



Overview of the 2nd Workshop on Asian Translation

Toshiaki Nakazawa Hideya Mino Isao Goto
Graham Neubig Sadao Kurohashi Eiichiro Sumita

The Features of WAT2015

- MT evaluation campaign focusing on **Asian languages** (Japanese, Chinese, Korean and English for this time)
- The **first** evaluation for **Chinese-Japanese** and **Korean-Japanese patent** translation.
- **Professional translators** evaluate the outputs of the top systems based on JPO adequacy.

Comparison of WAT 2014 and 2015

	WAT2014	WAT2015
Task	Scientific paper (ASPEC) <ul style="list-style-type: none">• Japanese ↔ English• Chinese ↔ Japanese	Scientific paper (ASPEC) <ul style="list-style-type: none">• Japanese ↔ English• Chinese ↔ Japanese JPO Patent corpus (JPC)  <ul style="list-style-type: none">• Chinese → Japanese• Korean → Japanese
Evaluation	Automatic Evaluation Human Evaluation <ul style="list-style-type: none">• Pairwise Crowdsourcing	Automatic Evaluation Human Evaluation <ul style="list-style-type: none">• Pairwise Crowdsourcing• JPO Adequacy (Top 3 systems)
Number of participants	12	12 

Notable Findings at WAT2015

- **Neural Network** based re-ranking was effective for human evaluation (NAIST, Kyoto-U, naver)
- The top **SMT** outperformed **RBMT** for **Chinese-Japanese** and **Korean-Japanese patent** translation
- **Korean-Japanese** patent translation achieved **high scores** for automatic and human evaluations.
- A **problem of automatic evaluation** was found in the **Korean-Japanese** translation.

Baseline Systems

System ID	System	Type	ASPEC				JPC	
			JE	EJ	JC	CJ	CJ	KJ
SMT Phrase	Moses' Phrase-based SMT	SMT	✓	✓	✓	✓	✓	✓
SMT Hiero	Moses' Hierarchical Phrase-based SMT		✓	✓	✓	✓	✓	✓
SMT S2T	Moses' String-to-Tree Syntax-based SMT and Berkeley parser		✓		✓			
SMT T2S	Moses' Tree-to-String Syntax-based SMT and Berkeley parser			✓		✓	✓	
RBMT X	The Honyaku V15 (Commercial system)	RBMT	✓	✓				
RBMT X	ATLAS V14 (Commercial system)		✓	✓				
RBMT X	PAT-Transer 2009 (Commercial system)		✓	✓				
RBMT X	J-Beijing 7 (Commercial system)				✓	✓	✓	
RBMT X	Hohrai 2011 (Commercial system)				✓	✓	✓	
RBMT X	J Soul 9 (Commercial system)							✓
RBMT X	Korai 2011 (Commercial system)							✓
Online X	Google translate (August, 2015)	(SMT)	✓	✓	✓	✓	✓	✓
Online X	Bing translator (August and September, 2015)	(SMT)	✓	✓	✓	✓	✓	✓

- **4** types of **SMT** systems, **7** **RBMT** systems, and **2** **online** systems
- The SYSTEM-IDs of the commercial RBMT and online systems are anonymized.
- The translation procedures for the SMT systems were published on the WAT web site.⁵

Evaluation Methods

Automatic Evaluation

- BLEU, RIBES
- The automatic evaluation server for WAT accepts translations any time.

Pairwise Crowdsourcing Evaluation

- Sentence-level evaluation comparing to the baseline Phrase-based SMT output

System output

VS

Baseline output

- The Lancers crowdsourcing platform (the same as that at WAT2014)
- Evaluators judge win (1), loss (-1), or tie (0)
- **5 evaluators** assessed for each sentence.
- The final judgment for each sentence is decided by voting based on the sum of judgments:
 - Win: $\text{sum} \geq 2$
 - Loss: $\text{sum} \leq -2$
 - Tie: otherwise

This voting criterion is **different from** that of **WAT2014**.

The Crowd scores between WAT2014 and WAT2015 **cannot** be compared.

Crowd score

400 sentences were evaluated for each system.

$$\text{Crowd score} = 100 \times \frac{W - L}{W + L + T}$$

(The score range is -100 to 100)

