INTERSPEECH 2004 - ICSLP Satellite Workshop

International Workshop on Spoken Language Translation

- Evaluation Campaign on Spoken Language Translation -

September 30 - October 1, 2004 Spoken Language Translation Research Laboratories Kyoto, Japan

Overview of the IWSLT04 Evaluation Campaign Results

September 30, 2004

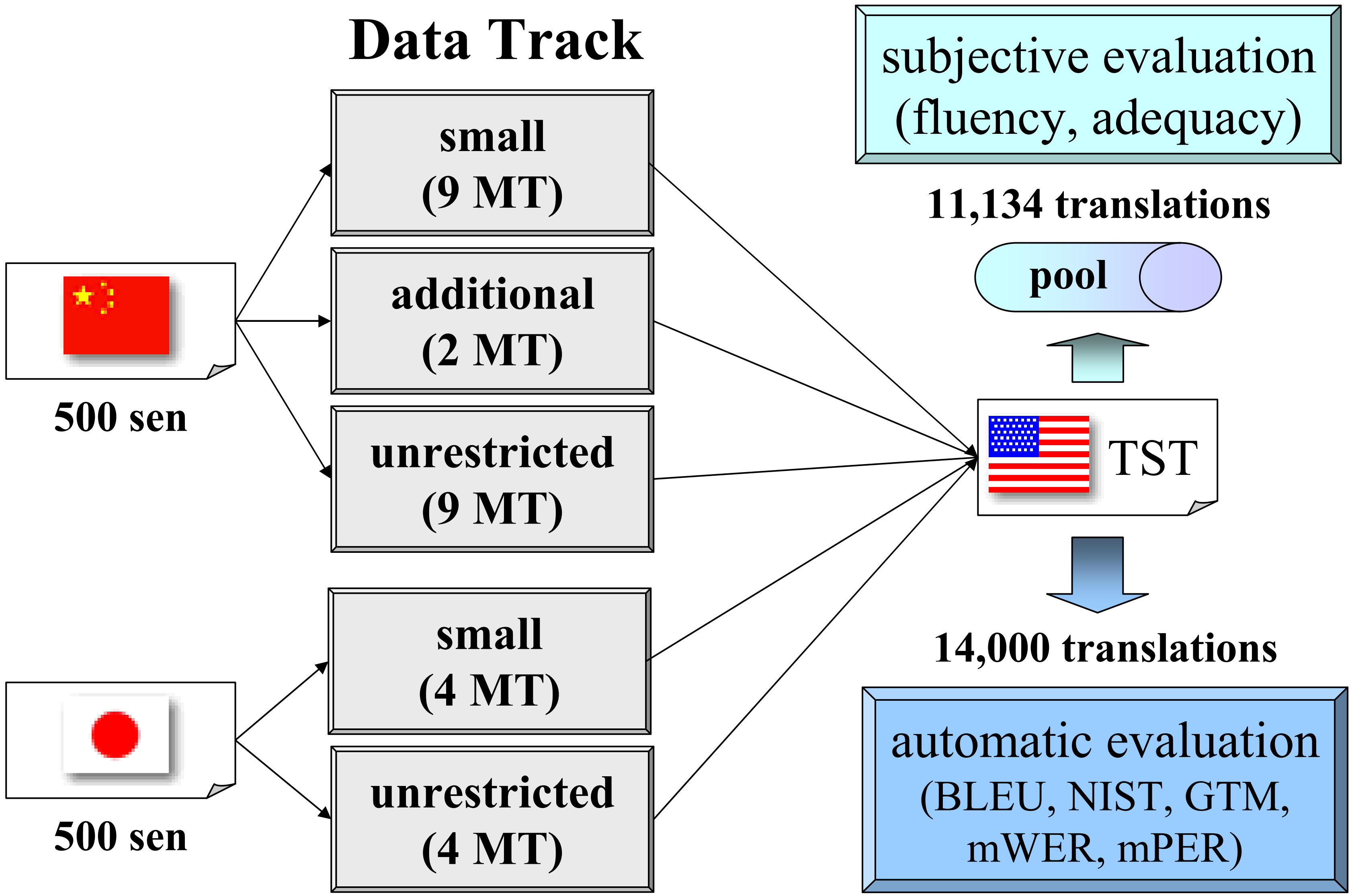
Yasuhiro Akiba[*1], Marcello Federico[*2], Noriko Kando[*3],
Hiromi Nakaiwa[*1], Michael Paul[*1], Jun'ichi Tsujii[*4]

[*1] ATR, [*2]ITC-irst, [*3]NII, [*4]University of Tokyo

Participants and their systems

- 14 institutions took part in the evaluation campaign.
  - – 7 SMT systems
    - ATR-SMT, IBM, IRST, ISI, ISL-SMT, RWTH, TALP
  - – 3 EBMT systems
    - HIT, ICT, UTokyo
  - – 1 RBMT system
    - CLIPS
  - – 4 Hybrid MT systems
    - ATR-HYBRID, IAI, ISL-EDTRL, NLPR

Evaluation Campaign



Data Track

small
(9 MT)

additional
(2 MT)

unrestricted
(9 MT)

500 sen

small
(4 MT)

unrestricted
(4 MT)

500 sen

TST

subjective evaluation
(fluency, adequacy)

11,134 translations

pool

14,000 translations

automatic evaluation
(BLEU, NIST, GTM,
mWER, mPER)

Data Track Conditions

Small Data Track (C-to-E, J-to-E):

- supplied corpus only

Additional Data Track (C-to-E):

- limits the use of bilingual resources
- no restrictions on monolingual resources
- besides supplied corpus, additional bilingual resources available from LDC are permitted

Unrestricted Data Track (C-to-E , J-to-E):

- no limitations on linguistic resources used to train the MT systems

Outline

1. **Evaluation Campaign**
    - participants and their systems
    - data track conditions
    - subjective/automatic evaluation metrics
2. **Evaluation Results**
    - subjective evaluation results
    - automatic evaluation results
    - correlation (automatic vs. subjective)
    - discussion

Evaluation Specifications

Evaluation Measures:

- *subjective*: fluency, adequacy (3 graders per translation)
- *automatic*: BLEU, NIST, GTM, mWER, mPER

Evaluation Parameter:

- case-insensitive
- no punctuation marks, i.e. remove '.' ',' '?' '!' '"'
- no word compounds, i.e. replace '-' with blank space
- spelling-out of numerals
- non-ASCII characters not permitted
- comparison using word and part-of-speech information

Subjective Evaluation

Fluency:

- degree to which translation is well-formed

Adequacy:

- degree to which translation communicates information present in reference translations

