JST 独立行政法人 科学技術振興機構 Japan Science and Technology Agency

NICT 独立行政法人 情報通信研究機構 National Institute of Information and Communications Technology

京都大学 KYOTO UNIVERSITY

Overview of WAT2014

Toshiaki Nakazawa Hideya Mino
Isao Goto Sadao Kurohashi
Eiichiro Sumita

4th October, 2014 @ WAT2014

http://lotus.kuee.kyoto-u.ac.jp/WAT/

WAT 2014

The 1st Workshop on Asian Translation

- MT evaluation campaign focusing on Asian languages (Japanese, Chinese and English for this time)
- The first evaluation workshop that uses scientific papers as a domain and Japanese-Chinese as a language pair
- Paragraph-based test set
  - investigate the viability of the context-aware MT
- All the data including test set are OPEN
  - contribute to continuous evolution of MT research by freely distributing the data (like PennTreebank sec. 23)

Automatic Evaluation in WAT2014

- Prepared an automatic evaluation server
- BLEU, RIBES
- several word segmentation tools

See Evaluation Results:

http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/

Submit Your Translations (need FREE registration):

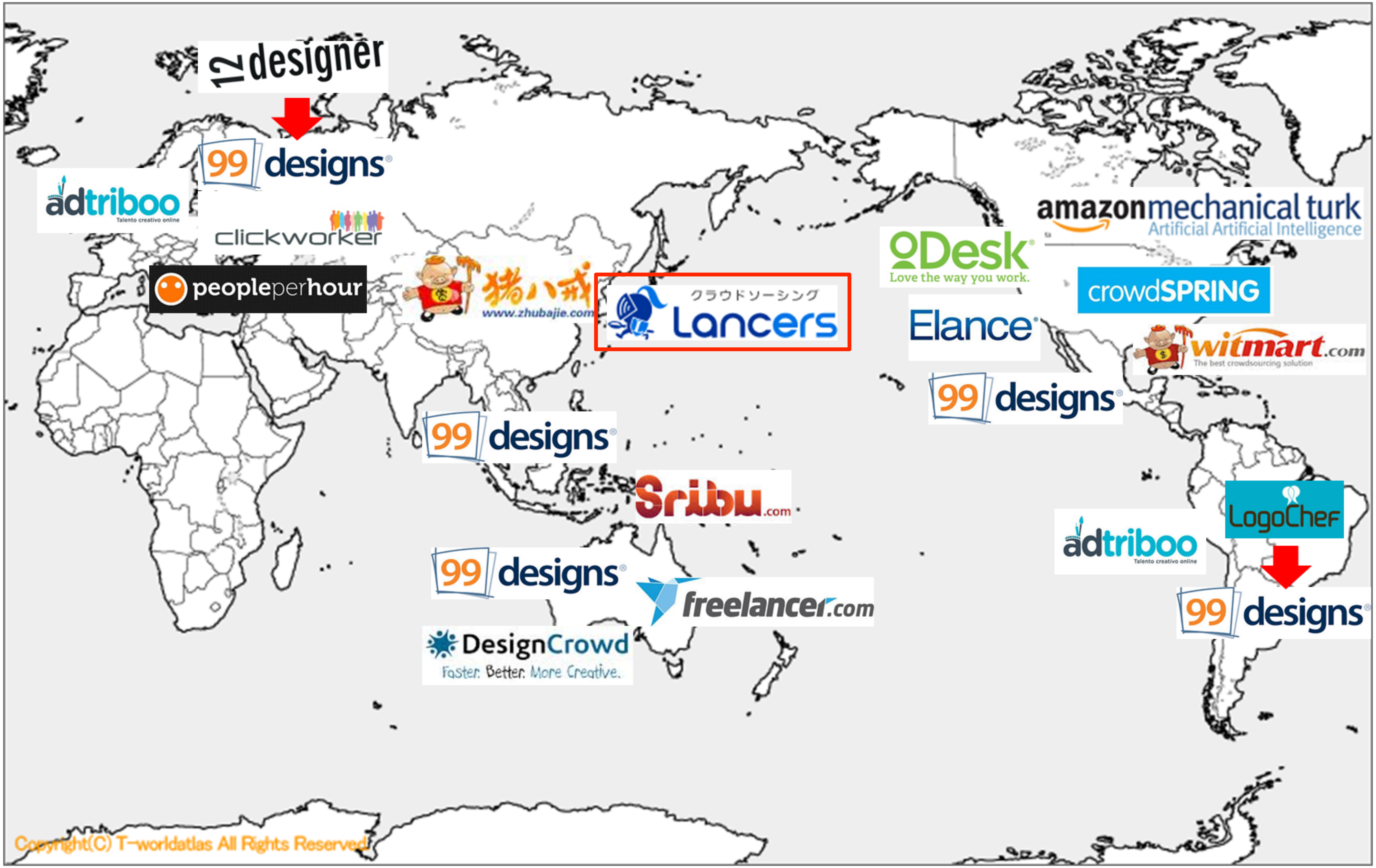
http://lotus.kuee.kyoto-u.ac.jp/WAT/submission/

Human Evaluation of MT

- Costs a lot of money and time
- Unstable results caused by the different criteria of each evaluators
- Many measures
  - Adequacy/Fluency
  - Ranking
  - Acceptability (NTCIR)
  - Patent Examination Evaluation (NTCIR)

Human Evaluation in WAT2014

- Costs a lot of money and time
  - using crowdsourcing to reduce them
- Unstable results caused by the different criteria of each evaluators
  - alleviate the divergence by voting
- Many measures
  - HUMAN score


12designer
99designs
adtriboo Talento creativo online
clickworker
peopleperhour
猪八戒 www.zhubajie.com
クラウドソーシング Lancers
oDesk Love the way you work.
amazonmechanical turk Artificial Artificial Intelligence
crowdSPRING
Elance
witmart.com The best crowdsourcing solution
99designs
99designs
Sribu.com
99designs
freelancer.com
DesignCrowd Faster. Better. More Creative.
adtriboo Talento creativo online
LogoChef
99designs
Copyright(C) T-worldatlas All Rights Reserved

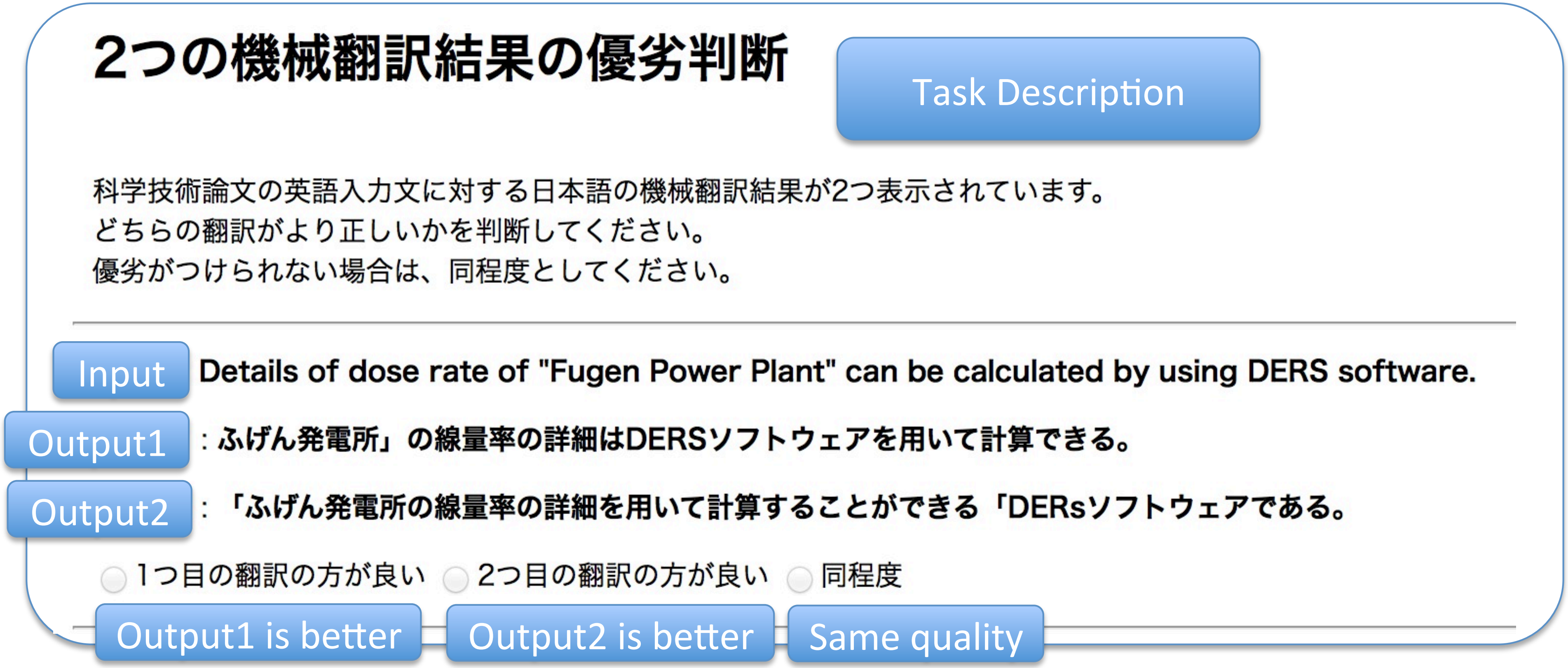
http://www.crowdinfo.jp/2014/02/02/world-crowdsourcing-service/