W: the number of wins compared to the baseline

L: the number of losses compared to the baseline

T: the number of ties

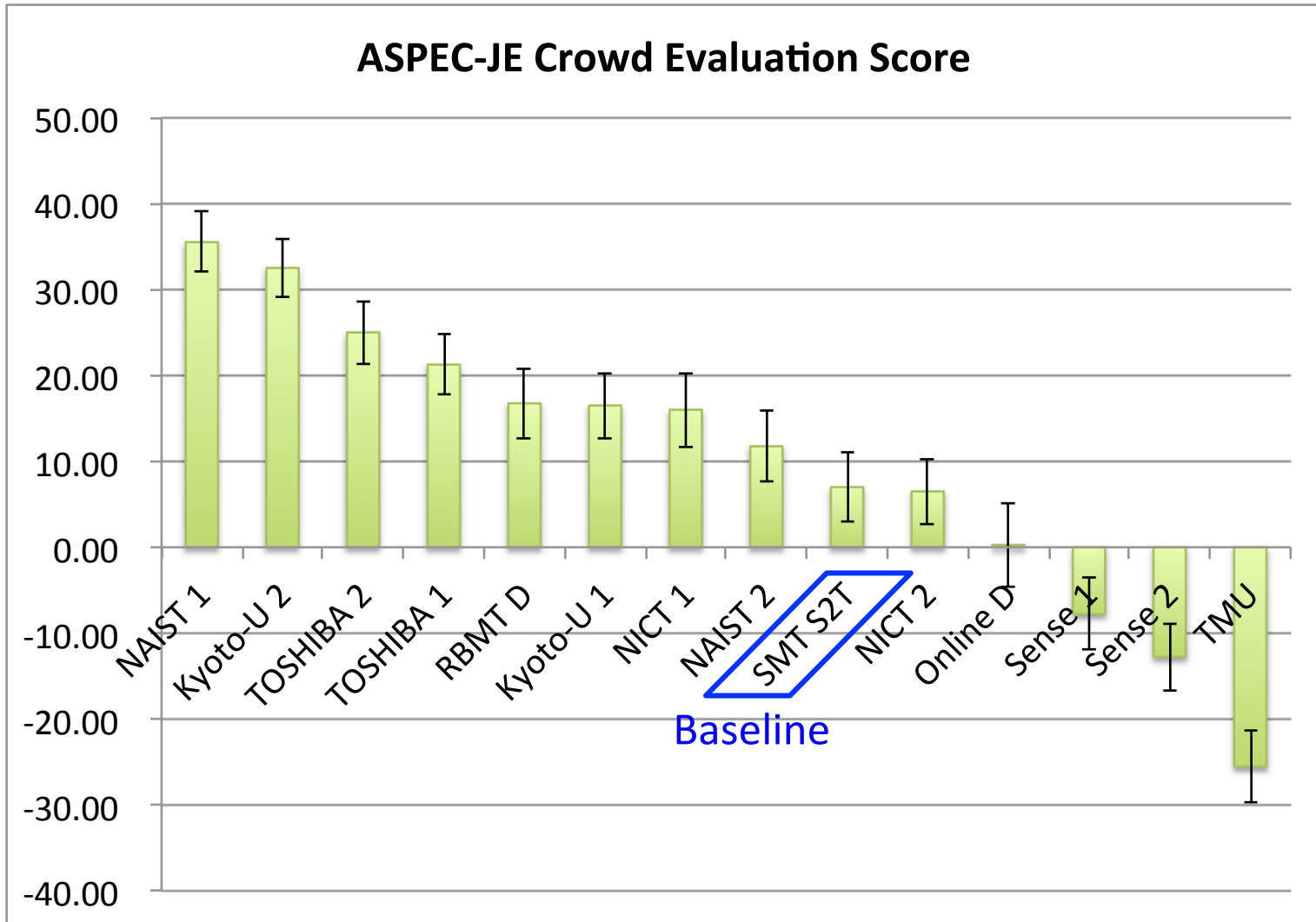
JPO Adequacy

- Sentence-level 5-scale criterion defined by Japan Patent Office

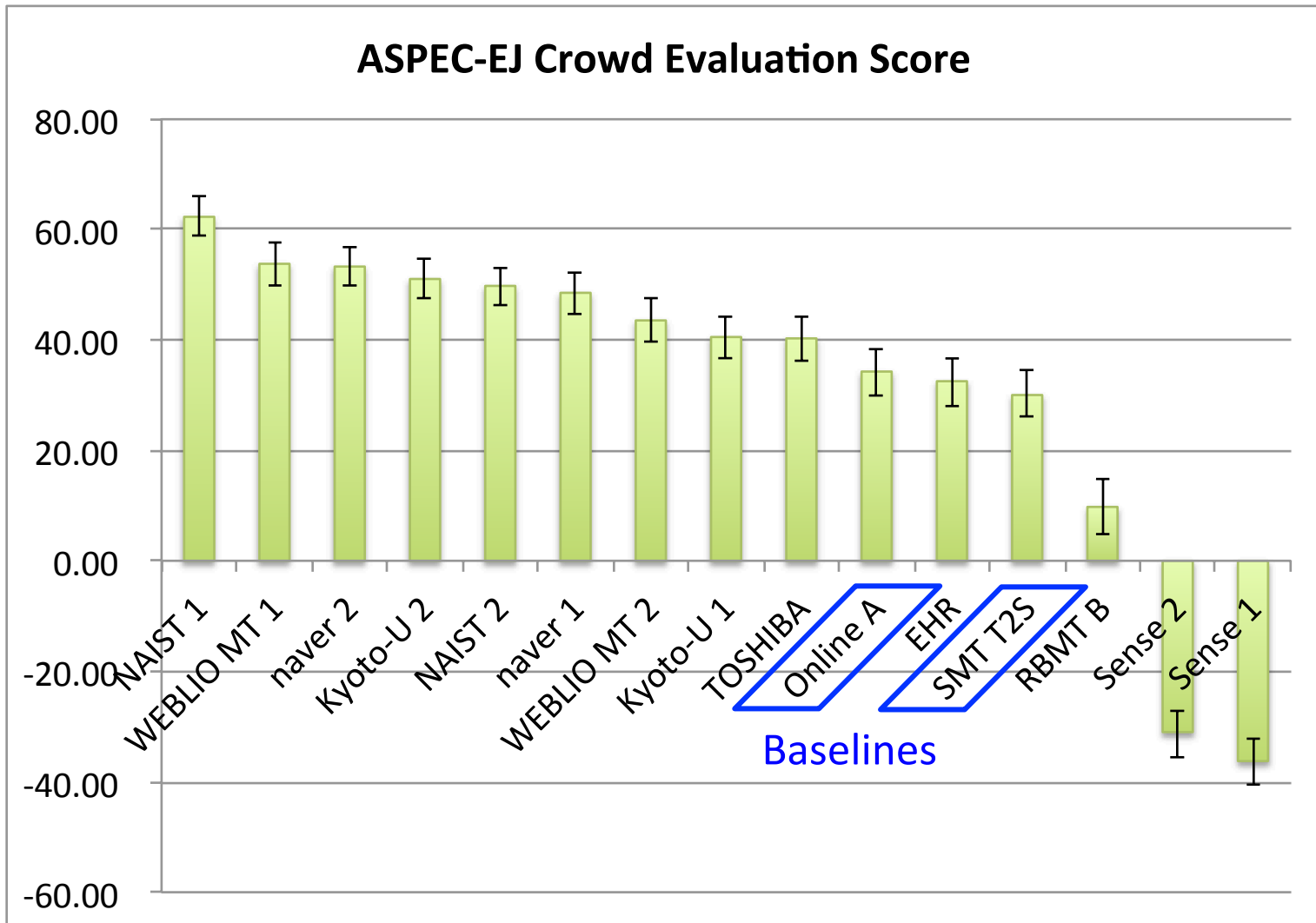
5	All important information is transmitted correctly. (100%)
4	Almost all important information is transmitted correctly. (80%~)
3	More than half of important information is transmitted correctly. (50%~)
2	Some of important information is transmitted correctly. (20%~)
1	Almost no important information is transmitted correctly. (~20%)

- **Professional translators** assessed.
- Top 3 systems of the Crowd scores were evaluated.
- One sentence was evaluated by **two evaluators**.
- **200 sentences** were evaluated for each system.