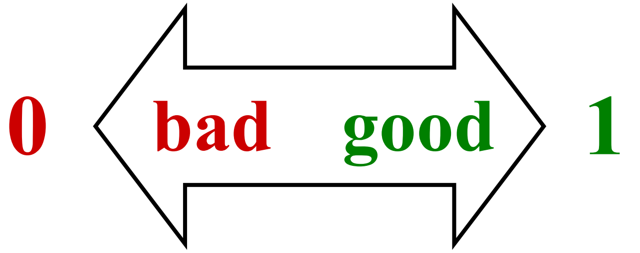
Fluency	
5	Flawless English
4	Good English
3	Non-native English
2	Disfluent English
1	Incomprehensible

Adequacy	
5	All Information
4	Most Information
3	Much Information
2	Little Information
1	None

Automatic Evaluation Measures

BLEU:

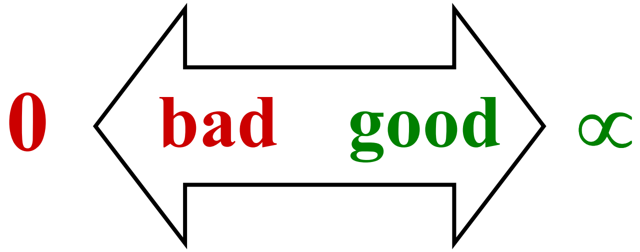
- the ***geometric mean of n-gram precision*** by the system output with respect to the reference translations.



0 bad good 1

NIST:

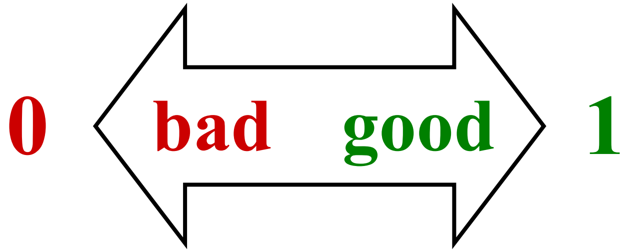
- a variant of BLEU using the ***arithmetic mean of weighted n-gram precision*** values.



0 bad good ∝

GTM:

- measures similarity between text using ***a unigram-based F-measure***

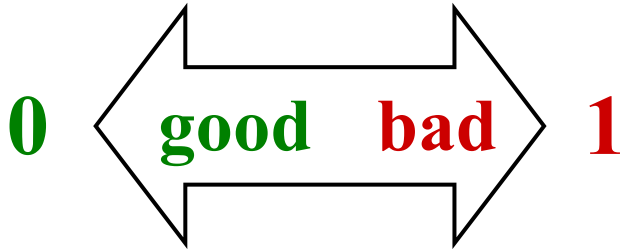


0 bad good 1

Automatic Evaluation Measures

mWER:

- ***multiple Word Error Rate***, the edit distance between the system output and the closest reference translation.



0 good bad 1

mPER:

- ***position-independent mWER***, a variant of mWER which disregards word ordering.

Outline

1. **Evaluation Campaign**
   - participants and their systems
   - data track conditions
   - subjective/automatic evaluation metrics
2. **Evaluation Results**
   - subjective evaluation results
   - automatic evaluation results
   - correlation (automatic vs. subjective)
   - discussion

Outline of Evaluation Results

2.1. Subjective evaluation results

→ ***fluency***, ***adequacy***

A) How consistently did a group of three graders evaluate them?

B) How consistently did each grader evaluate translations?

C) How were MT systems ranked according to the subjective evaluation?

2.2. Automatic evaluation results

→ ***mWER***, ***mPER***, ***BLEU***, ***NIST***, ***GTM***

2.3. Correlation between subjective and automatic evaluation results

Workload of Graders

TEST	1st grader	2nd grader	3rd grader	# input data
Team 1	G0	G2	G9	200
Team 2	G4	G5	G8	160
Team 3	G1	G3	G6	80
Team 4	G0	G3	G7	60

additional evaluation set I (*COMMON*):

- 100 translations randomly selected from all MT outputs
- the common data set was evaluated by all graders
- compare the grading differences between graders

additional evaluation set II (*GRADER*):

- 100 translations randomly selected from MT outputs assigned to grader
- the grader-specific data set was evaluated a second time
- validate self-consistency of each grader

Consistency of Median Grades

- 100 translations randomly selected from all MT outputs
- Team of three graders for each median grade

COMMON	Fluency			Adequacy		
	T2	T3	T4	T2	T3	T4
Team 1 (T1)	0.49	0.75	0.47	0.54	0.61	0.34
Team 2 (T2)	–	0.68	0.66	–	0.59	0.48
Team 3 (T3)	–	–	0.44	–	–	0.51
Average	**0.58**			**0.51**		

→ the expected difference of fluency/adequacy were **0.55**

→ the quality of two MT systems whose difference is **less than 1.1** cannot be distinguished

Self-Consistency of Graders

- 100 translations randomly selected from MT outputs assigned to each grader
- Expected difference between two assessments by the same grader

GRADER	Fluency	Adequacy
G0 – G9	0.12 – 0.77	0.33 – 0.64
Average	**0.39**	**0.44**

→ the quality of two MT systems whose difference in either fluency or adequacy is **less than 0.8** cannot be distinguished

Minimization of Grading Errors

- Reducing the error rates of the graders
  - subjective evaluation is a classification task.
  - merging two classes that are difficult to distinguish reduces the error rate.

- Examine the following binary classifications
  - → “5” vs. “less than 5”
  - → “larger than or equal to 4” vs. “less than 4”
  - → “larger than or equal to 3” vs. “less than 3”
  - → “larger than or equal to 2” vs. “less than 2”

Minimization of Grading Errors

- Error rates of binary classifications

GRADER		**5 or <5**	≥4 or <4	≥3 or <3	≥2 or <2	**5- grade**
avg.	fluency	**0.07**	0.08	0.15	0.09	**0.32**
	adequacy	**0.10**	0.12	0.13	0.09	**0.36**
min.	fluency	**0.01**	0.03	0.06	0.02	**0.14**
	adequacy	**0.05**	0.06	0.07	0.04	**0.23**

→ error rates of binary classification are **much smaller** than the 5-grade classification

Minimization of Grading Errors

- Error rates of assessments by the grader with the **smallest error for each team**

COMMON	Fluency	Adequacy
Team 1	0.01	0.07
Team 2	0.05	0.09
Team 3	0.05	0.05
Team 4	0.01	0.05
Total	**0.03**	**0.07**

TEST # data
200
160
80
60
500

Ranking MT Systems

- **Regular ranking lists**
  - MT system ranking using ***5-grade classification***
  - Scores are in the range of [0, 5]
  - Higher scores indicate better systems

  → ***NOTE***: self-consistency of each grader is low!