Human Evaluation in WAT2014

Phrase-based SMT

- Pairwise evaluation compared to the baseline
  - reduce the number of sentences to be evaluated
  - enable the evaluation of new translation results after the workshop
- 400 sentences selected from the test set by document-based sampling
- Reference translations are not shown

Sample of the Task



2つの機械翻訳結果の優劣判断

科学技術論文の英語入力文に対する日本語の機械翻訳結果が2つ表示されています。
どちらの翻訳がより正しいかを判断してください。
優劣がつけられない場合は、同程度としてください。

Details of dose rate of "Fugen Power Plant" can be calculated by using DERS software.

：ふげん発電所」の線量率の詳細はDERSソフトウェアを用いて計算できる。

：「ふげん発電所の線量率の詳細を用いて計算することができる「DERsソフトウェアである。

○ 1つ目の翻訳の方が良い ○ 2つ目の翻訳の方が良い ○ 同程度

- The order of the baseline and subject outputs are at random

Pairwise Evaluation by Voting

- To guarantee the quality of the evaluation, each pair is evaluated 3 different workers
- The evaluation result of each pair is decided by the voting of 3 judgments
  - e.g. MT A vs. MT B

Worker 1	A	A	A	A	A	A	Tie	Tie	Tie	B
Worker 2	A	A	A	Tie	Tie	B	Tie	Tie	B	B
Worker 3	A	Tie	B	Tie	B	B	Tie	B	B	B
Decision	A	A	A	A	Tie	B	Tie	B	B	B

Comparison to the Baseline

BASELINE
(Phrase-based SMT)

BASELINE-MT-1 ○		SYSTEM1-MT-1 ✕
BASELINE-MT-2		SYSTEM1-MT-2
BASELINE-MT-3 ✕		SYSTEM1-MT-3 ○
BASELINE-MT-4 ✕		SYSTEM1-MT-4 ○
BASELINE-MT-5 ✕	VS.	SYSTEM1-MT-5 ○
BASELINE-MT-6		SYSTEM1-MT-6
BASELINE-MT-7 ✕		SYSTEM1-MT-7 ○
BASELINE-MT-8 ○		SYSTEM1-MT-8 ✕
BASELINE-MT-9		SYSTEM1-MT-9
BASELINE-MT-10 ✕		SYSTEM1-MT-10 ○

SYSTEM1

5 wins, 2 losses, 3 ties

Human Evaluation Score

- Suppose *W* = # of wins, *L* = # of losses and *T* = # of ties, the HUMAN score is

$$HUMAN = 100 \times \frac{W - L}{W + L + T}$$

e.g. sample of the previous page

$$100 \times \frac{5 - 2}{5 + 2 + 3} = 30$$

- Estimate the confidence interval by bootstrap resampling [Koehn, 2004]
  - calculate the human evaluation score on 300 sentences randomly sampled from 400 sentences
  - iterate calculation 1000 times
  - sort the 1000 scores and discard top and bottom 25 scores to get the 95% confidence interval

Cost of Crowdsourcing

- One judgment by one worker costs 5 JPY
- Each sentence requires 3 judgments
- We have 400 sentences for the human evaluation

- One evaluation of one submission costs

5 × 3 × 400 = 6,000 JPY

OFFICIAL HUMAN EVALUATION RESULTS

Participants List

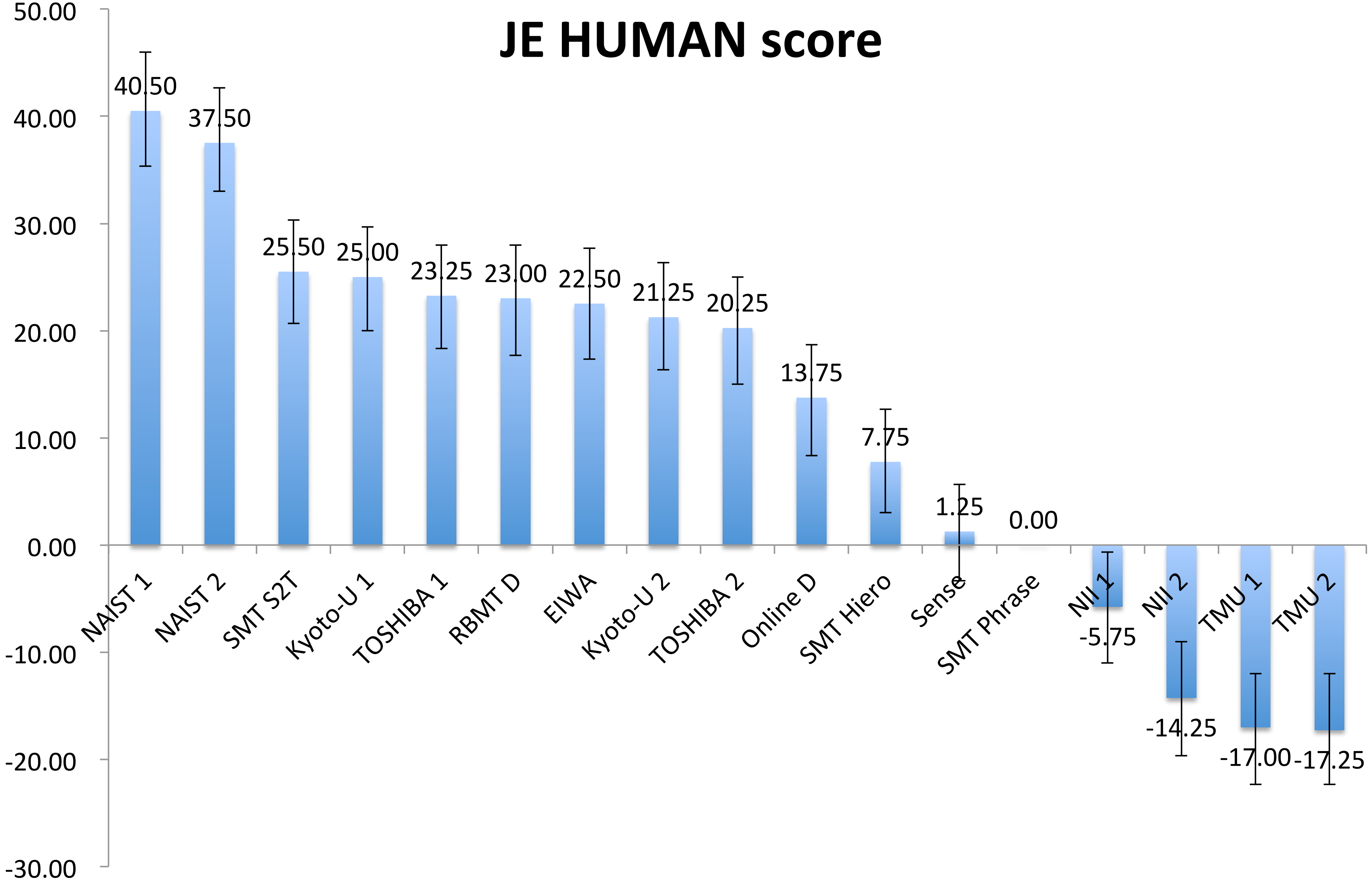
Team ID	J->E	E->J	J->C	C->J
NAIST	✓	✓	✓	✓
EIWA	✓			✓
Kyoto-U	✓	✓	✓	✓
WEBLIO-EJ1		✓		
TMU	✓			
BJTUNLP			✓	