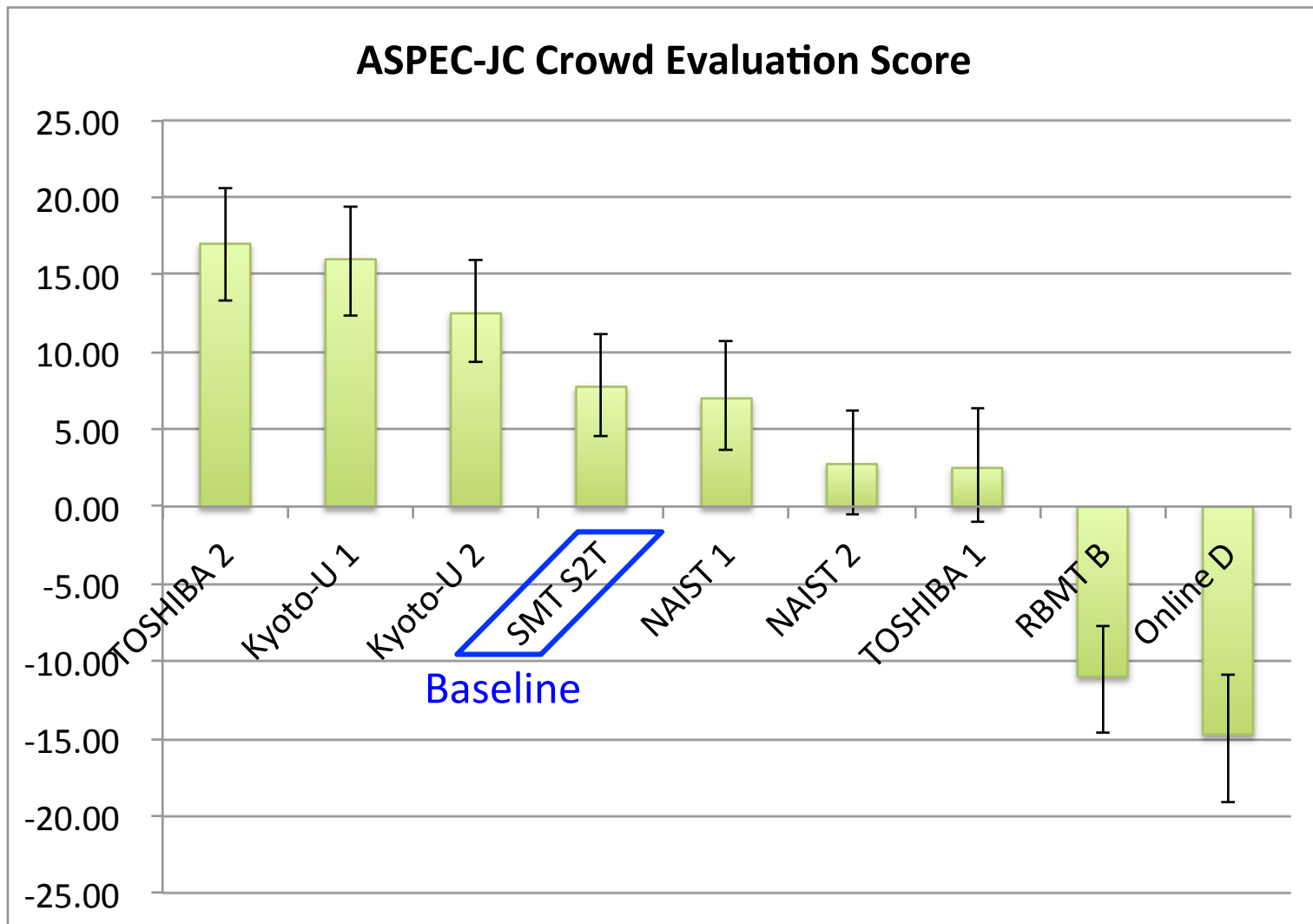
Crowd Evaluation Results



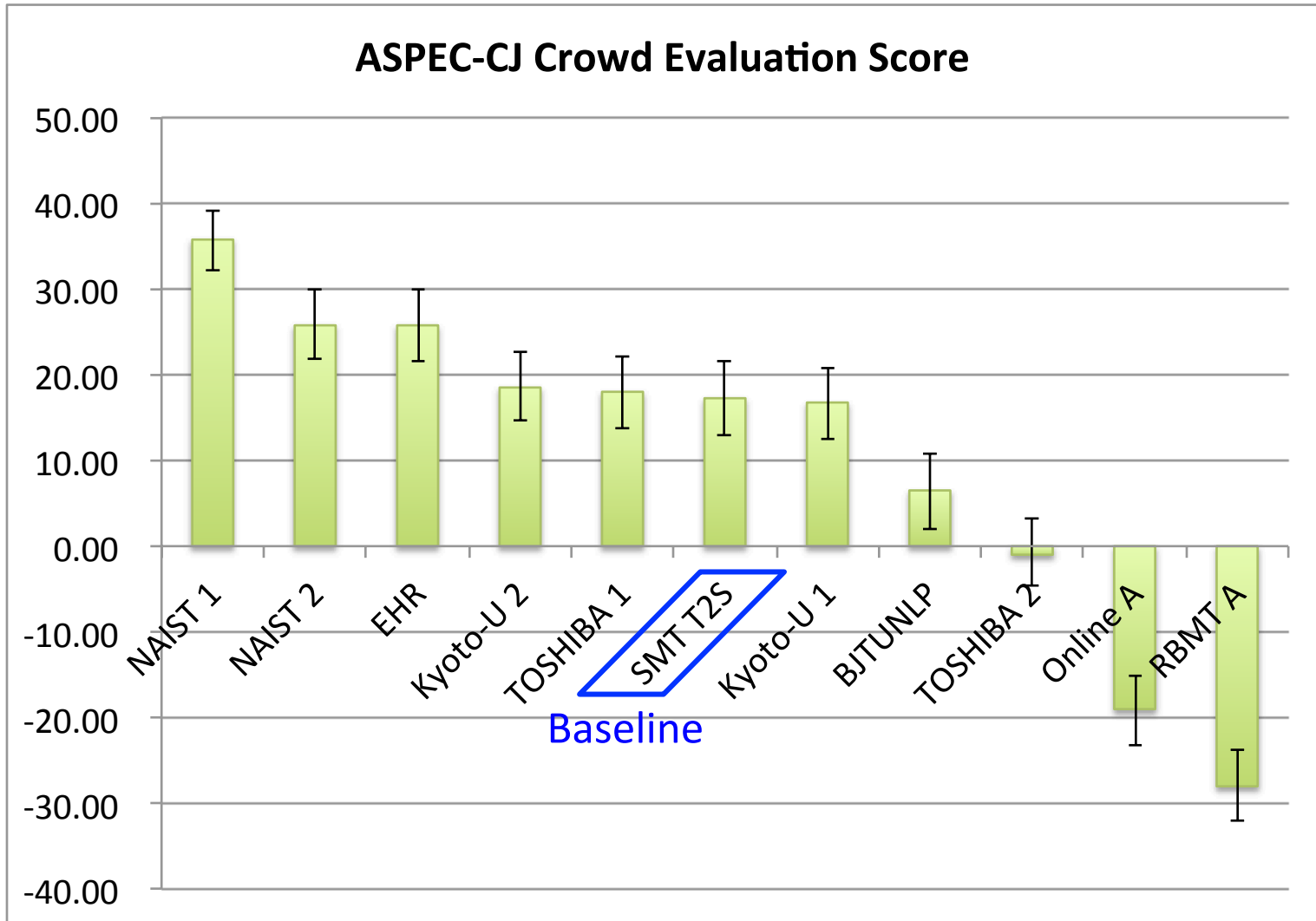
Many participants outperformed SMT S2T, which