- **Alternative ranking lists**
  - MT system ranking using **"5 or <5"** ***classification***
  - Scores are in the range of [0, 1]
  - Higher scores indicate better systems

MT Ranking List (Fluency)

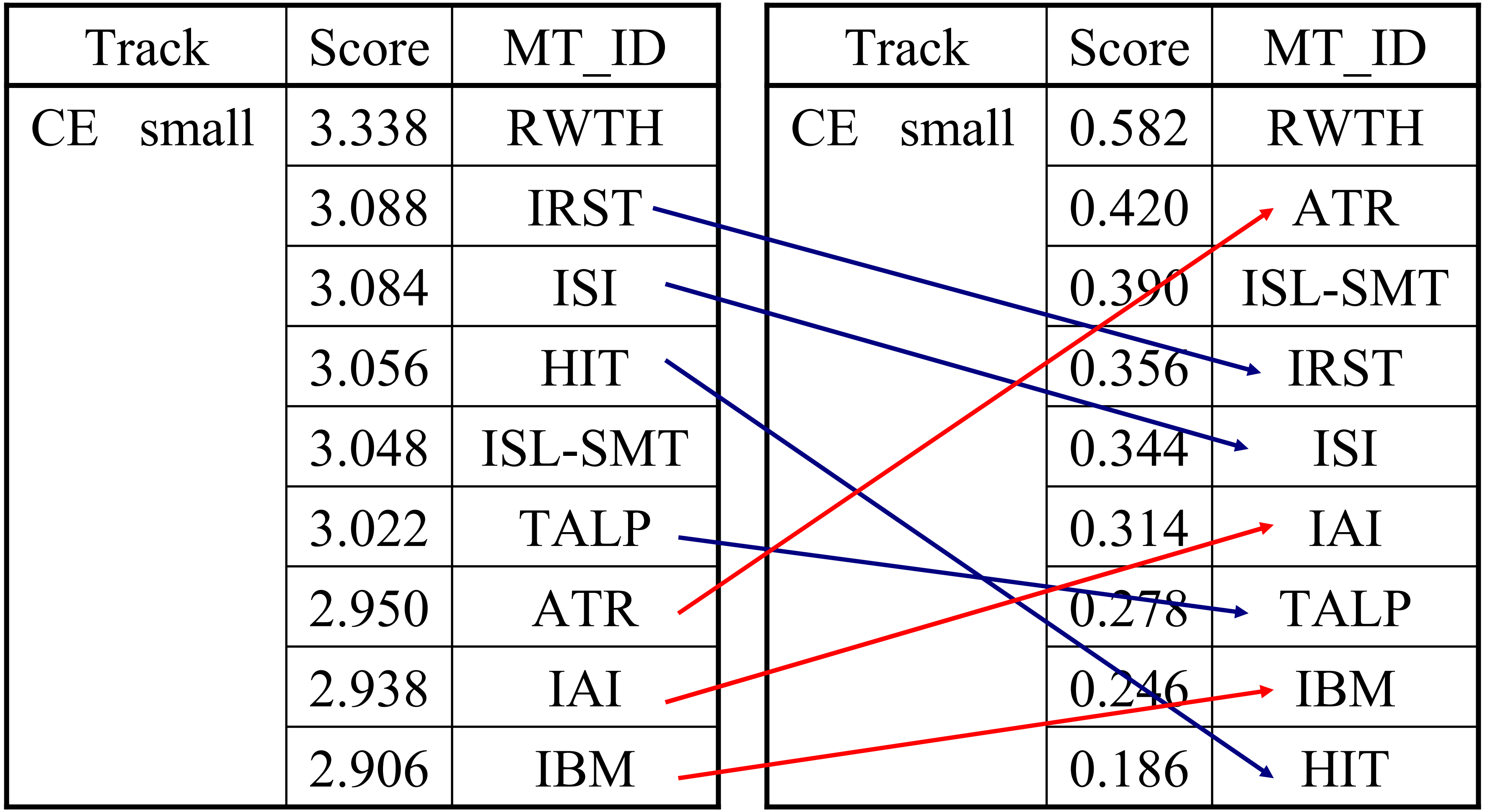
Regular Ranking

Track	Score	MT_ID
CE small	3.820	ATR
	3.356	RWTH
	3.332	ISL-SMT
	3.120	IRST
	3.074	ISI
	2.948	IBM
	2.914	IAI
	2.792	TALP
	2.504	HIT

Alternative Ranking

Track	Score	MT_ID
CE small	0.582	ATR
	0.420	ISL-SMT
	0.390	RWTH
	0.356	IRST
	0.344	ISI
	0.314	IBM
	0.278	IAI
	0.246	TALP
	0.186	HIT

MT Ranking List (Adequacy)



Regular Ranking

Track	Score	MT_ID
CE small	3.338	RWTH
	3.088	IRST
	3.084	ISI
	3.056	HIT
	3.048	ISL-SMT
	3.022	TALP
	2.950	ATR
	2.938	IAI
	2.906	IBM

Alternative Ranking

Track	Score	MT_ID
CE small	0.582	RWTH
	0.420	ATR
	0.390	ISL-SMT
	0.356	IRST
	0.344	ISI
	0.314	IAI
	0.278	TALP
	0.246	IBM
	0.186	HIT

Reliability of Difference in MT Quality (Fluency)

Regular Ranking

Track	Score	MT_ID
CE small	3.820	ATR
	3.356	RWTH
	3.332	ISL-SMT
	3.120	IRST
	3.074	ISI
	2.948	IBM
	2.914	IAI
	2.792	TALP
	2.504	HIT

Alternative Ranking

Track	Score	MT_ID
CE small	0.582	ATR
	0.420	ISL-SMT
	0.390	RWTH
	0.356	IRST
	0.344	ISI
	0.314	IBM
	0.278	IAI
	0.246	TALP
	0.186	HIT

Reliability of Difference in MT Quality (Adequacy)

Regular Ranking

Track	Score	MT_ID
CE small	3.338	RWTH
	3.088	IRST
	3.084	ISI
	3.056	HIT
	3.048	ISL-SMT
	3.022	TALP
	2.950	ATR
	2.938	IAI
	2.906	IBM

Alternative Ranking

Track	Score	MT_ID
CE small	**0.582**	**RWTH**
	0.420	**ATR**
	0.390	**ISL-SMT**
	0.356	**IRST**
	0.344	ISI
	0.314	IAI
	0.278	TALP
	0.246	IBM
	0.186	HIT

Outline of Evaluation Results

3.1. Subjective evaluation results

→ ***fluency***, ***adequacy***

A) How consistently did each grader evaluate translations?

