Spoken Language Communication
Research Laboratories

Overview of the IWSLT 2006 Evaluation Campaign

Michael Paul

ATR Spoken Language Communication Research Laboratories
Kyoto, Japan

ATR
Spoken Language Communication
Research Laboratories

History of IWSLT

2000 … 2003 2004 2005 2006

	CSTAR03	IWSLT04	IWSLT05	IWSLT06
(translation)	▪ text	▪ text	▪ read speech	▪ spontaneous speech
(input data)	▪ text	▪ text	▪ ASR output	▪ speech input
(evaluation campaign)	▪ closed evaluation of MT engines	▪ feasibility of MT technologies, evaluation metrics	▪ feasibility of speech translation	▪ robustness of speech translation technologies
(language pairs)	▪ C (3) → E I (1) → E J (1) → E K (1) → E	▪ C (13) → E J (6) → E	▪ A (8) → E C (12) → E J (11) → E K (5) → E E (3) → C	▪ A (9) → E C (12) → E I (10) → E J (12) → E
(participants) (submissions)	▪ 6 participants, 6 runs	▪ 14 participants, 28 runs	▪ 19 participants, 84 runs	▪ 19 participants, 73 runs

Outline of Talk

1. Evaluation Campaign:

- data preparations
- translation input and data track conditions
- run submissions
- evaluation specifications

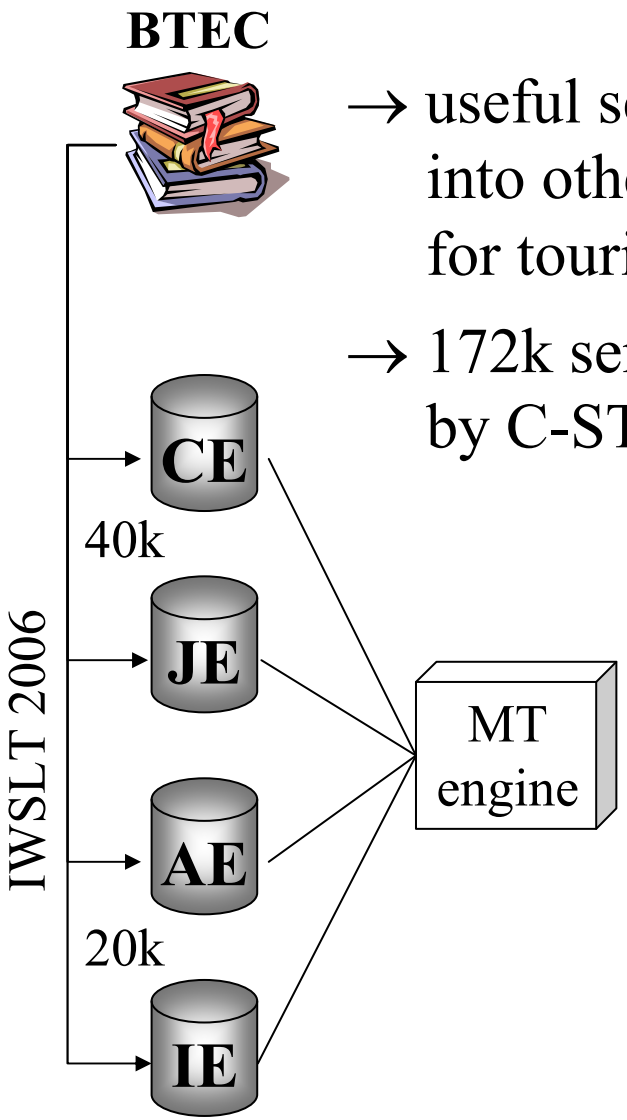
2. Evaluation Results:

- subjective/automatic evaluation
- correlation between evaluation metrics

3. Discussions:

- Challenge Task 2006
- source language effects
- robustness towards recognition errors
- innovative idea's explored by participants

Basic Travel Expression Corpus



→ useful sentences, together with the translation into other languages usually found in phrasebooks for tourists going abroad

→ 172k sentence pairs collected/translated by C-STAR partners (A,C,E,I,J,K)

J: フィルムを買いたいです。
E: I want to buy a roll of film.