ranked at the second at WAT2014.



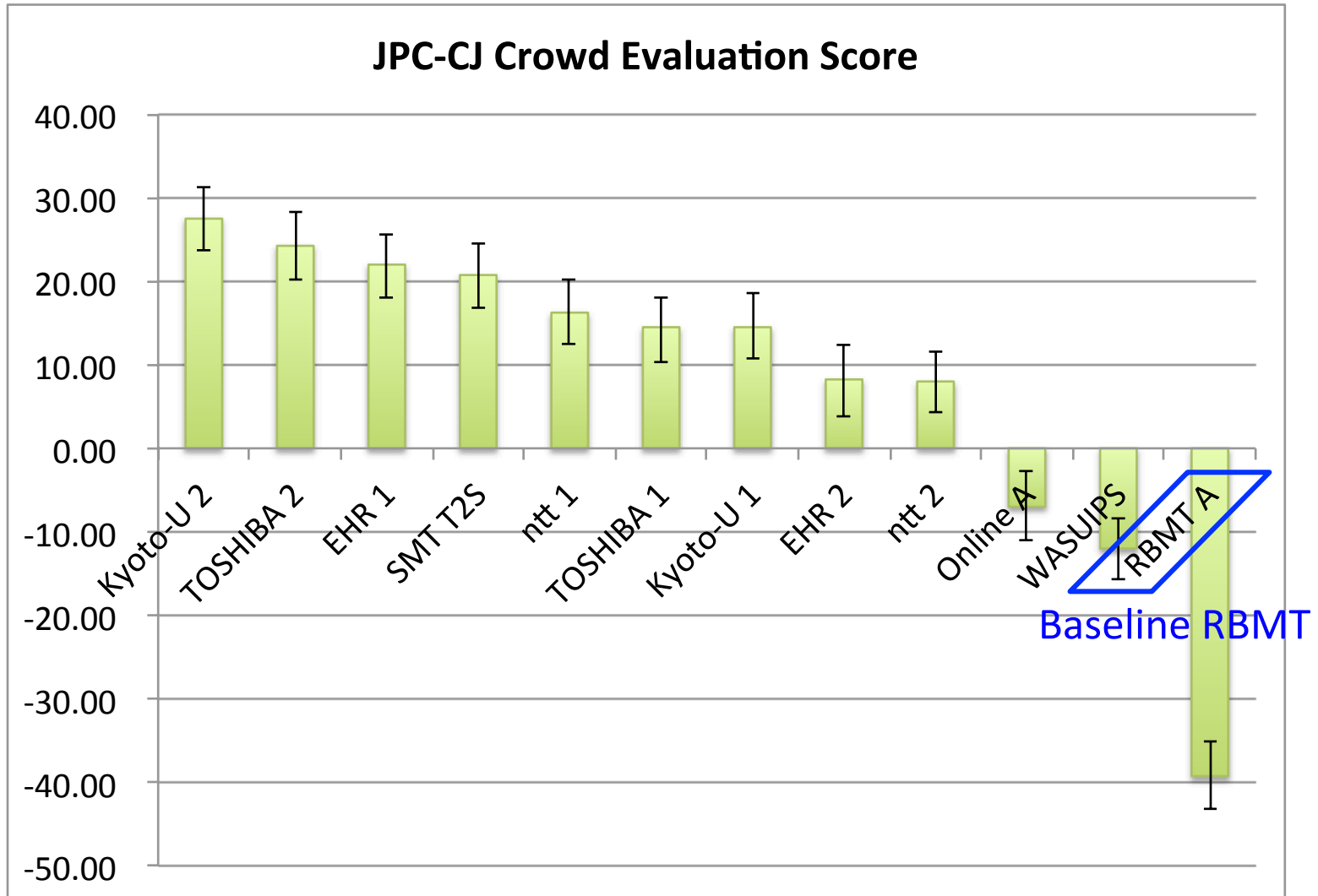
Many participants outperformed the baselines of Online A and SMT T2S