B) How consistently did a group of three graders evaluate them?

C) How were MT systems ranked according to the subjective evaluation?

3.2. Automatic evaluation results

→ ***mWER***, ***mPER***, ***BLEU***, ***NIST***, ***GTM***

3.3. Correlation between subjective and automatic evaluation results

MT Ranking List (automatic evaluation metrics)

CE small

mWER		mPER		BLEU		NIST		GTM	
score	MT	score	MT	score	MT	score	MT	score	MT
0.455	RWTH	0.390	RWTH	0.454	ATR	8.55	RWTH	0.720	RWTH
0.469	ATR	0.404	ISL-S	0.414	ISL-S	8.34	ISL-S	0.694	ISL-S
0.471	ISL-S	0.420	ATR	0.408	RWTH	7.85	IAI	0.685	IAI
0.488	ISI	0.425	ISI	0.374	ISI	7.74	ISI	0.672	ISI
0.507	IRST	0.430	IRST	0.349	IRST	7.48	ATR	0.670	ATR
0.532	IAI	0.451	IAI	0.346	IBM	7.12	IBM	0.665	IBM
0.538	IBM	0.452	IBM	0.338	IAI	7.09	IRST	0.647	TALP
0.556	TALP	0.465	TALP	0.278	TALP	6.77	TALP	0.644	IRST
0.616	HIT	0.500	HIT	0.209	HIT	5.95	HIT	0.601	HIT

MT Ranking List (automatic evaluation metrics)

JE unrestricted

mWER		mPER		BLEU		NIST		GTM	
score	MT	score	MT	score	MT	score	MT	score	MT
0.263	ATR	0.233	ATR	0.630	ATR	11.25	RWTH	0.824	RWTH
0.305	RWTH	0.249	RWTH	0.619	RWTH	10.72	ATR	0.796	ATR
0.485	UTokyo	0.420	UTokyo	0.397	UTokyo	7.88	UTokyo	0.672	UTokyo
0.730	CLIPS	0.597	CLIPS	0.132	CLIPS	5.64	CLIPS	0.568	CLIPS

Outline of Evaluation Results

3.1. Subjective evaluation results

→ ***fluency***, ***adequacy***

A) How consistently did each grader evaluate translations?

B) How consistently did a group of three graders evaluate them?

C) How were MT systems ranked according to the subjective evaluation?

3.2. Automatic evaluation results

→ ***mWER***, ***mPER***, ***BLEU***, ***NIST***, ***GTM***

3.3. Correlation between subjective and automatic evaluation results

Correlation Coefficients

Regular Ranking vs. Automatic Ranking (all MT systems)

		mWER	mPER	**BLEU**	**NIST**	GTM
Fluency	CE	-0.7124	-0.5830	**0.8505**	0.5995	0.5132
	JE	-0.8867	-0.7836	**0.9404**	0.5995	0.6387
Adequacy	CE	-0.4324	-0.4404	0.4376	**0.5318**	0.3711
	JE	-0.8978	-0.9376	0.7884	**0.9701**	0.9401

Correlation Coefficients

Alternative Ranking vs. Automatic Ranking (all MT systems)

		mWER	mPER	BLEU	NIST	GTM
Fluency	CE	-0.7214	-0.6010	**0.8600**	0.5950	0.5214
	JE	-0.8252	-0.7032	**0.9070**	0.4871	0.5383
Adequacy	CE	-0.6427	-0.5779	**0.7407**	0.6820	0.5136
	JE	**-0.9690**	-0.9641	0.9157	0.9176	0.9152

Correlation Coefficients

Alternative Ranking vs. Automatic Ranking (partial MT systems)

		mWER	**mPER**	**BLEU**	NIST	GTM
Fluency	CE	-0.8734	-0.6743	**0.9548**	0.5736	0.5454
	JE	-0.8376	-0.7223	**0.9288**	0.5089	0.5632
Adequacy	CE	-1.0000	**-1.0000**	1.0000	1.0000	1.0000
	JE	-0.9894	**-0.9984**	0.9195	0.9907	0.9977

partial = MT systems whose score difference is at least twice at large as the word error rates

Discussion

1. Evaluation scheme

- description of grades are ambiguous (e.g. “non-native English” vs. “disfluent English”)
- separate grading of translations of same input
- selection of single reference translations for *adequacy*
- splitting of workload between 3-4 graders

→ Improve consistency of subjective evaluation grades

- unambiguous definition of evaluation grades
- simultaneous grading of all MT outputs for given input
- select grader with high self-consistency
- evaluation of all MT outputs by same graders

Discussion

2. Grading ambiguity

- partially correct translations
- additional information NOT in reference translation
- importance of specific pieces of information (e.g. numeral expressions)
- missing context for very short utterances

→ **Identify and evaluate important pieces of information**

- mark information units to be evaluated
- provide more context for highly ambiguous utterances
- careful selection of appropriate reference translations

The End

Basic Travel Expression Corpus (BTEC*)

useful phrases, together with the translation into other languages

usually found in phrasebooks for tourists going abroad

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.

Status of BTEC* Corpus

language	*sentence count*	*word token*	*word type*	*words per sentence*
Japanese	162k	1,114,186	18,781	6.9
English		952,300	12,404	5.9
Korean		1,211,129	21,837	7.5
Chinese		959,846	15,516	5.9
Italian	48k	361,250	14,871	7.4

🚧 **Spanish, French, German** 🚧

IWSLT04 Corpus

		sentence count		*avg.*	*word*	*word*
		total	*unique*	*length*	*tokens*	*types*
training ↓ *JE ≠ CE*	Chinese	20,000	19,288	9.1	182,904	7,643
	English		19,949	9.4	188,935	8,191
	Japanese	20,000	19,046	10.5	209,012	9,277
	English		19,923	9.4	188,712	8,074
develop ↓ *JE =CE*	Chinese	506	495	6.9	3,515	870
	Japanese	506	502	8.6	4,374	954
	English*	8,089	7,173	7.5	67,410	2,435
test ↓ *JE =CE*	Chinese	500	492	7.6	3,794	893
	Japanese	500	491	8.7	4,370	979
	English*	8,000	6,907	8.4	66,994	2,496