J: 8人分予約したいです。
E: I 'd like to reserve a table for eight.

J: 友人が車にひかれ大けがを しました。
E: My friend was hit by a car and badly injured.

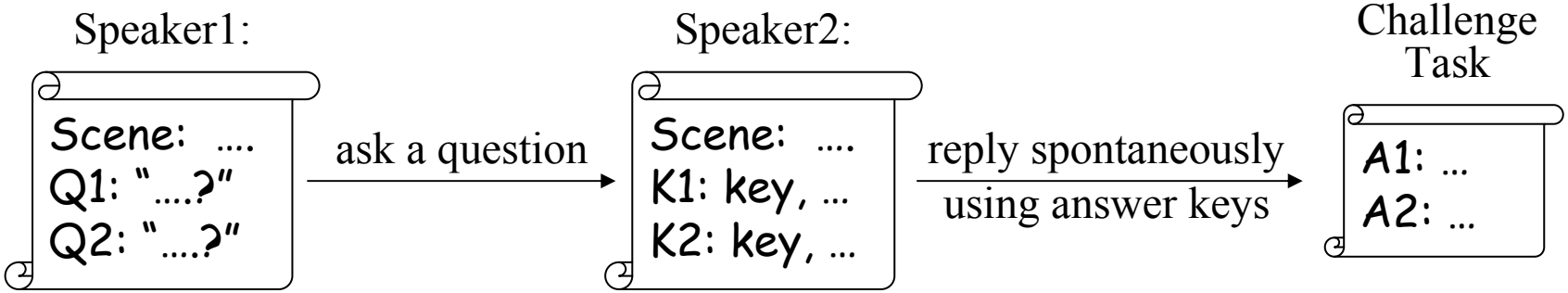
ATR
Spoken Language Communication
Research Laboratories

Supplied Resources

type	language	sentence count	word token	word type	words per sentence
training	C/E	40k	342k / 367k	11k / 7k	8.6 / 9.2
	J/E		398k / 367k	11k / 7k	10.0 / 9.2
	A/E	20k	154k / 183k	18k / 5k	7.7 / 9.2
	I/E		171k / 183k	10k / 5k	8.6 / 9.2
development (dev1, dev2, dev3)	C/E_{16}	1.5k	11k / 198k	2k / 3k	7.0 / 8.2
	J/E_{16}		12k / 198k	2k / 3k	8.2 / 8.2
	A/E_{16}		9k / 198k	3k / 3k	6.3 / 8.2
	I/E_{16}		10k / 198k	2k / 3k	6.8 / 8.2

Challenge Task 2006

- speech input with certain level of "spontaneity"
- **spontaneous answers to questions in tourism domain**



[airplane] passenger asks flight attendance for help

Q1: Okay. Where can I put my luggage? Is it here okay?