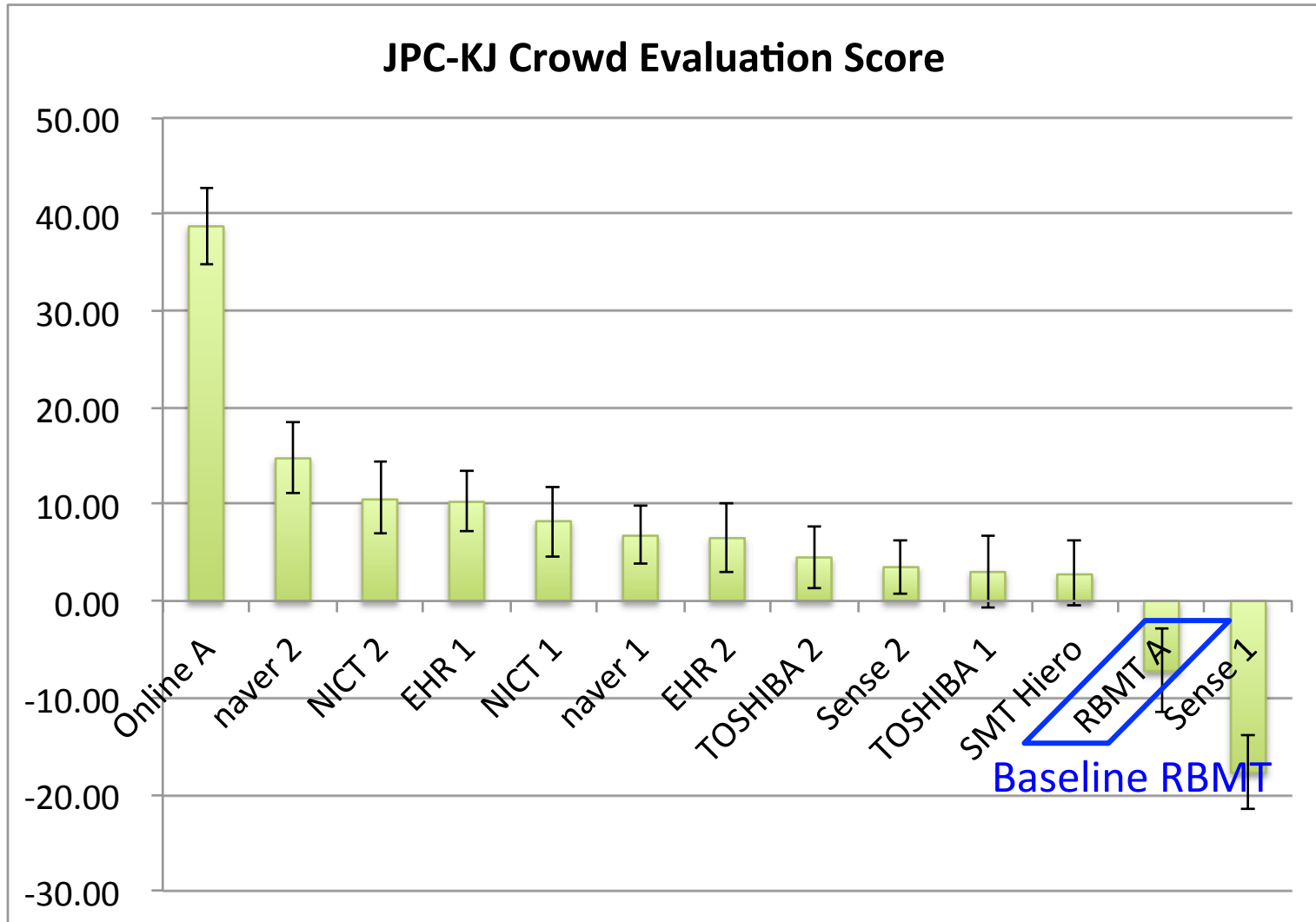
Two teams outperformed SMT S2T, which ranked at the second at WAT2014. NAIST 1 ranked at middle. However this ranking was different from JPO adequacy.



NAIST team achieved the best score.



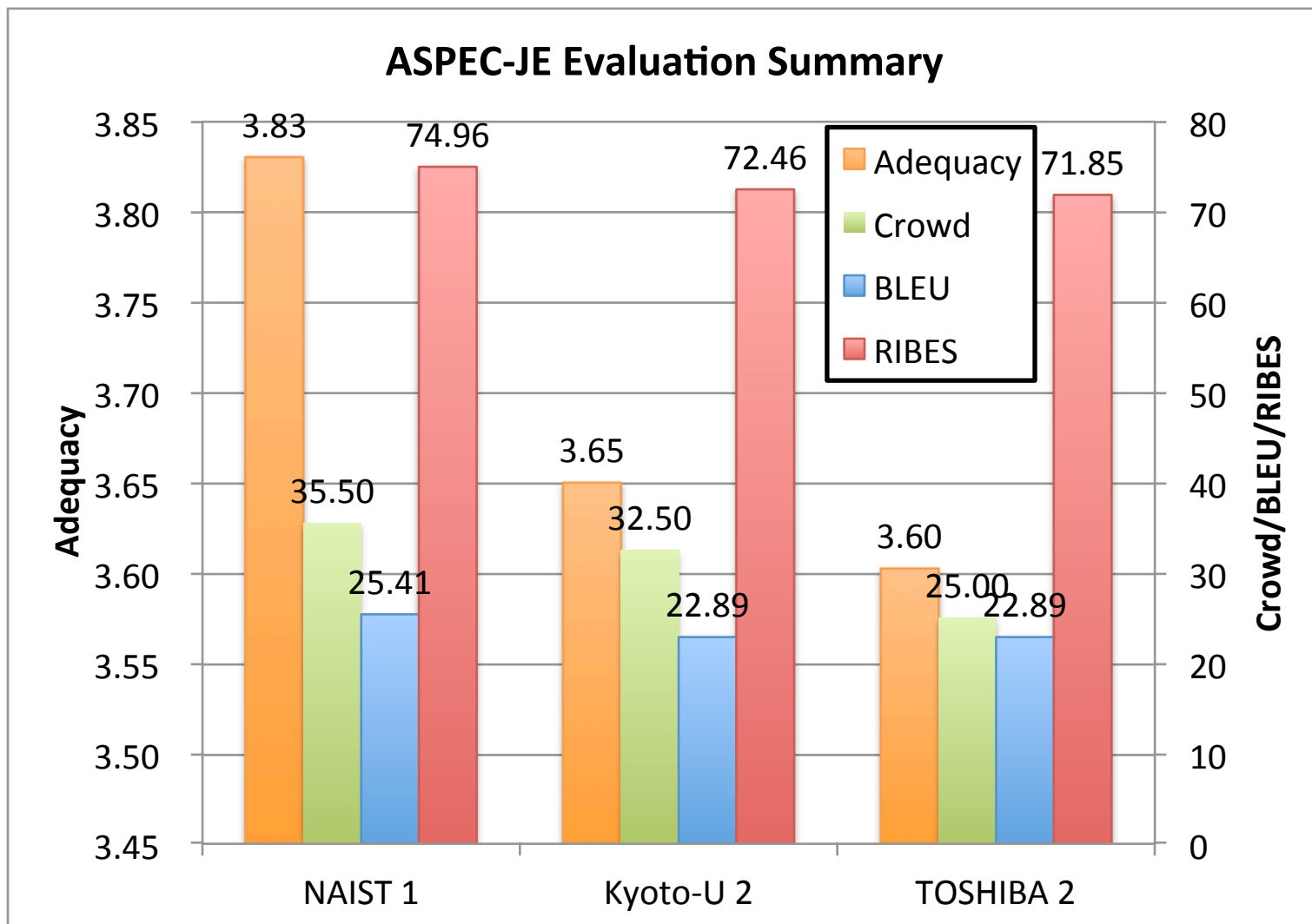
SMT systems outperformed the baseline RBMT for CJ patent translation.