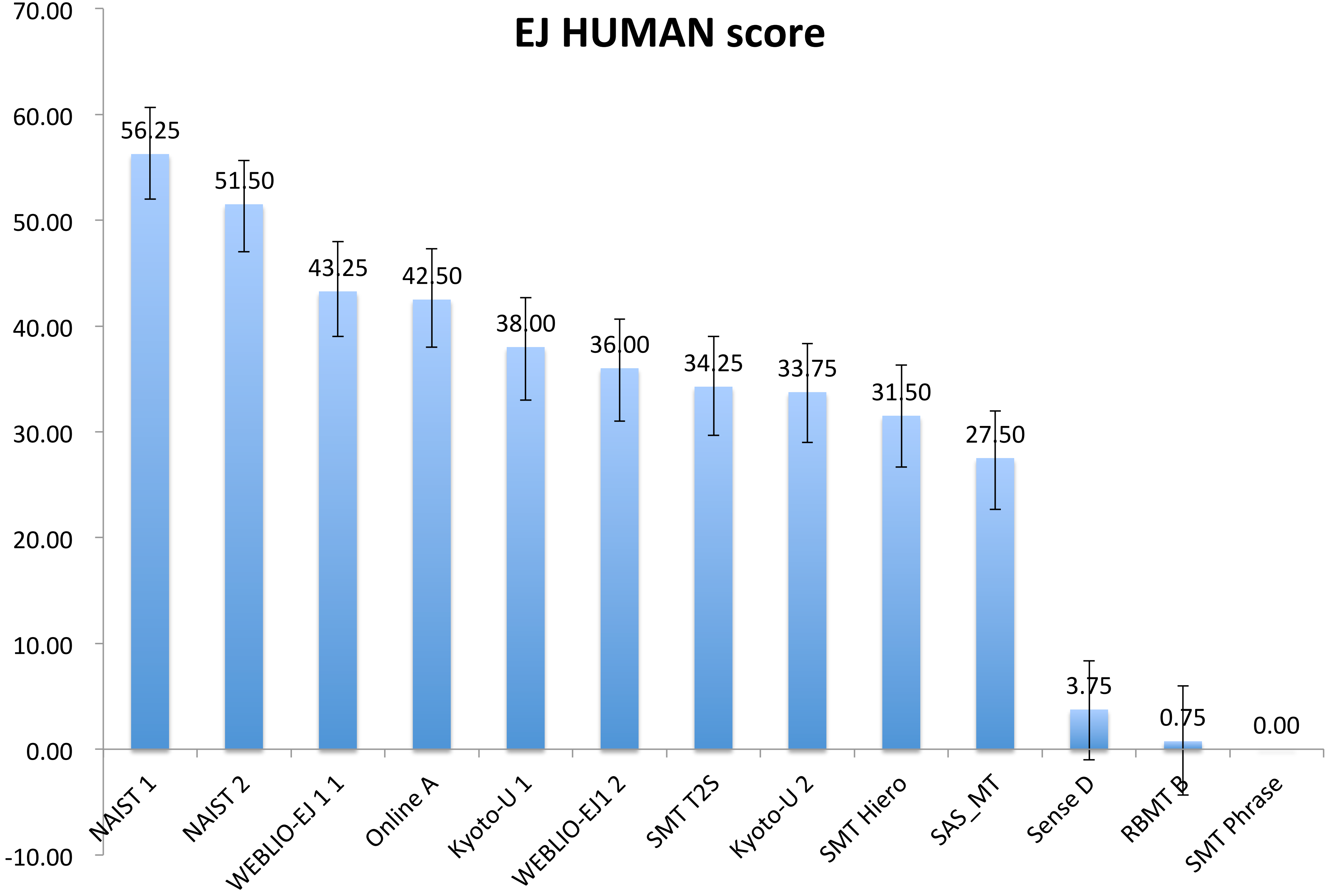
Team ID	J->E	E->J	J->C	C->J
NII	✓			
SAS_MT		✓		✓
Sense	✓	✓	✓	✓
NICT			✓	
TOSHIBA	✓		✓	
WASUIPS			✓ *	✓ *