* up to 16 multiple reference translations

Permitted LDC Resources

LDC2000T46	Hong Kong News Parallel Text
LDC2000T47	Hong Kong Laws Parallel Text
LDC2000T50	Hong Kong Hansards Parallel Text
LDC2001T11	Chinese Treebank 2.0
LDC2001T57	TDT2 Multilanguage Text V4.0
LDC2001T58	TDT3 Multilanguage Text V2.0
LDC2002L27	Chinese English Translation Lexicon V3.0
LDC2002T01	Multiple-Translation Chinese Corpus
LDC2003T16	SummBank 1.0
LDC2003T17	Multiple-Translation Chinese Part 2
LDC2004T05	Chinese Treebank 4.0
LDC2004T09	ACE 2003 Multilingual Training Data

Permitted Linguistic Resources

Resources	Data Track		
	Small	Additional	Unrestricted
IWSLT04 corpus	✓	✓	✓
LDC resources	✘	✓	✓
tagger	✘	✓	✓
chunker	✘	✓	✓
parser	✘	✓	✓
external bilingual dictionaries	✘	✘	✓
other resources	✘	✘	✓

IWSLT 2004 Participants

participant		MT system	CE	JE
ATR	ATR Spoken Language Translation (JPN)	SMT, Hybrid	✓	✓
CLIPS	Universite Joseph Fourier (FRA)	RBMT	✓	✓
HIT	Harbin Institute of Technology (CHN)	EBMT	✓	✗
IAI	IAI (DEU), RALI (CAN), NUK (TWN)	Hybrid	✓	✗
IBM	IBM (USA)	SMT	✓	✓
ICT	Institute of Computing Technology (CHN)	EBMT	✓	✗
IRST	ITC-irst (ITA)	SMT	✓	✗
ISI	Information Sciences Institute/USC (USA)	SMT	✓	✓
ISL-EDTRL	University of Karlsruhe (DEU)	Hybrid	✓	✗
ISL-SMT	Carnegie Mellon University (USA)	SMT	✓	✗
NLPR	National Lab. of Pattern Recognition (CHN)	Hybrid	✓	✗
RWTH	Rheinisch Westf. Techn. Hochschule (DEU)	SMT	✓	✓
TALP	Univerisitat Polytecnica de Catalunya (ESP)	SMT	✓	✗
UTokyo	University of Tokyo (JPN)	EBMT	✗	✓

IWSLT 2004 Participants

Data Track	**CE**	**JE**
Small	9	4
Additional	2	—
Unrestricted	9	4
Organization	13	6

MT Engine	
SMT	7
EBMT	3
RBMT	1
Hybrid	4

Hybrid
SMT+EBMT
SMT+TM
SMT+IF
RBMT+IF

Evaluation Procedure

Run Submission:

- at most one MT system of each participant per data track
- multiple run submissions for each track permitted

Evaluation:

- automatic evaluation applied to all submissions
- human assessment only for one run
  (selected by participants themselves)

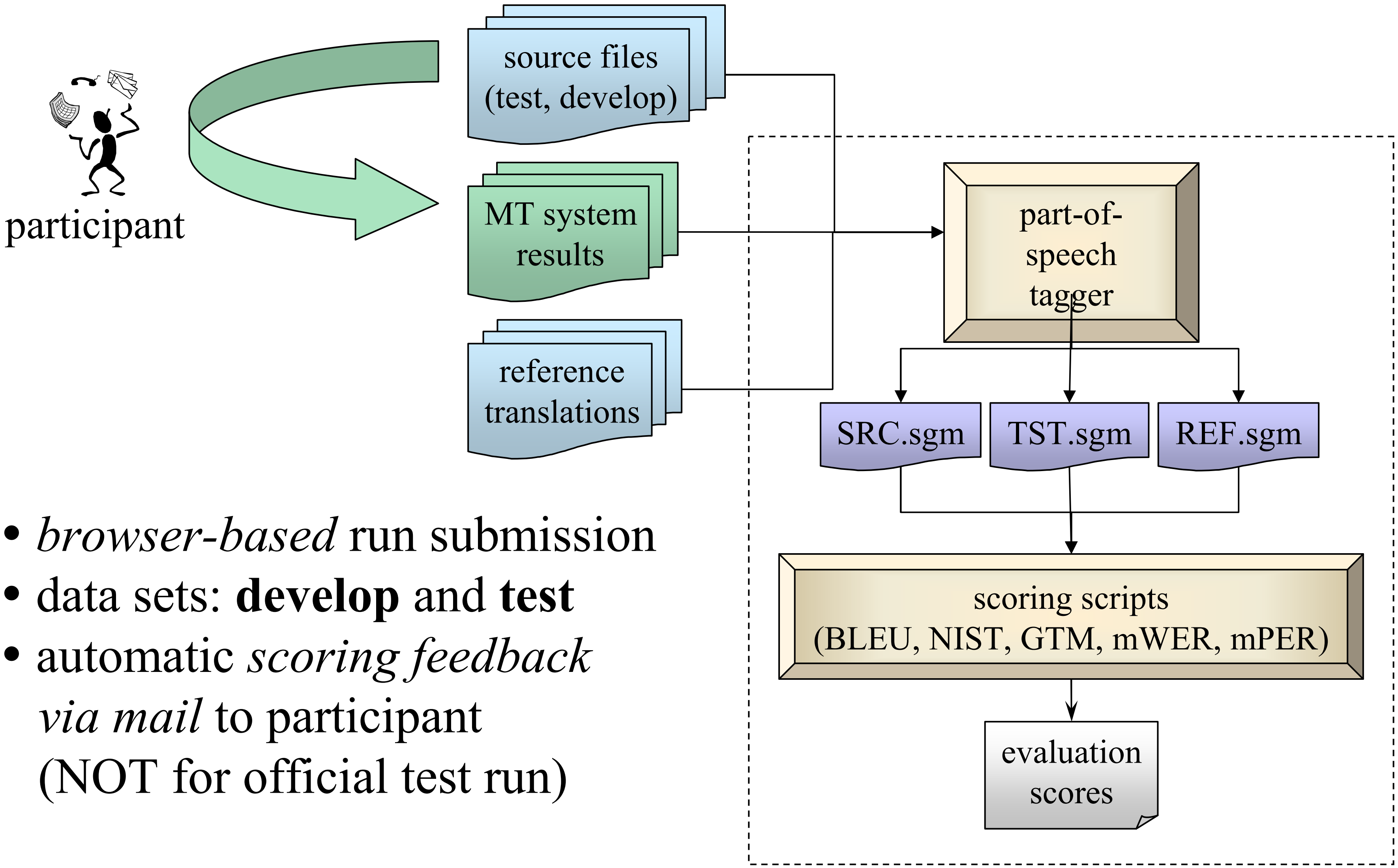
Evaluation Results:

- automatic scoring and subjective grading results
- MT system ranking lists for each data track
- correlation between automatic and subjective evaluation

Evaluation Campaign Schedule

Event	Date
Evaluation Specifications	February 15, 2004
Application Submission	April 15, 2004
Notification of Acceptance	April 30, 2004
Sample Corpus Release	May 7, 2004
Training Corpus Release	May 21, 2004
Develop Corpus Release	July 15, 2004
Test Corpus Release	August 9, 2004
Run Submission	August 12, 2004
Result Feedback	September 10, 2004
Camera-ready Submission	September 17, 2004
Workshop	September 30 – October 1, 2004

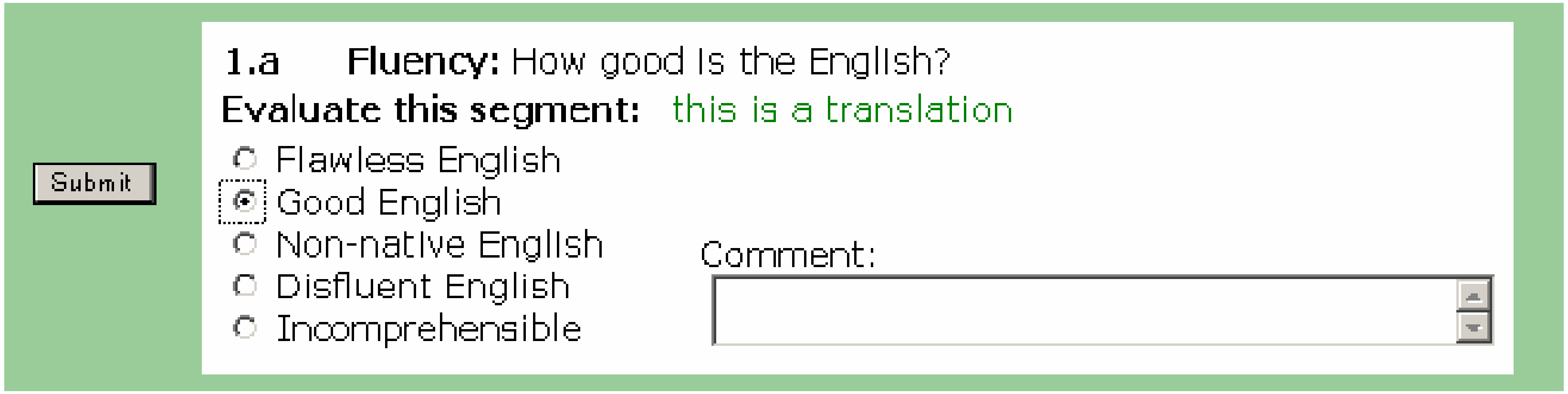
Automatic Evaluation Tool



- *browser-based* run submission
- data sets: **develop** and **test**
- automatic *scoring feedback via mail* to participant (NOT for official test run)

Subjective Evaluation Interface

Step 1:



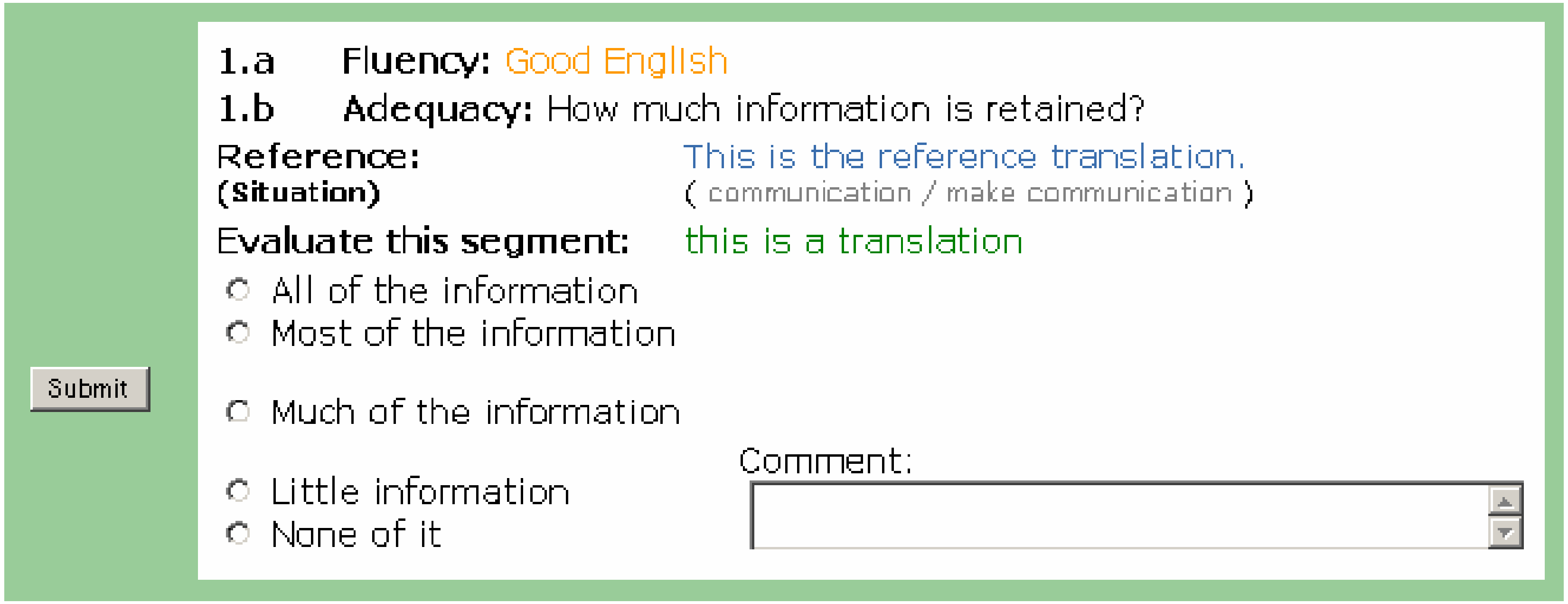
1.a **Fluency:** How good is the English?

Evaluate this segment: this is a translation

○ Flawless English
● Good English
○ Non-native English
○ Disfluent English
○ Incomprehensible

Comment:

Step 2:



1.a **Fluency:** Good English

1.b **Adequacy:** How much information is retained?