K1: *[not here], [overhead compartment]*

A1: **sorry you'd better put it in the overhead compartment**

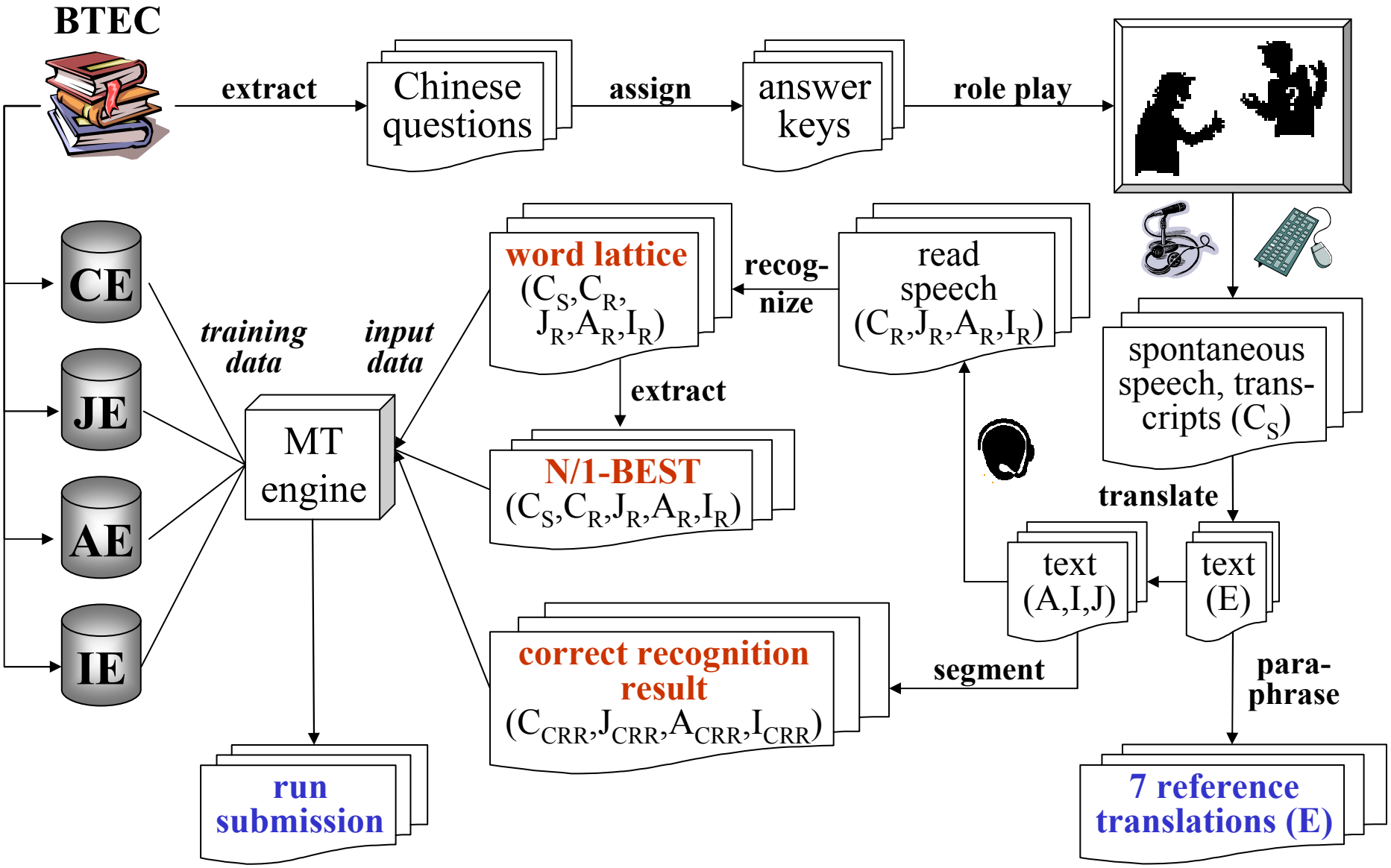
[airport] customer asks taxi driver for directions

Q2: Take me to this address. How long will it take?

K2: *[depending on traffic condition], [around 20 minutes]*

A2: **it's hard to say it depends on the traffic condition it should take only twenty minutes or so if there's no traffic jam**

Data Preparation



ATR
Spoken Language Communication
Research Laboratories

Data Statistics

task	*lang uage*	*sentence count*	*word token*	*word type*	*words per sentence*
eval	C/E_7	500	6.0k / 50k	1.3k / 1.6k	**12.1** / **14.4**
	J/E_7		7.4k / 50k	1.2k / 1.6k	**14.8** / **14.4**
	A/E_7		5.2k / 50k	1.9k / 1.6k	**10.4** / **14.4**
	I/E_7		6.7k / 50k	1.5k / 1.6k	**13.4** / **14.4**

task		*Out-Of-Vocabulary rates (%)*		
		CRR	*1BEST*	*NBEST*
eval	C_S	2.6	2.1	2.4
	C_R	2.6	2.4	2.5
	J_R	2.2	1.6	2.3
	A_R	**14.3**	**16.0**	**17.1**
	I_R	4.3	2.5	2.6
	E_7	2.7 (for AE,IE) / 1.9 (for CE/JE)		