* Only submitted to the automatic evaluations

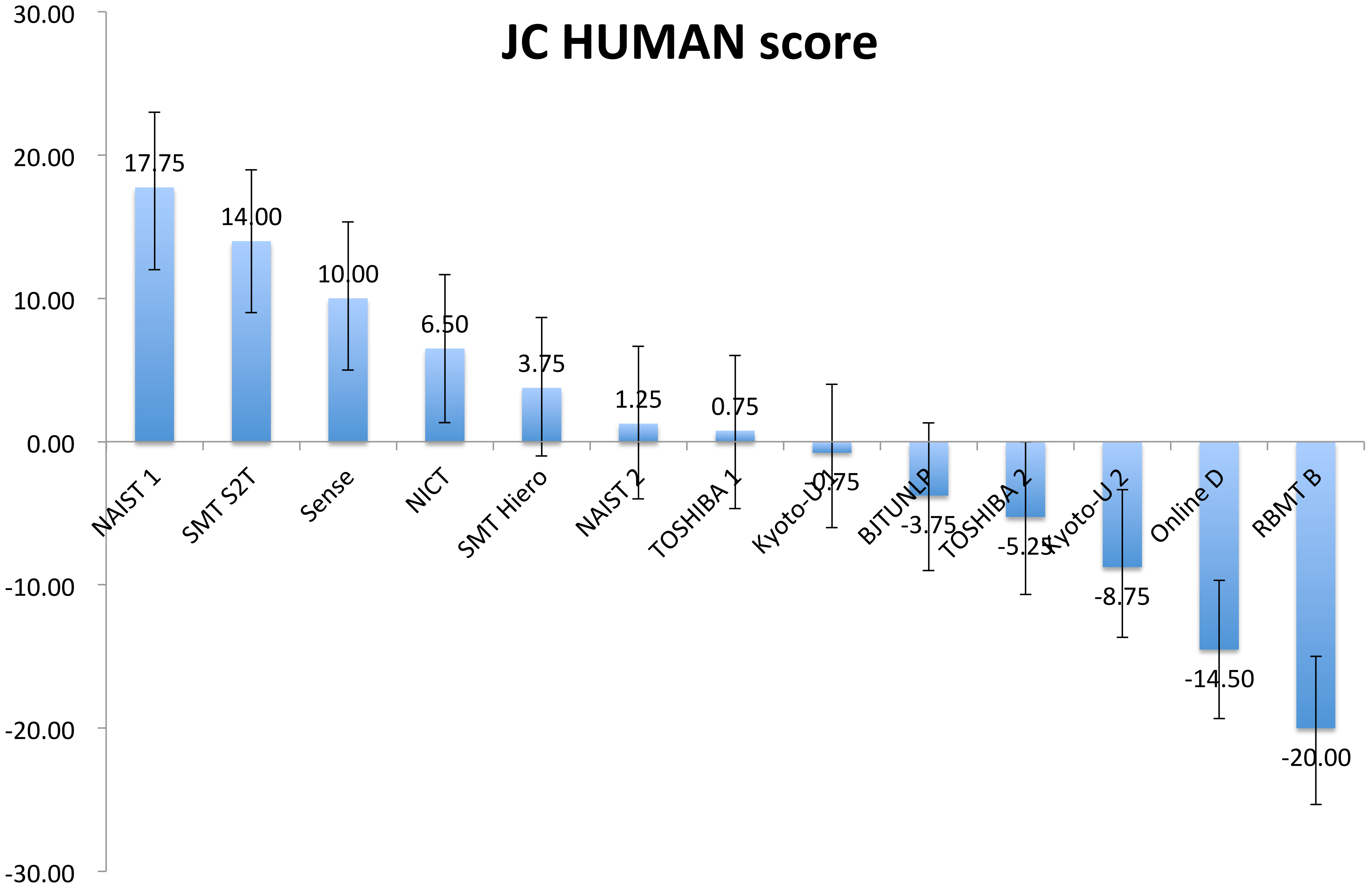
Company

Outside Japan

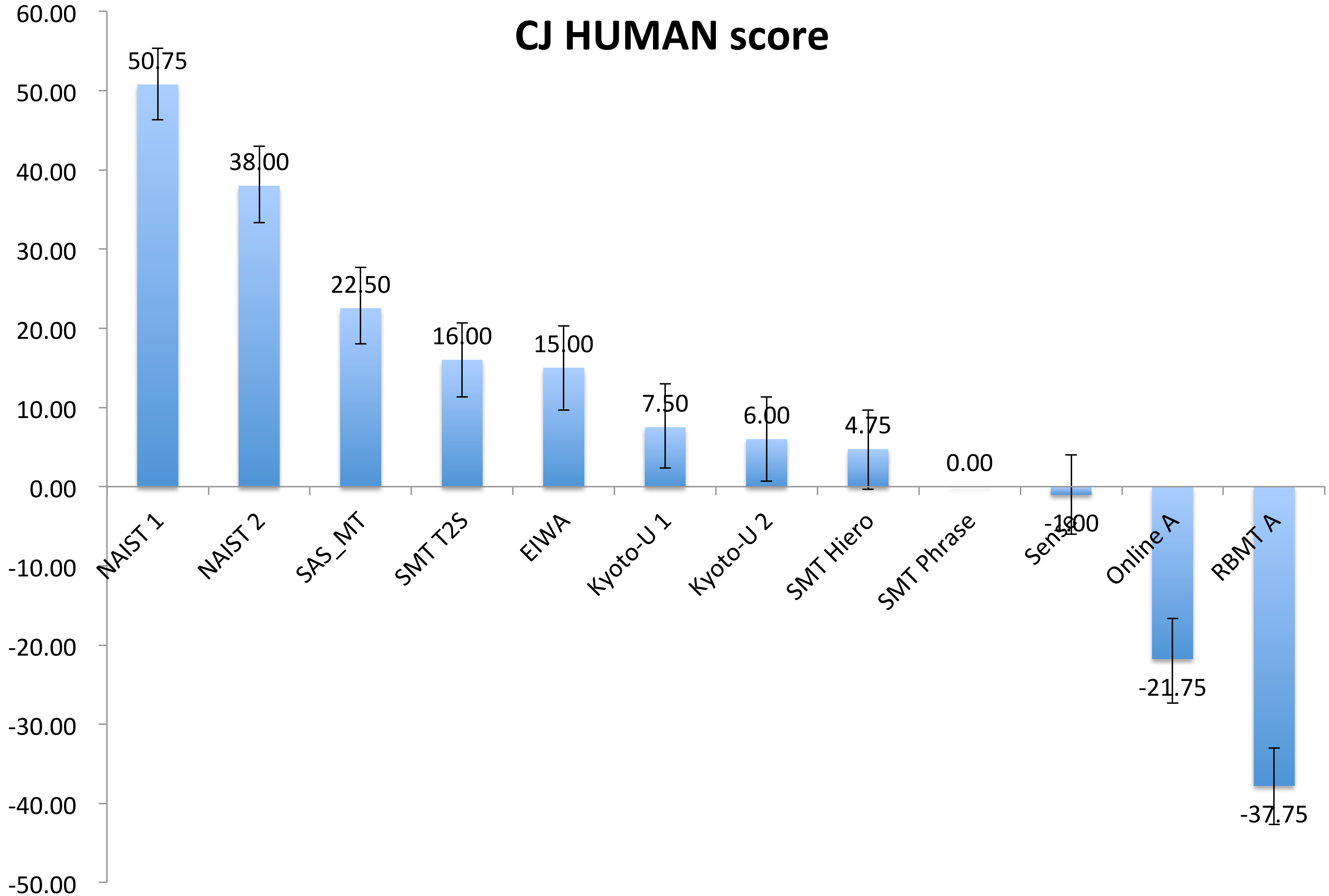

JE HUMAN score
50.00
40.00
30.00
20.00
10.00
0.00
-10.00
-20.00
-30.00
40.50
37.50
25.50
25.00
23.25
23.00
22.50
21.25
20.25
13.75
7.75
1.25
0.00
-5.75
-14.25
-17.00
-17.25
NAIST 1
NAIST 2
SMT S2T
Kyoto-U 1
TOSHIBA 1
RBMT D
EIWA
Kyoto-U 2
TOSHIBA 2
Online D
SMT Hiero
Sense
SMT Phrase
NII 1
NII 2
TMU 1
TMU 2


EJ HUMAN score
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
-10.00
56.25
51.50
43.25
42.50
38.00
36.00
34.25
33.75
31.50
27.50
3.75
0.75
0.00
NAIST 1
NAIST 2
WEBLIO-EJ 1 1
Online A
Kyoto-U 1
WEBLIO-EJ1 2
SMT T2S
Kyoto-U 2
SMT Hiero
SAS_MT
Sense D
RBMT B
SMT Phrase