SMT systems outperformed the baseline RBMT for KJ patent translation. Online A was top-ranked. However this rank was different from JPO adequacy.

JPO Adequacy Evaluation Results and Reliability of Crowd/BLEU/RIBES

Comparison with JPO Adequacy (ASPEC-JE)

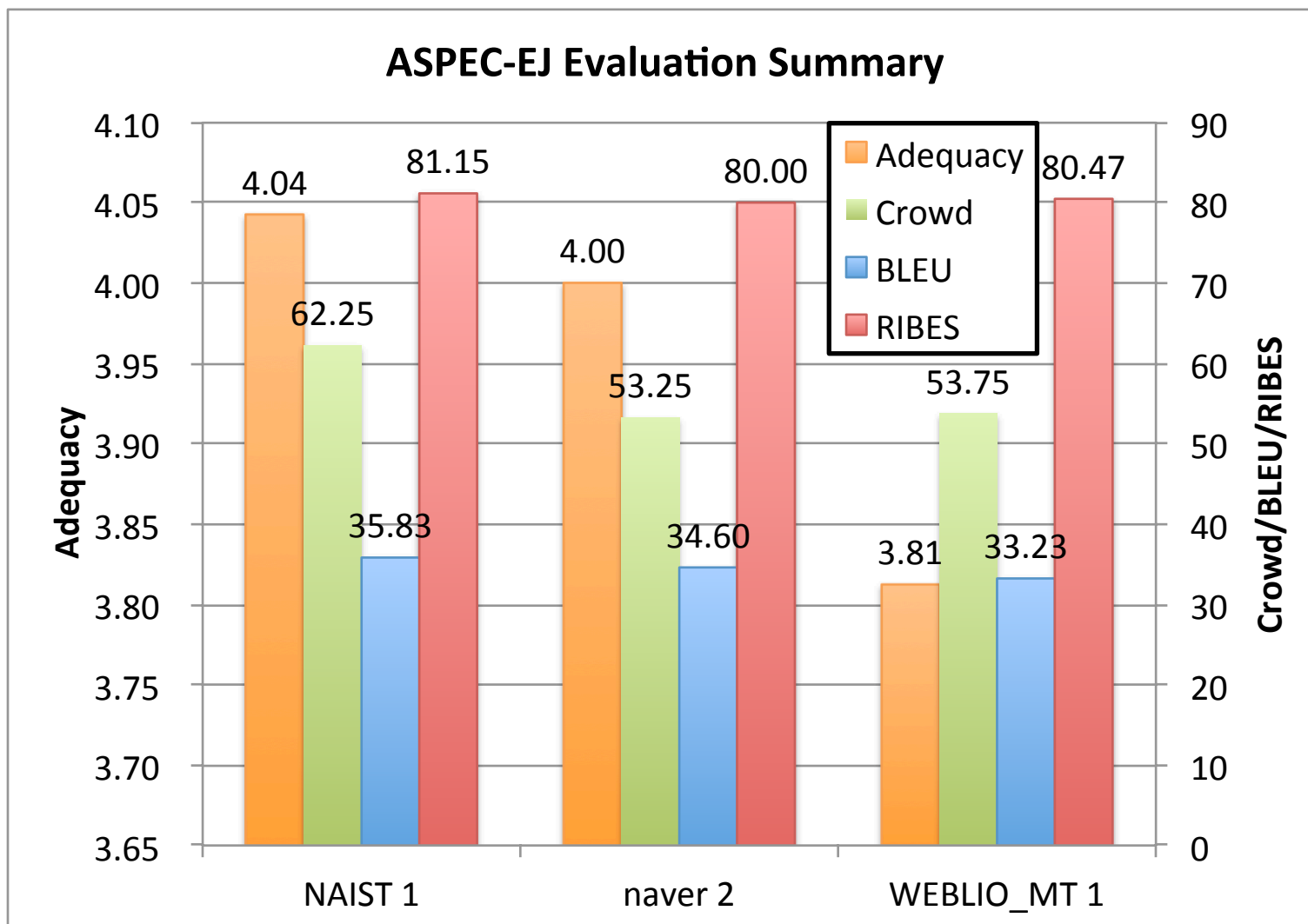


The ranking of **Crowd** is **consistent**.

The ranking of BLEU is partially consistent.

The ranking of **RIBES** is **consistent**.