Recognition Accuracy

task		word (%)		sentence (%)	
		lattice	*1BEST*	*lattice*	*1BEST*
eval	C_S	79.08	**68.11**	22.80	**16.60**
	C_R	82.07	**73.64**	28.40	**22.80**
	J_R	90.48	**85.14**	52.60	**38.00**
	A_R	88.20	**73.88**	41.60	**16.60**
	I_R	72.90	**70.88**	5.40	**4.60**

- performance differences between source languages
  - closed language model used for Arabic and Chinese
- differences between lattice and 1BEST accuracies

Translation Conditions

Translation Directions

- Arabic
- Chinese
- Italian
- Japanese

→ English

Data Tracks

- **OPEN**:
  - in-domain training data restricted to supplied BTEC resources
- **CSTAR**:
  - no restrictions

Input Conditions

Cleaned Transcripts

plain text
(text normalized according to ASR engine)

⇓

correct recognition results of supplied ASR engines

ASR Output

word lattice
NBEST
1BEST

⇓

output of ASR engine supplied by CSTAR partner

Speech Input

audio data
(C: spontaneous speech)
(A,C,I,J: read speech)

⇓

each participant uses its own ASR engine

Spoken Language Communication Research Laboratories

Participants

Research Group		System	Type	Input
US	AT&T Research	**ATT**	SMT	C_S,C_R
EU	CLIPS-GETA	**CLIPS**	RBMT	C_S^*,J_R^*,A_R^*,I_R^*
EU	Dublin City University	**DCU**	EBMT	A_R,I_R^*
ZH	Hong Kong University	**HKUST**	SMT	C_S,C_R,J_R,A_R,I_R
US	IBM	**IBM**	SMT	A_R
EU	Instituto Trentino di Cultura	**ITC-irst**	SMT	C_S,C_R,J_R,A_R,I_R
US/EU	JHU.Summer Workshop 2006	**JHU_WS06**	SMT	C_S,C_R
JP	Kyoto University	**Kyoto-U**	EBMT	J_R
US	MIT Lincoln Lab / Air Force Research Lab	**MIT-LL-AFRL**	SMT	C_S,C_R,J_R,I_R
JP	National Institute of Science & Technology	**NAIST**	SMT	J_R
JP	NICT / ATR-SLC	**NiCT-ATR**	SMT	C_S,C_R,J_R,A_R,I_R
ZH	NLPR, Chinese Academy of Science	**NLPR**	RBMT,SMT	C_S,C_R
JP	NTT Communication Research	**NTT**	SMT	C_S,C_R,J_R,A_R,I_R
EU	Rheinisch Westfählische Hochschule	**RWTH**	SMT	C_S,C_R,J_R
EU	SHARP Laboratories of Europe	**SLE**	EBMT	J_R
EU	TALP-UPC Research Center **(2x)**	**TALP**	SMT	C_R,J_R,A_R,I_R
US	InterACT,CMU / Karlsruhe University **(2x)**	**UKACMU**	SMT	C_S,C_R,J_R,A_R,I_R
US	University of Washington	**Washington-U**	SMT	I_R
ZH	Xiamen University	**Xiamen-U**	SMT	C_S,C_R

Run Submissions

Type	Input	Lang	Group	Runs OPEN / C-STAR
text	correct recognition result	A_{CRR} C_{CRR} I_{CRR} J_{CRR}	19	*mandatory for all participants*
read speech	ASR output	A_R C_R I_R J_R	9 12 10 12	11 (14) / 1 (1) 14 (17) / 3 (3) 12 (14) / 1 (3) 14 (14) / 2 (3)
spontaneous speech	ASR output	C_S	12	12 (11) / 3 (3)
TOTAL			**19**	**63 (70) / 10 (13)**

