CE} > BT_{CE} > BT_{AE} > CT_{EC} > BT_{CS} > PV_{CS}$
  - ° adequacy: $BT_{CE} > CT_{CE} \approx CT_{EC} > BT_{AE} > BT_{CS} > PV_{CS}$

Ranking

"rank each whole sentence translation from Best to Worst relative to the other choices (ties are allowed)" [Callison-Burch 07]

° graders compared up to 5 MT system outputs per page and assigned a grade of "1" to "5" to each system output

° *metrics*:
    average number of times that a system was judged better than any other systems

Ranking

0 <bad   good> 1

	CT_{EC}	CT_{CE}	BT_{CE}	BT_{AE}	BT_{CS}	PV_{CS}
ranking	0.3966	0.5274	0.5255	0.4415	0.4773	0.4932
system	tch	nlpr	nlpr	mitll	tch	tch

Paired Comparison

NICT
Language
Translation

ATR
Spoken Language
Communication

"paired comparison evaluation based on the *Ranking* results"

∘ *metrics*: gain ($\frac{better - worse}{graded}$) of MT_1 towards MT_2 in %

BT$_{CE}$

$MT_2 \rightarrow$

$MT_1 \searrow$

	tch	ict	i2r	rwth	cmu	dcu	fbk	⋯
nlpr	5.3	9.12	13.09	19.10	12.93	36.60	39.93	
tch		1.40	16.48	21.48	**11.11**	30.86	30.34	
ict			**-3.52**	9.77	17.25	25.38	28.09	
i2r				8.90	16.23	22.63	22.92	
rwth					**-3.51**	12.00	**11.90**	
cmu						18.03	**10.34**	
dcu							**-2.78**	
fbk								

- systems ordered according to *Ranking score*
- negative gains for certain system combinations

→ **inconsistency of *Ranking* metrics?**

Best Rank Difference

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

○ *metrics*: gain ($\frac{better - worse}{graded}$) of the top MT towards any other system in %

BestRankDiff

0 ⟷ good bad ⟶ 1

Ranking	BT$_{CE}$		worse	same	better
nlpr	nlpr	0.00	—	—	—
tch	tch	5.28	27.81	39.10	33.09
ict	ict	9.12	26.31	38.26	35.43
i2r	**cmu**	**12.92**	25.17	36.74	38.09
rwth	i2r	13.04	22.82	41.32	35.86
cmu	**rwth**	**19.10**	22.47	35.96	41.57
dcu	fbk	36.60	15.47	32.46	52.07
fbk	dcu	40.14	11.31	37.24	51.45

Correlation between Automatic Evaluation and Ranking

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

$$\rho = -1 \quad \rightarrow \text{perfect inverse correlation}$$
$$\rho = 0 \quad \rightarrow \text{no correlation}$$
$$\rho = +1 \quad \rightarrow \text{perfect correlation}$$

task	metric	(B+M)/2
CT_{EC} (7)	Ranking	0.1071
	BestRankDiff	0.1071
BT_{CS} (8)	Ranking	0.1190
	BestRankDiff	-0.1667
PV_{CS} (8)	Ranking	0.0238
	BestRankDiff	**0.6190**

task	metric	(B+M)/2
CT_{CE} (11)	Ranking	0.4000
	BestRankDiff	**0.8727**
BT_{CE} (14)	Ranking	0.3736
	BestRankDiff	**0.5769**
BT_{AE} (10)	Ranking	-0.0667
	BestRankDiff	**0.9152**

Grader Consistency

NICT
Language
Translation

ATR
Spoken Language
Communication

agreement: $\kappa < 0$ ~0.2   ~0.4   ~0.6   ~0.8   ~1.0
         none   slight   fair   moderate   substantial   perfect

Intra-Grader	κ Coefficient					
	CT_{EC}	CT_{CE}	BT_{CE}	BT_{AE}	BT_{CS}	PV_{CS}
fluency	0.64	0.71	**0.75**	0.71	0.52	0.54
adequacy	0.74	0.68	**0.81**	0.61	0.67	0.70
ranking	**0.73**	0.56	0.69	0.56	0.52	0.56

- mainly substantial agreement for *adequacy* and *fluency*
- mainly moderate agreement for *ranking*

→ intra-grader agreement: English > Chinese > Spanish

Grader Consistency

agreement: $\kappa<0$ ~0.2 ~0.4 ~0.6 ~0.8 ~1.0

           none   slight   fair   moderate   substantial   perfect

Inter-Grader	κ Coefficient					
	CT_{EC}	CT_{CE}	BT_{CE}	BT_{AE}	BT_{CS}	PV_{CS}
fluency	0.41	0.38	0.41	**0.44**	**0.19**	**0.25**
adequacy	0.40	0.41	0.46	**0.47**	**0.26**	**0.30**
ranking	**0.57**	0.52	0.56	0.54	0.47	0.51

- mainly moderate agreement between different graders
- best agreement for *ranking*
- only **slight~fair agreement** for Spanish

$\rightarrow$ inter-grader agreement: Chinese = English >> Spanish

Outline of Talk

1. Evaluation Campaign:
- schedule
- participants
- language resources
- evaluation specifications

2. Evaluation Results:
- automatic evaluation
- subjective evaluation
- correlation between evaluation metrics
- grader consistency

3. Discussions:
- Challenge Task 2008
- innovative idea's explored by participants

Challenge Task 2008

° machine-mediated spontaneous speech collected using inexperienced users in a real situation

   → **much easier than Challenge Task 2006**
      (spontaneous Chinese utterances of simulated Q&A scenarios)

° comparison to BTEC Task:

   · **CT$_{CE}$ sentences shorter and less complex** than BT$_{CE}$ sentences

      … but: **ASR output translation quality lower**

   · **CT$_{CE}$ and BT$_{CE}$ recognition performance similar**
      for lattice input (word: 95%, sentence: 80%) and
      1BEST word accuracy (84%)

      … but: **1BEST sentence-level accuracy 10% worse**

   → **most participants used only 1BEST** ☹

Innovative Ideas
Explored by Participants

° additional (synthetic) training resources by translating
  in-domain monolingual resources

° advanced techniques for phrase-extraction from NBEST
  alignments

° improved statistical modeling techniques

° new rescoring/reranking methods of NBEST lists

° improved system combinations using hybrid MT engines

Acknowledgements

- *data preparation*
  - ° NICT/ATR: Gen Itoh, Shigeki Matsuda
  - ° CMU: Mark Fuhs

- *automatic evaluation software*
  - ° JHU: Chris Callison-Burch
  - ° CMU: Matthias Eck

THANK YOU !

- *human assessment*
  - ° TOSHIBA team (Chinese)
  - ° UPC-TALP team (Spanish)
  - ° FBK team (English)
  - ° NICT/ATR team (English, Chinese)

- *participation*
  - ° **all of you !**