Comparison with JPO Adequacy (ASPEC-EJ)

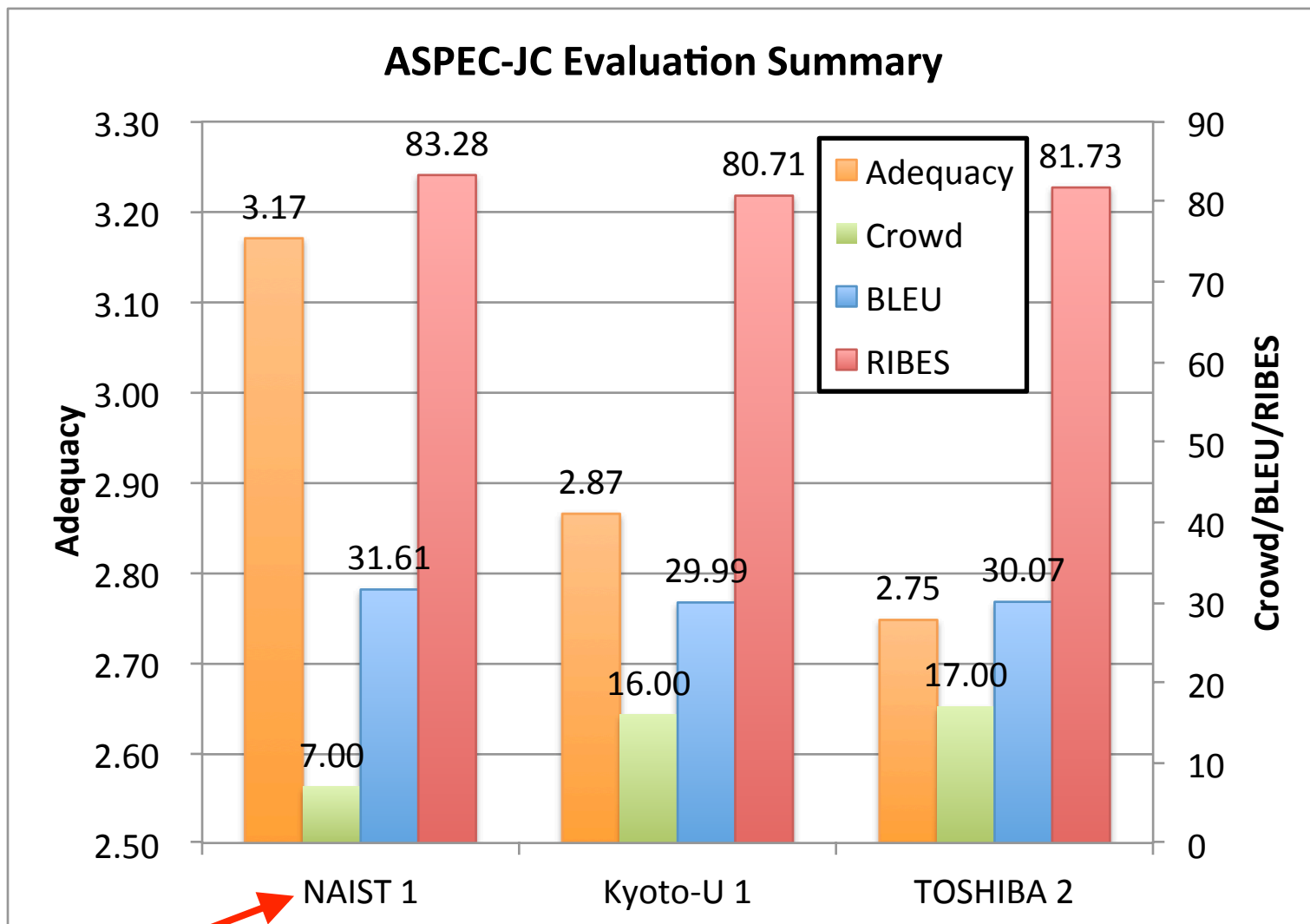


The ranking of Crowd is partially consistent.

The ranking of **BLEU** is **consistent**.

The ranking of RIBES is partially consistent.

Comparison with JPO Adequacy (ASPEC-JC)