Evaluation Specifications

- **case-sensitive, with punctuation marks** (*official*)
- case-insensitive, without punctuation marks (*additional*)

Automatic Evaluation:

- all run submissions
- metrics:

BLEU: 0 (bad — good) 1

NIST: 0 (bad — good) ∝

METEOR: 0 (bad — good) 1

Subjective Evaluation:

- CE, ASR Output, Open Data Track
- 7 MT engines with C_S,C_R,C_{CRR} run submissions
- median of 3 human grades
- metrics:

adequacy, fluency: 0 (bad — good) 4

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

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

Outline of Talk

1. Evaluation Campaign:

- data preparations
- translation input and data track conditions
- run submissions
- evaluation specifications

2. Evaluation Results:

- **subjective/automatic evaluation**
- **correlation between evaluation metrics**

3. Discussions:

- Challenge Task 2006
- source language effects
- robustness towards recognition errors
- innovative idea's explored by participants

Evaluation Results

Subjective Evaluation

· *adequacy/fluency* : **p.11** (scores, system rankings)

Automatic Evaluation

· *BLEU/NIST/METEOR*: **pp.12-13** (scores), **pp.14-15** (system rankings)

- test significance of differences in translation quality between two MT systems using **"bootStrap"** method:

  (1) perform a random sampling with replacement from the *eval* data

  (2) calculate respective evaluation metric scores of each MT engine and differences between the two MT engine scores

  (3) repeat sampling/scoring steps iteratively

  (4) apply Student's t-test at a significant level of 95% to test whether score differences are significant

→ **horizontal lines omitted in ranking tables, if system performance difference is NOT significant**

Subjective Evaluation Results

TOP Scores (MT Engine)

metric	C_{CRR}	C_R	C_S
adequacy	**1.4319** (MIT-LL-AFRL)	**1.0297** (JHU-WS06)	**0.9647** (JHU-WS06)
fluency	**1.6498** (RWTH)	**1.2952** (RWTH)	**1.3734** (MIT-LL-AFRL)

Combination of Subjective Evaluation Rankings

C_{CRR}	C_R	C_S
MIT-LL-AFRL	**JHU-WS06**	**JHU-WS06**
RWTH	**MIT-LL-AFRL**	RWTH
NTT	RWTH	NTT
JHU-WS06	NTT	**MIT-LL-AFRL**
NiCT-ATR	NiCT-ATR	UKACMU_SMT
UKACMU_SMT	UKACMU_SMT	NiCT-ATR

Spoken Language Communication Research Laboratories

Automatic Evaluation Results

TOP Scores (MT Engine) for ASR Output

input	BLEU	NIST	METEOR
C_S	**0.1898** (RWTH)	**5.1513** (JHU-WS)	**0.4238** (HKUST)
C_R	**0.2111** (RWTH)	**5.4154** (MIT-LL-AFRL)	**0.4456** (HKUST)
J_R	**0.2142** (RWTH)	**5.6502** (RWTH)	**0.4574** (NiCT-ATR)
A_R	**0.2274** (IBM)	**5.9216** (NiCT-ATR)	**0.4867** (NiCT-ATR)
I_R	**0.2989** (NiCT-ATR)	**6.9318** (Washington-U)	**0.5853** (Washington-U)