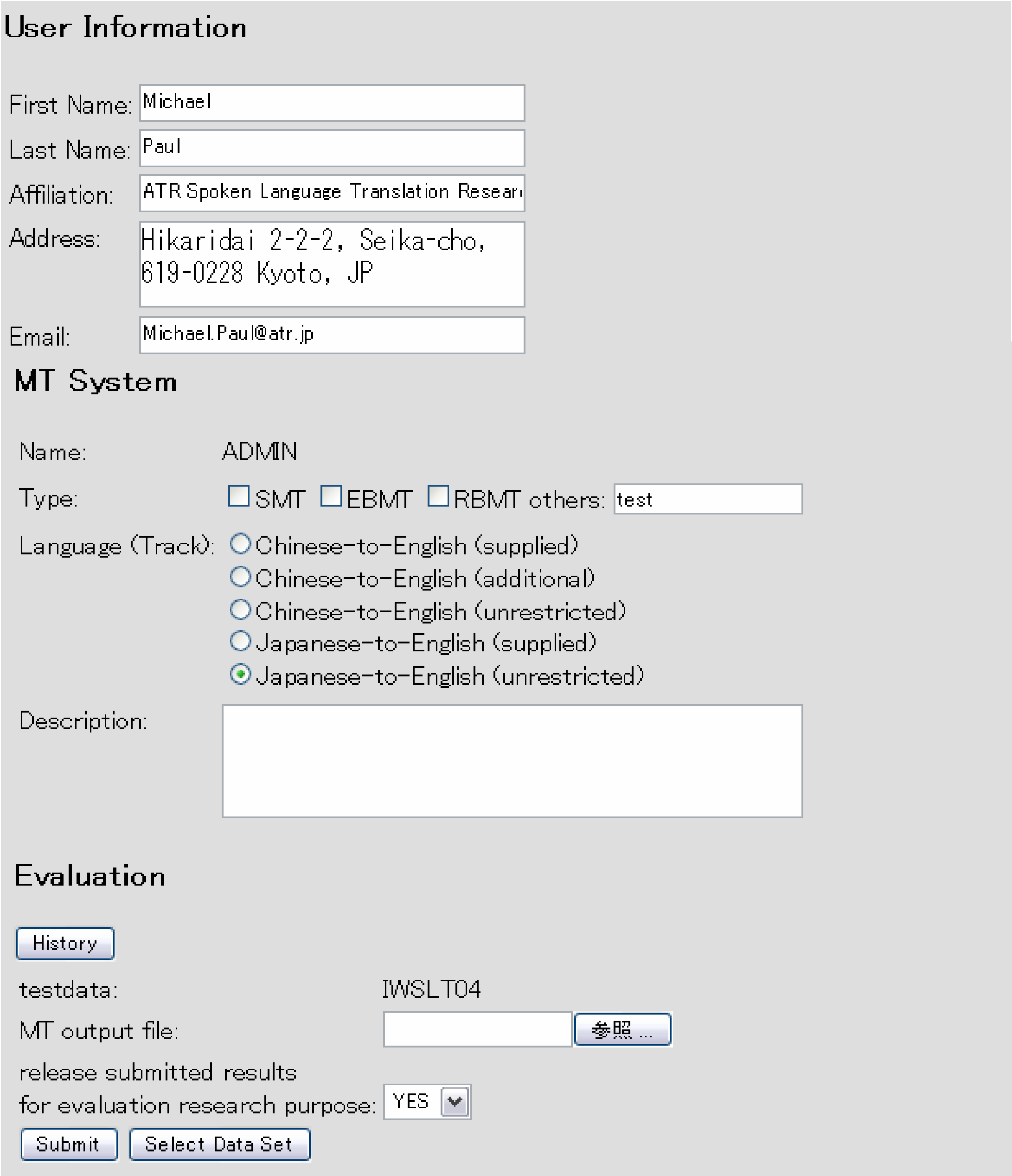
Reference: This is the reference translation.
(Situation) (communication / make communication)

Evaluate this segment: this is a translation

○ All of the information
○ Most of the information
○ Much of the information
○ Little information
○ None of it

Comment:

Automatic Evaluation Tool



User Information

First Name: Michael

Last Name: Paul

Affiliation: ATR Spoken Language Translation Resear

Address: Hikaridai 2-2-2, Seika-cho,
619-0228 Kyoto, JP

Email: Michael.Paul@atr.jp

MT System

Name: ADMIN

Type: □ SMT □ EBMT □ RBMT others: test

Language (Track):
○ Chinese-to-English (supplied)
○ Chinese-to-English (additional)
○ Chinese-to-English (unrestricted)
○ Japanese-to-English (supplied)
● Japanese-to-English (unrestricted)

Description:

Evaluation

testdata: IWSLT04

MT output file:

release submitted results
for evaluation research purpose: YES

Self-Consistency of Graders

- Error rates of the graders

GRADER	Fluency	Adequacy
G0 – G9	**0.14** – 0.53	**0.23** – 0.55
Average	0.32	0.36

→ even in the smallest case, the error rates are around **20%**, which were considerably larger than expected.

MT Ranking List
(Fluency)

Regular Ranking

Track	score	MT_ID
CE additional	3.256	IRST
	2.846	ISI

Alternative Ranking

Track	score	MT_ID
CE additional	0.410	IRST
	0.284	ISI

MT Ranking List (Adequacy)

Regular Ranking

Track	score	MT_ID
CE additional	3.110	IRST
	2.724	ISI

Alternative Ranking

Track	score	MT_ID
CE additional	0.316	IRST
	0.212	ISI

MT Ranking List (Fluency)

Regular Ranking

Track	score	MT_ID
CE unrestricted	3.776	IRST
	3.776	ISL-SMT
	3.400	NLPR
	3.036	IBM
	2.954	ISI
	2.934	ISL-EDTRL
	2.718	ICT
	2.648	HIT
	2.570	CLIPS

Alternative Ranking

Track	score	MT_ID
CE unrestricted	0.558	IRST
	0.532	ISL-SMT
	0.406	NLPR
	0.326	IBM
	0.296	ISL-EDTRL
	0.286	ISI
	0.224	HIT
	0.222	ICT
	0.180	CLIPS

MT Ranking List (Adequacy)

Regular Ranking

Track	score	MT_ID
CE unrestricted	3.662	ISL-SMT
	3.526	IRST
	3.254	ISL-EDTRL
	3.188	HIT
	3.082	ICT
	2.996	IBM
	2.960	CLIPS
	2.800	NLPR
	2.784	ISI

Alternative Ranking

Track	score	MT_ID
CE unrestricted	0.446	ISL-SMT
	0.394	IRST
	0.294	ISL-EDTRL
	0.258	ICT
	0.250	IBM
	0.228	NLPR
	0.226	HIT
	0.178	ISI
	0.164	CLIPS