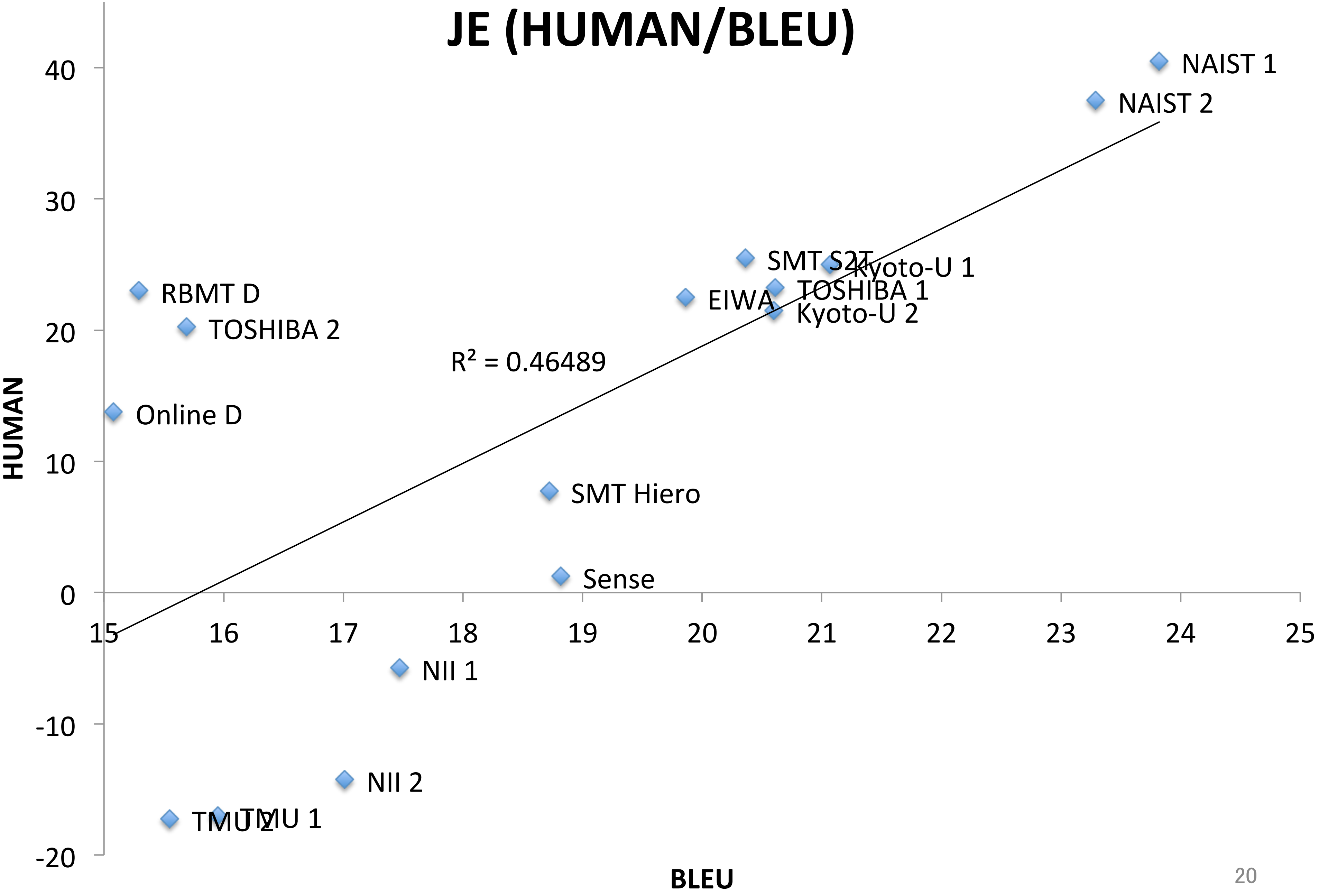
JC HUMAN score

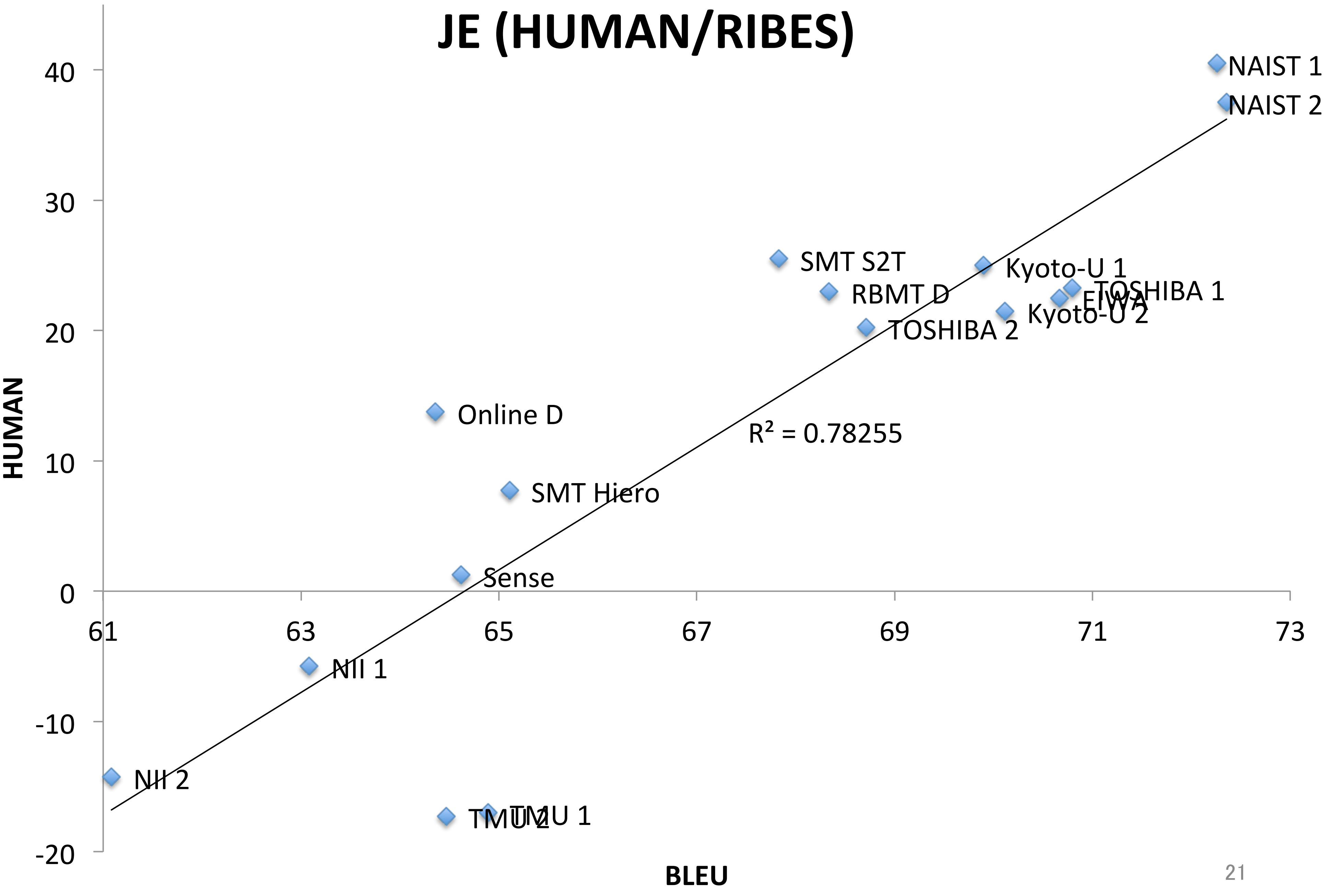


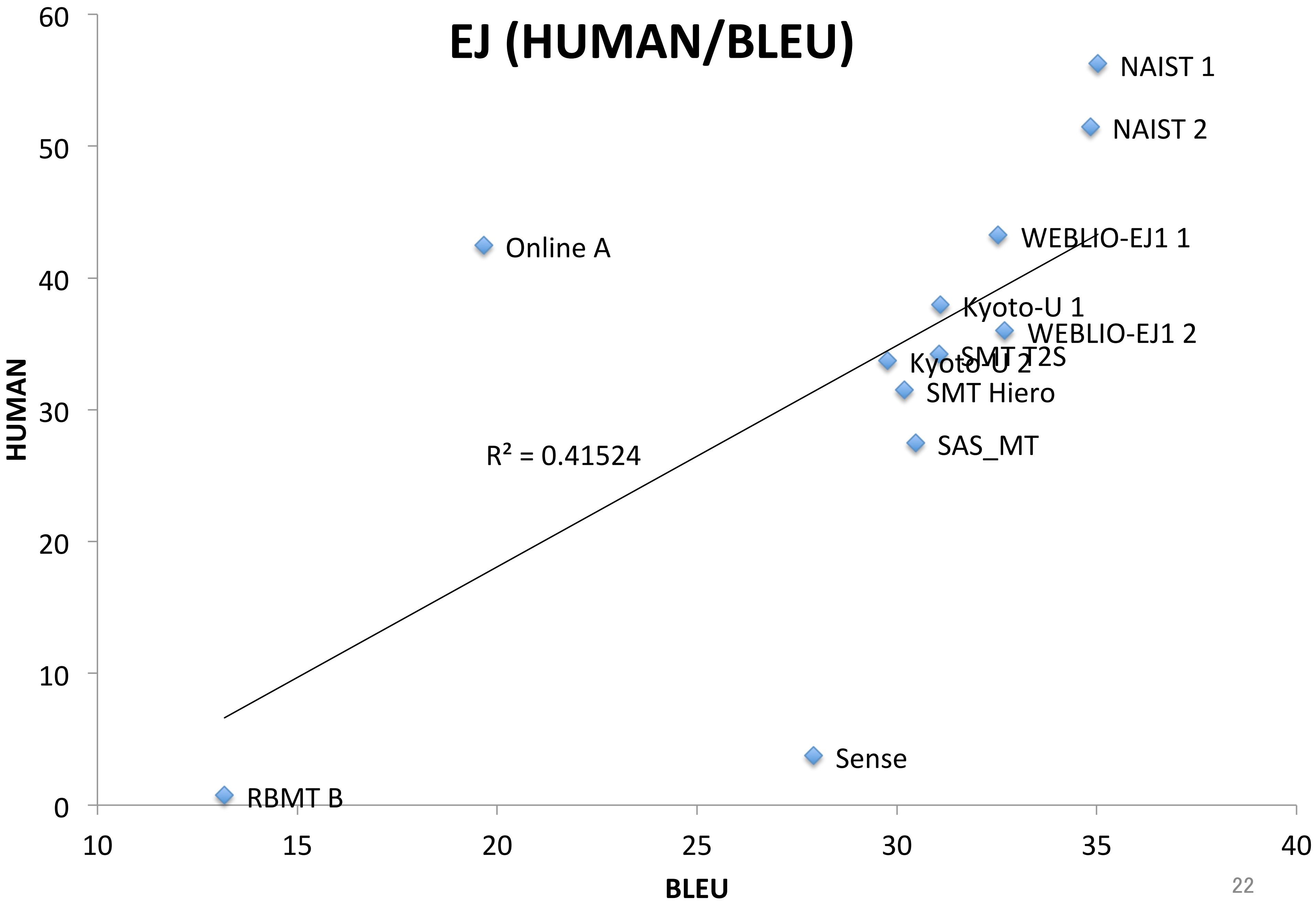
CJ HUMAN score

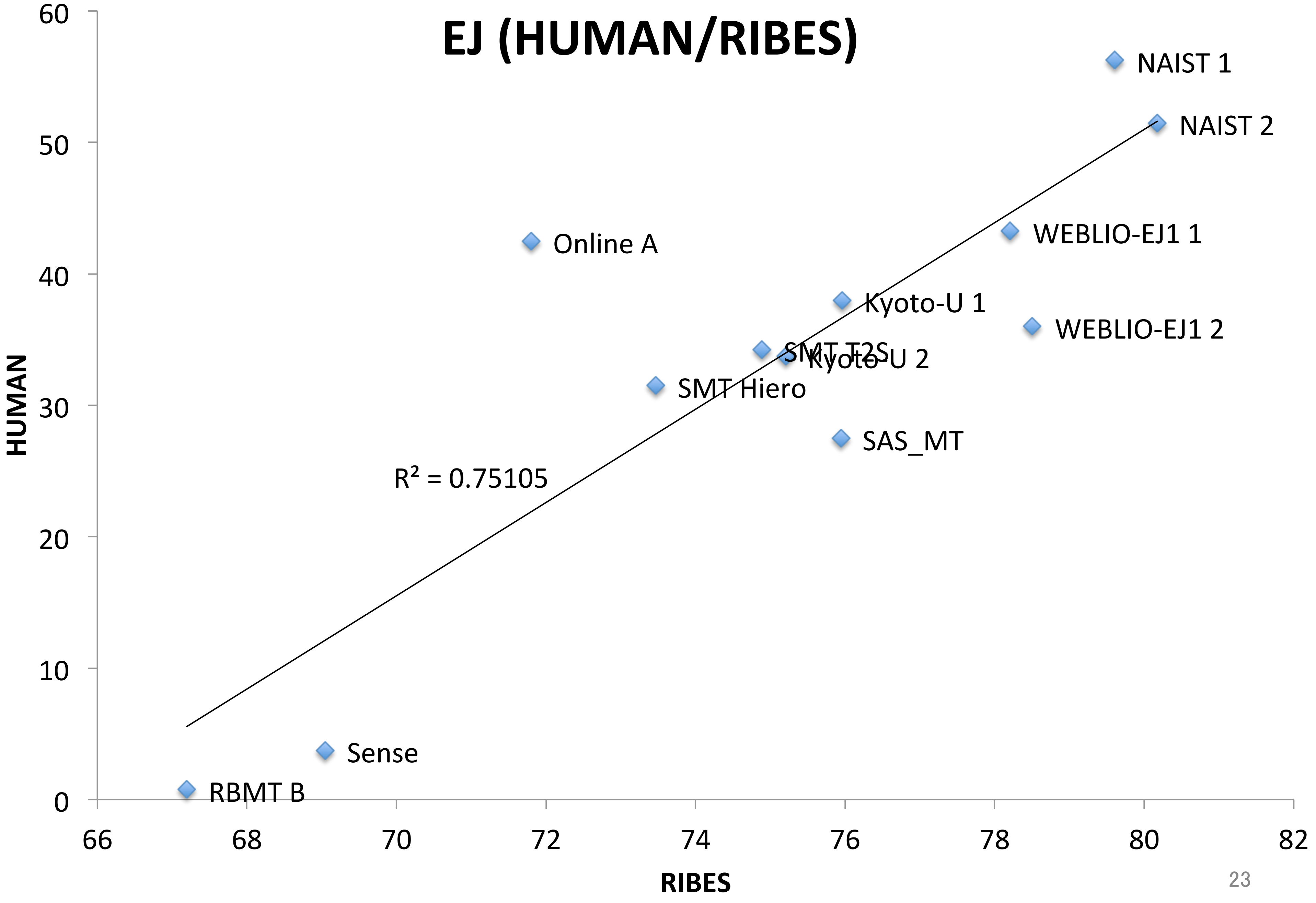


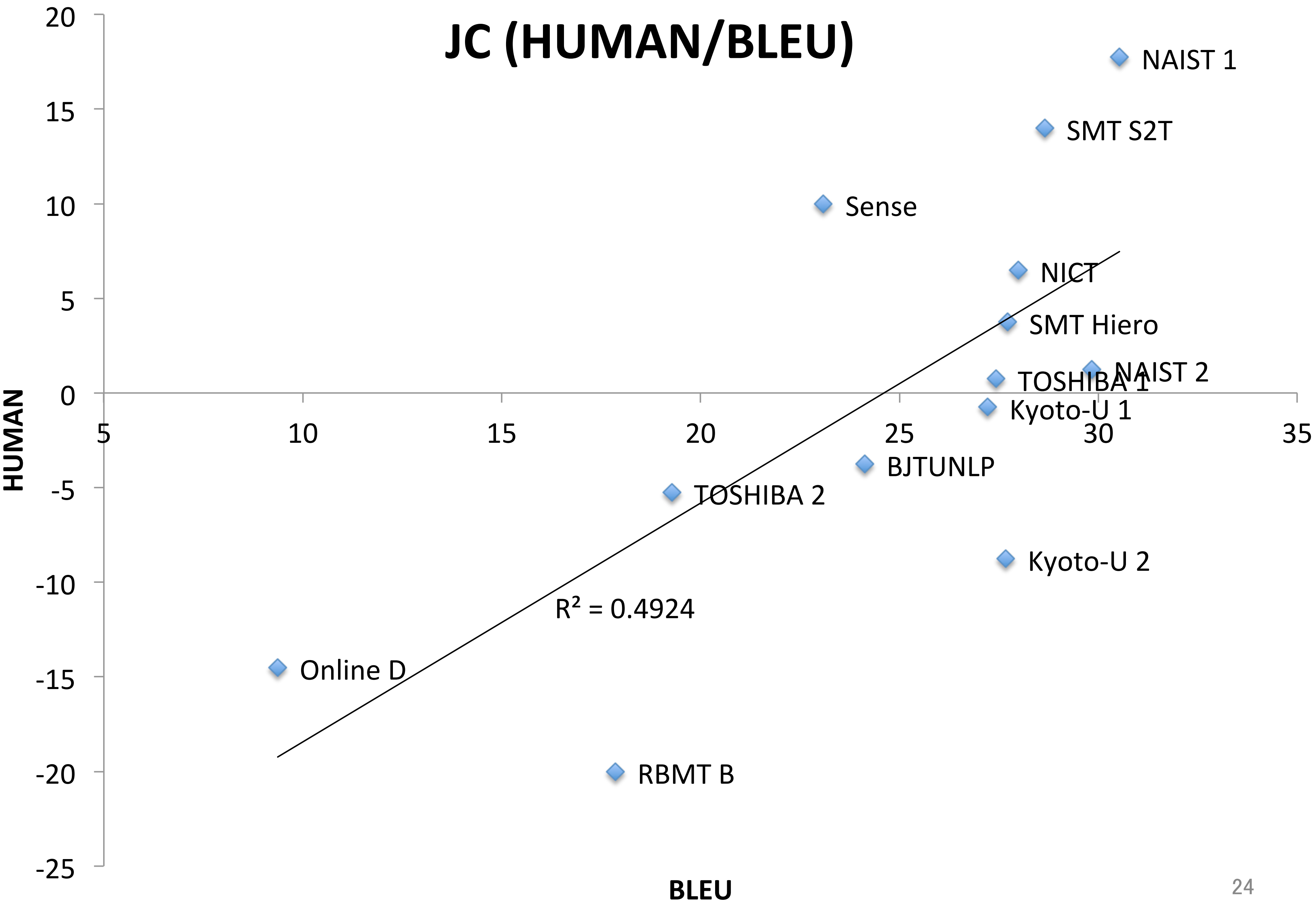
CORRELATION BETWEEN BLEU/ RIBES AND HUMAN SCORE


JE (HUMAN/BLEU)
HUMAN
BLEU
40
30
20
10
0
-10
-20
15
16
17
18
19
20
21
22
23
24
25
NAIST 1
NAIST 2
SMT S2T
Kyoto-U 1
TOSHIBA 1
Kyoto-U 2
EIWA
RBMT D
TOSHIBA 2
R² = 0.46489
Online D
SMT Hiero
Sense
NII 1
NII 2
TMU 2
TMU 1