Spoken Language Communication Research Laboratories

Automatic Evaluation Results

Combination of Automatic Evaluation Rankings

C_S	C_R	J_R	A_R	I_R
RWTH	RWTH	RWTH	IBM	Washington-U
JHU-WS06	MIT-AFRL	NiCT-ATR	NiCT-ATR	NiCT-ATR
NiCT-ATR	NiCT-ATR	UKACMU	TALP-tuples	TALP-tuples
UKACMU	JHU-WS06	NTT	TALP-comb	MIT-AFRL
HKUST	ITC-irst	MIT-AFRL	NTT	TALP-comb
ITC-irst	TALP-tuples	ITC-irst	UKACMU	ITC-irst
MIT-AFRL	TALP-phrases	SLE	TALP-phrases	TALP-phrases
NTT	UKACMU	HKUST	ITC-irst	NTT
Xiamen-U	HKUST	TALP-tuples	DCU	DCU
ATT	TALP-comb	NAIST	HKUST	UKACMU
NLPR	NTT	Kyoto-U	CLIPS	HKUST
CLIPS	Xiamen-U	TALP-comb		CLIPS
	NLPR	TALP-phrases		
	ATT	CLIPS		

Correlation between Automatic and Subjective Evaluation Metrics

input	metric	BLEU	NIST	METEOR
C_{CRR}	fluency	**0.96**	0.84	0.93
	adequacy	0.95	0.82	**0.96**
C_R	fluency	**0.89**	0.63	0.66
	adequacy	0.83	0.64	**0.89**
C_S	fluency	**0.88**	0.55	0.72
	adequacy	0.34	**0.57**	0.54

Outline of Talk

1. Evaluation Campaign:

- data preparations
- translation input and data track conditions
- run submissions
- evaluation specifications

2. Evaluation Results:

- subjective/automatic evaluation
- correlation between evaluation metrics

3. Discussions:

- **Challenge Task 2006**
- **source language effects**
- **robustness towards recognition errors**
- **innovative idea's explored by participants**

ATR
Spoken Language Communication
Research Laboratories

Challenge Task 2006

- more difficult than previous IWSLT tasks

translation task	*training data*	
	40k (CE/JE)	*20k (AE/IE)*
dev_1 / dev_2 / dev_3	27.5 / 31.4 / 32.9	32.6 / 36.7 / 38.8
dev_4 / eval	**85.6 / 105.9**	**98.3 / 113.9**

- quite low MT performance for all systems for all conditions
  - discrepancy between training and evaluation data
  - high OOV figures
  - number of reference translations differed (16 vs. 7)

Source Language Effects

Italian:
- highest scores despite worst recognition accuracy
  → close language relationship

Arabic:
- largest OOV rates
  → re-segmentation led to improved coverage & translation quality

Japanese:
- highest recognition accuracy, but low scores
  → one of the most difficult translation tasks
- largest number of non-SMT run submissions

Chinese:
- recognition accuracy similar to Arabic, but much lower scores
- largest number of participants

task complexity:

CE ≈ JE > AE ≫ IE

Robustness Towards Recognition Errors

TOP-scoring systems		recognition errors			
		none	low	medium	high
C_S	%	16.6	39.6	34.4	9.4
	adequacy	1.52	1.14	0.69	0.27
	fluency	1.81	1.22	1.21	0.58
C_R	%	22.8	45.5	22.4	9.3
	adequacy	1.80	1.02	0.63	0.17
	fluency	2.05	1.13	0.91	0.30

ATR
Spoken Language Communication
Research Laboratories

Innovative Ideas Explored by Participants

- **additional training resources**
  - · *in-domain* → large gain (CSTAR data track)
  - · *out-of-domain* → partially effective (IBM, Washington-U)
- **distortion modeling** (ITC-irst, TALP)
- **topic-dependent model adaptation** (NiCT-ATR)
- **efficient decoding of word lattices** (JHU_WS06, ITC-irst)
- **rescoring/ranking features** (NTT, RWTH, Washington-U)

closer coupling of ASR and MT technologies required to overcome problems of speech translation tasks

ATR
Spoken Language Communication
Research Laboratories

The End

谢谢大家注意
听我的发言。

ご静聴、ありがとう
ございました。

Thank you
for your attention!

شكراً على انتباهكم.

Grazie
per la vostra
attenzione!