MT Ranking List (Fluency)

Regular Ranking

Track	score	MT_ID
JE unrestricted	4.308	ATR
	4.036	RWTH
	3.650	UTokyo
	2.472	CLIPS

Track	score	MT_ID
JE small	3.484	ATR
	3.480	RWTH
	3.106	IBM
	3.102	ISI

Alternative Ranking

Track	score	MT_ID
JE unrestricted	0.698	ATR
	0.608	RWTH
	0.506	UTokyo
	0.170	CLIPS

Track	score	MT_ID
JE small	0.520	ATR
	0.440	RWTH
	0.368	ISI
	0.334	IBM

MT Ranking List (Adequacy)

Regular Ranking

Track	score	MT_ID
JE unrestricted	4.208	ATR
	4.066	RWTH
	3.316	UTokyo
	2.602	CLIPS

Track	score	MT_ID
JE small	3.412	RWTH
	3.086	ISI
	2.990	IBM
	1.942	ATR

Alternative Ranking

Track	score	MT_ID
JE unrestricted	0.600	ATR
	0.564	RWTH
	0.360	UTokyo
	0.120	CLIPS

Track	score	MT_ID
JE small	0.358	RWTH
	0.304	ISI
	0.262	IBM
	0.126	ATR

Reliability of Difference in MT Quality (Fluency)

Regular Ranking

Track	score	MT_ID
CE additional	3.256	IRST
	2.846	ISI

Alternative Ranking

Track	score	MT_ID
CE additional	0.410	IRST
	0.284	ISI

Reliability of Difference in MT Quality (Adequacy)

Regular Ranking

Track	score	MT_ID
CE additional	3.110	IRST
	2.724	ISI

Alternative Ranking

Track	score	MT_ID
CE additional	0.316	IRST
	0.212	ISI

Reliability of Difference in MT Quality (Fluency)

Regular Ranking

Track	score	MT_ID
CE unrestricted	**3.776**	**IRST**
	3.776	**ISL-SMT**
	3.400	NLPR
	3.036	IBM
	2.954	ISI
	2.934	ISL-EDTRL
	2.718	ICT
	2.648	HIT
	2.570	CLIPS

Alternative Ranking

Track	score	MT_ID
CE unrestricted	**0.558**	**IRST**
	0.532	**ISL-SMT**
	0.406	NLPR
	0.326	IBM
	0.296	ISL-EDTRL
	0.286	ISI
	0.224	HIT
	0.222	ICT
	0.180	CLIPS

Reliability of Difference in MT Quality (Adequacy)

Regular Ranking

Track	score	MT_ID
CE	**3.662**	**ISL-SMT**
unrestricted	**3.526**	**IRST**
	3.254	ISL-EDTRL
	3.188	HIT
	3.082	ICT
	2.996	IBM
	2.960	CLIPS
	2.800	NLPR
	2.784	ISI

Alternative Ranking

Track	score	MT_ID
CE	0.446	ISL-SMT
unrestricted	0.394	IRST
	0.294	ISL-EDTRL
	0.258	ICT
	0.250	IBM
	0.228	NLPR
	0.226	HIT
	0.178	ISI
	0.164	CLIPS

Reliability of Difference in MT Quality (Fluency)

Regular Ranking

Track	score	MT_ID
JE unrestricted	4.308	ATR
	4.036	RWTH
	3.650	UTokyo
	2.472	CLIPS

Track	score	MT_ID
JE small	3.484	ATR
	3.480	RWTH
	3.106	IBM
	3.102	ISI

Alternative Ranking

Track	score	MT_ID
JE unrestricted	0.698	ATR
	0.608	RWTH
	0.506	UTokyo
	0.170	CLIPS

Track	score	MT_ID
JE small	0.520	ATR
	0.440	RWTH
	0.368	ISI
	0.334	IBM

Reliability of Difference in MT Quality (Adequacy)

Regular Ranking

Track	score	MT_ID
JE unrestricted	4.208	ATR
	4.066	RWTH
	3.316	UTokyo
	2.602	CLIPS

Track	score	MT_ID
JE small	3.412	RWTH
	3.086	ISI
	2.990	IBM
	1.942	ATR

Alternative Ranking

Track	score	MT_ID
JE unrestricted	0.600	ATR
	0.564	RWTH
	0.360	UTokyo
	0.120	CLIPS

Track	score	MT_ID
JE small	0.358	RWTH
	0.304	ISI
	0.262	IBM
	0.126	ATR

MT Ranking List (automatic evaluation metrics)

CE unrestricted

mWER		mPER		BLEU		NIST		GTM	
score	MT	score	MT	score	MT	score	MT	score	MT
0.379	ISL-S	0.319	ISL-S	0.524	ISL-S	9.56	ISL-S	0.748	ISL-S
0.457	IRST	0.393	IRST	0.440	IRST	7.50	ISL-E	0.684	IBM
0.525	IBM	0.427	ISL-E	0.350	IBM	7.36	IBM	0.671	IRST
0.531	ISL-E	0.422	IBM	0.311	NLPR	7.24	IRST	0.666	ISL-E
0.573	ISI	0.487	HIT	0.275	ISL-E	6.13	HIT	0.611	HIT
0.578	NLPR	0.499	ISI	0.243	HIT	6.00	CLIPS	0.602	ISI
0.594	HIT	0.531	NLPR	0.243	ISI	5.92	NLPR	0.584	CLIPS
0.658	CLIPS	0.542	CLIPS	0.162	CLIPS	5.42	ISI	0.563	NLPR
0.846	ICT	0.765	ICT	0.079	ICT	3.64	ICT	0.386	ICT

MT Ranking List (automatic evaluation metrics)

JE small

mWER		mPER		BLEU		NIST		GTM	
score	MT	score	MT	score	MT	score	MT	score	MT
0.418	RWTH	0.337	RWTH	0.453	RWTH	9.49	RWTH	0.764	RWTH
0.484	ISI	0.379	ISI	0.400	ISI	8.46	ISI	0.732	ISI
0.527	IBM	0.430	IBM	0.366	IBM	7.97	IBM	0.698	IBM
0.614	ATR	0.570	ATR	0.364	ATR	3.41	ATR	0.539	ATR

Correlation Coefficients

Alternative Ranking vs. Automatic Ranking (partial MT systems)

CE+JE	**mWER**	mPER	BLEU	NIST	GTM
Fluency	**-0.9936**	-0.9850	0.9839	0.9497	0.9206
Adequacy	**-0.9980**	-0.9946	0.9969	0.9703	0.9729

partial = MT systems whose score difference is at least twice at large as the word error rates