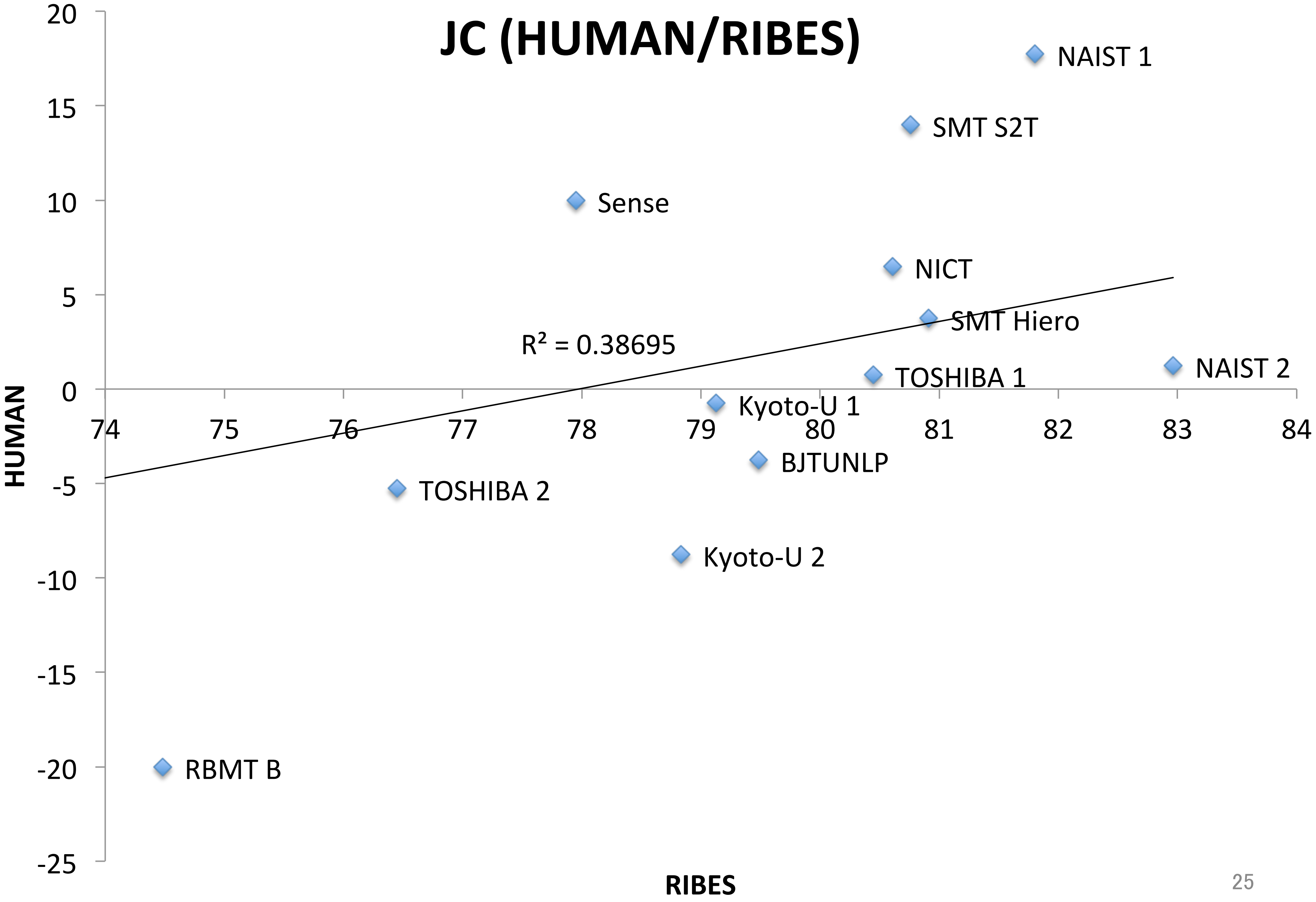

JE (HUMAN/RIBES)
HUMAN
BLEU
40
30
20
10
0
-10
-20
61
63
65
67
69
71
73
NAIST 1
NAIST 2
SMT S2T
Kyoto-U 1
TOSHIBA 1
RBMT D
EIWA
Kyoto-U 2
TOSHIBA 2
Online D
R² = 0.78255
SMT Hiero
Sense
NII 1
NII 2
TMU 2
TMU 1

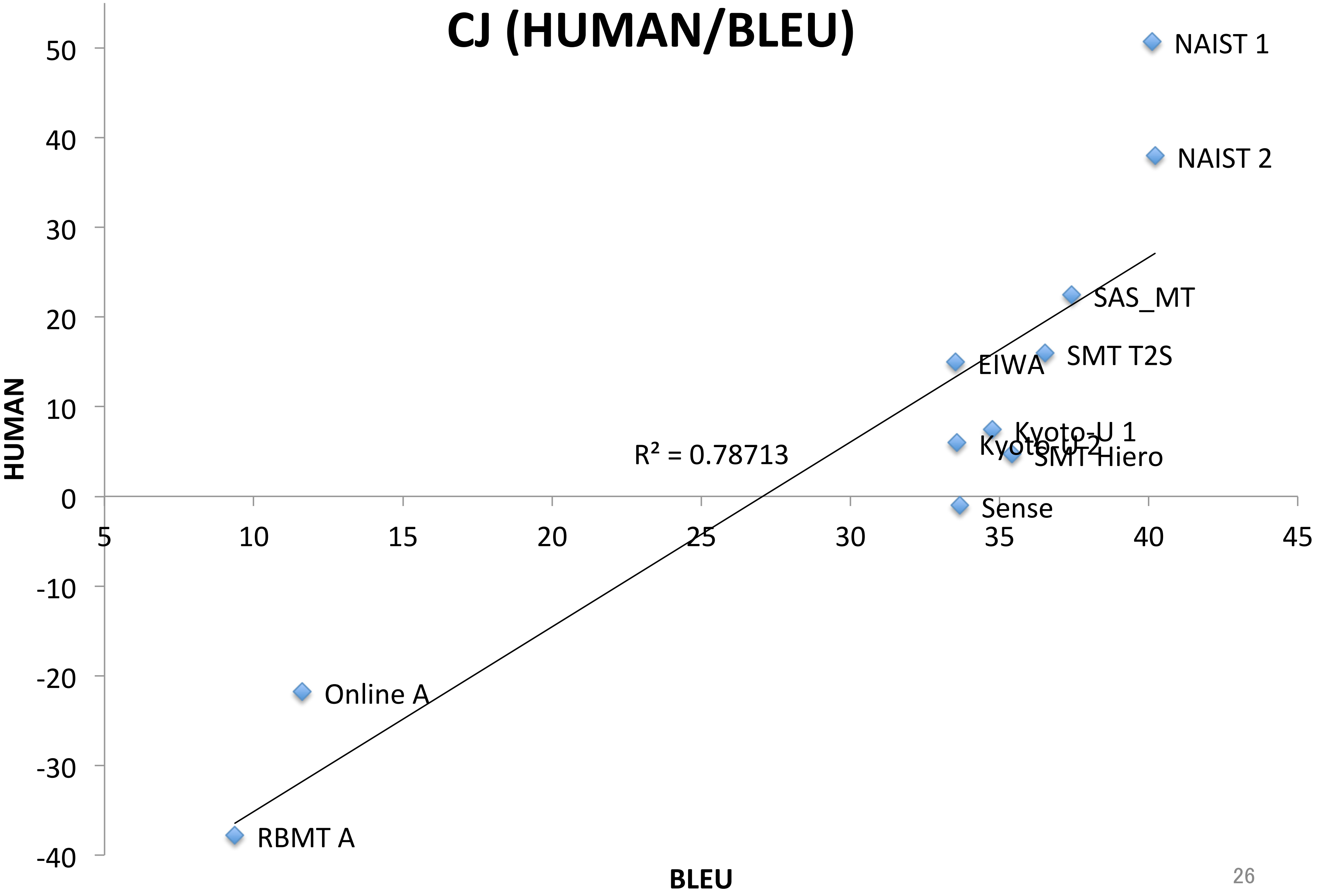

EJ (HUMAN/BLEU)
HUMAN
BLEU
0
10
20
30
40
50
60
10
15
20
25
30
35
40
NAIST 1
NAIST 2
Online A
WEBLIO-EJ1 1
Kyoto-U 1
WEBLIO-EJ1 2
SMT T2S
Kyoto-U 2
SMT Hiero
SAS_MT
R² = 0.41524
Sense
RBMT B


EJ (HUMAN/RIBES)
60
50
40
30
20
10
0
HUMAN
66
68
70
72
74
76
78
80
82
RIBES
NAIST 1
NAIST 2
WEBLIO-EJ1 1
Online A
Kyoto-U 1
WEBLIO-EJ1 2
SMT T2S
Kyoto-U 2
SMT Hiero
SAS_MT
R² = 0.75105
Sense
RBMT B

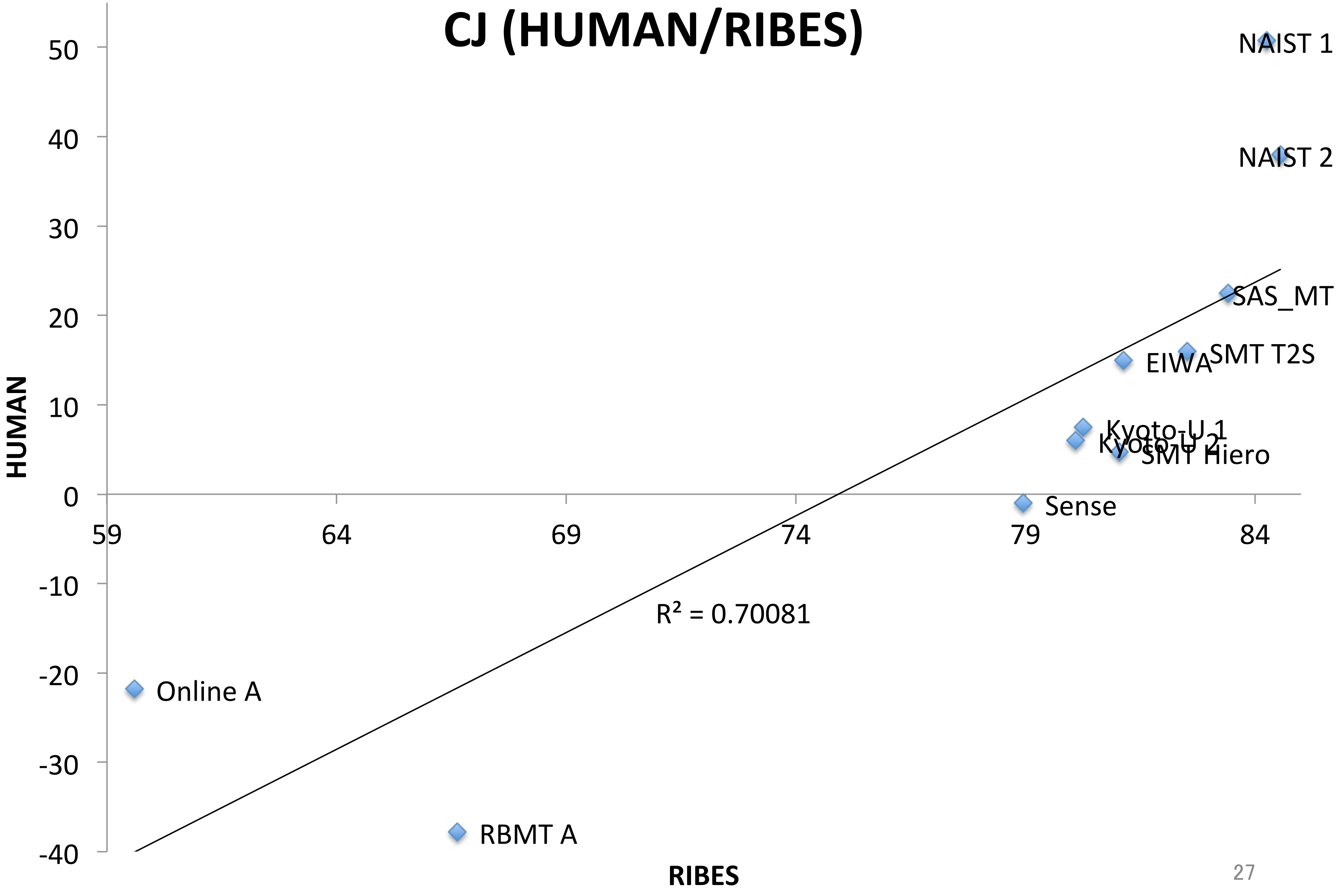
# JC (HUMAN/BLEU)

JC (HUMAN/RIBES)




CJ (HUMAN/BLEU)
HUMAN
BLEU
50
40
30
20
10
0
-10
-20
-30
-40
5
10
15
20
25
30
35
40
45
NAIST 1
NAIST 2
SAS_MT
SMT T2S
EIWA
Kyoto-U 1
Kyoto-U 2
SMT Hiero
Sense
Online A
RBMT A
R² = 0.78713

CJ (HUMAN/RIBES)



Better Correlation among Corpus-based MT?

- Less correlations between automatic and human evaluations for RBMT and Online

	All systems	Corpus-based only
JE BLEU	0.46489	0.95098
JE RIBES	0.78255	0.83691
EJ BLEU	0.41524	0.84418
EJ RIBES	0.75105	0.85730
JC BLEU	0.49240	0.07937
JC RIBES	0.38695	0.10198
CJ BLEU	0.78713	0.82592
CJ RIBES	0.70081	0.83209

* R^2 values

INTER ANNOTATOR AGREEMENT

Inter Annotator Agreement

JE	
System ID	Kappa
NAIST 1	0.162
NAIST 2	0.047
SMT S2T	0.099
Kyoto-U 1	0.070
TOSHIBA 1	0.098
RBMT D	0.075
EIWA	0.083
Kyoto-U 2	0.139
TOSHIBA 2	0.078
Online D	0.055
SMT Hiero	0.119
Sense	0.245
NII 1	0.119
NII 2	0.086
TMU 1	0.091
TMU 2	0.136
Ave.	**0.106**

EJ	
System ID	Kappa
NAIST 1	0.280
NAIST 2	0.250
WEBLIO-EJ1 1	0.238
Online A	0.219
Kyoto-U 1	0.216
WEBLIO-EJ1 2	0.240
SMT T2S	0.240
Kyoto-U 2	0.229
SMT Hiero	0.277
SAS_MT	0.248
Sense	0.395
RBMT B	0.217
Ave.	**0.254**

JC	
System ID	Kappa
NAIST 1	0.077
SMT S2T	0.069
Sense	0.087
NICT	0.066
SMT Hiero	0.202
NAIST 2	0.093
TOSHIBA 1	0.089
Kyoto-U 1	0.091
BJTUNLP	0.198
TOSHIBA 2	0.066
Kyoto-U 2	0.163
Online D	0.035
RBMT B	0.083
Ave.	**0.101**

CJ	
System ID	Kappa
NAIST 1	0.168
NAIST 2	0.203
SAS_MT	0.167
SMT T2S	0.236
EIWA	0.175
Kyoto-U 1	0.199
Kyoto-U 2	0.180
SMT Hiero	0.274
Sense	0.228
Online A	0.239
RBMT A	0.130
Ave.	**0.200**

* Fleiss’s Kappa values

Inter Annotator Agreement

JE	
System ID	**Kappa**
NAIST 1	0.162
NAIST 2	0.047
SMT S2T	0.099
Kyoto-U 1	0.070
TOSHIBA 1	0.098
RBMT D	0.075
Online D	0.055
SMT Hiero	0.119
Sense	0.245
NII 1	0.119
NII 2	0.086
TMU 1	0.091
TMU 2	0.136
Ave.	**0.106**

EJ	
System ID	**Kappa**
NAIST 1	0.280
NAIST 2	0.250
WEBLIO-EJ1 1	0.238
Online A	0.219
Kyoto-U 1	0.216
WEBLIO-EJ1 2	0.240
SAS_MT	0.248
Sense	0.395
RBMT B	0.217
Ave.	**0.254**

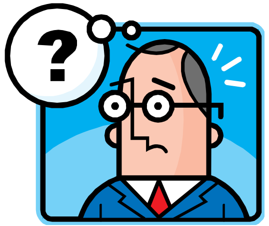
JC	
System ID	**Kappa**
NAIST 1	0.077
SMT S2T	0.069
Sense	0.087
NICT	0.066
SMT Hiero	0.202
NAIST 3	0.093
TOSHIBA 2	0.066
Kyoto-U 2	0.163
Online D	0.035
RBMT B	0.083
Ave.	**0.101**

CJ	
System ID	**Kappa**
NAIST 1	0.168
NAIST 2	0.203
SAS_MT	0.167
SMT T2S	0.236
EIWA	0.175
Kyoto-U 1	[illegible]
Online A	0.239
RBMT A	0.130
Ave.	**0.200**

X->J evaluations are easier than J->X evaluations
(because the workers are almost Japanese?)

* Fleiss's Kappa values

Case Study



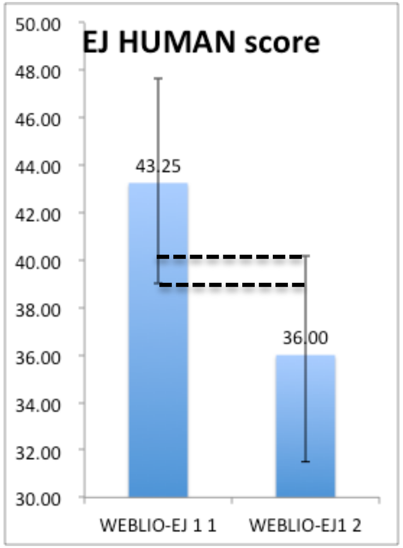
Submission ID	BLEU	RIBES	HUMAN	Description
WEBLIO-EJ1 1	32.53	0.782	43.25	w/o forest input
WEBLIO-EJ1 2	32.69	0.785	36.00	w/ forest input

BASELINE
(WEBLIO-EJ1 1)

VS.

WEBLIO-EJ1 2

HUMAN	Kappa
2.50 ± 4.17	0.528



- No significant difference
- Much higher Kappa value
  - similar outputs can be easily and faithfully judged

Conclusion

- 12 participants for the evaluation task
  - including 3 companies and 3 teams outside Japan
- Human evaluation using crowdsourcing
- NAIST team achieved the best results for all the subtasks (congratulations!!)
- Shared the findings of MT for scientific papers
  - http://lotus.kuee.kyoto-u.ac.jp/WAT/papers/papers-2014.html

Future Perspective

- Automatic evaluation server will keep running even after the workshop
  - promote continuous evolution of MT research
- WAT will be held annually
  - include more languages, domains…
- Let's share your resource!
  - monolingual/bilingual corpora, dictionaries…

Future Perspective

- Need more investigation to acquire reliable human evaluation results at low cost
- Need to find a better way to compare two systems efficiently and reliably
- Discuss the importance of both sentence internal/external information

Thank you very much
for attending WAT2014

Stay tuned for
the next WAT workshop!

Future Perspective

- WAT will be held annually
  - include more languages, domains…
- Let's share your resource!
  - monolingual/bilingual corpora, dictionaries…
- Discuss the importance of both sentence internal/external information