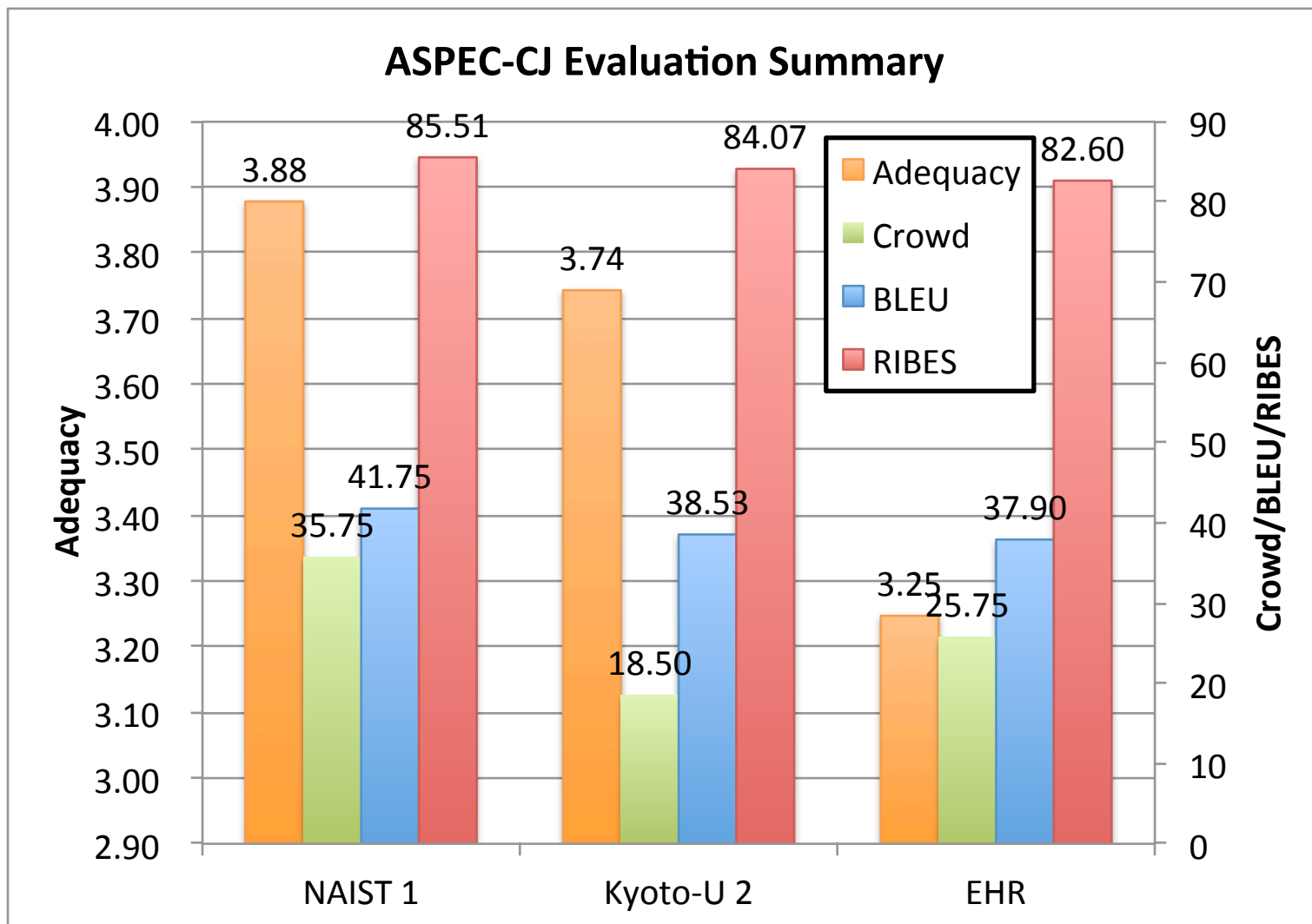


Different ranking from Crowd

The ranking of Crowd is not consistent.
The ranking of BLEU is partially consistent.
The ranking of RIBES is partially consistent.

JC is difficult to evaluate!

Comparison with JPO Adequacy (ASPEC-CJ)

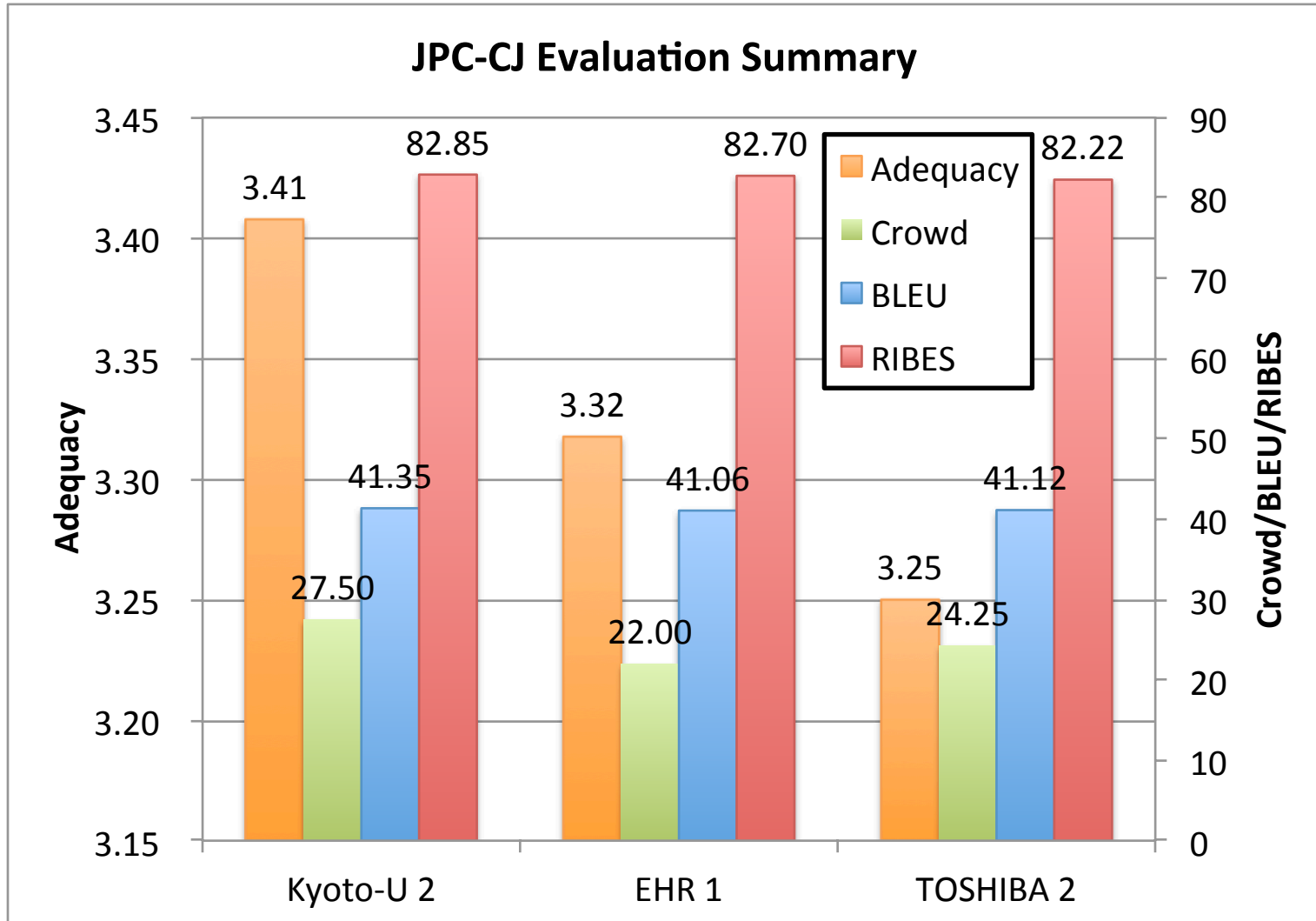


The ranking of Crowd scores is partially consistent.

The ranking of **BLEU** scores is **consistent**.

The ranking of **RIBES** scores is **consistent**.

Comparison with JPO Adequacy (JPC-CJ)

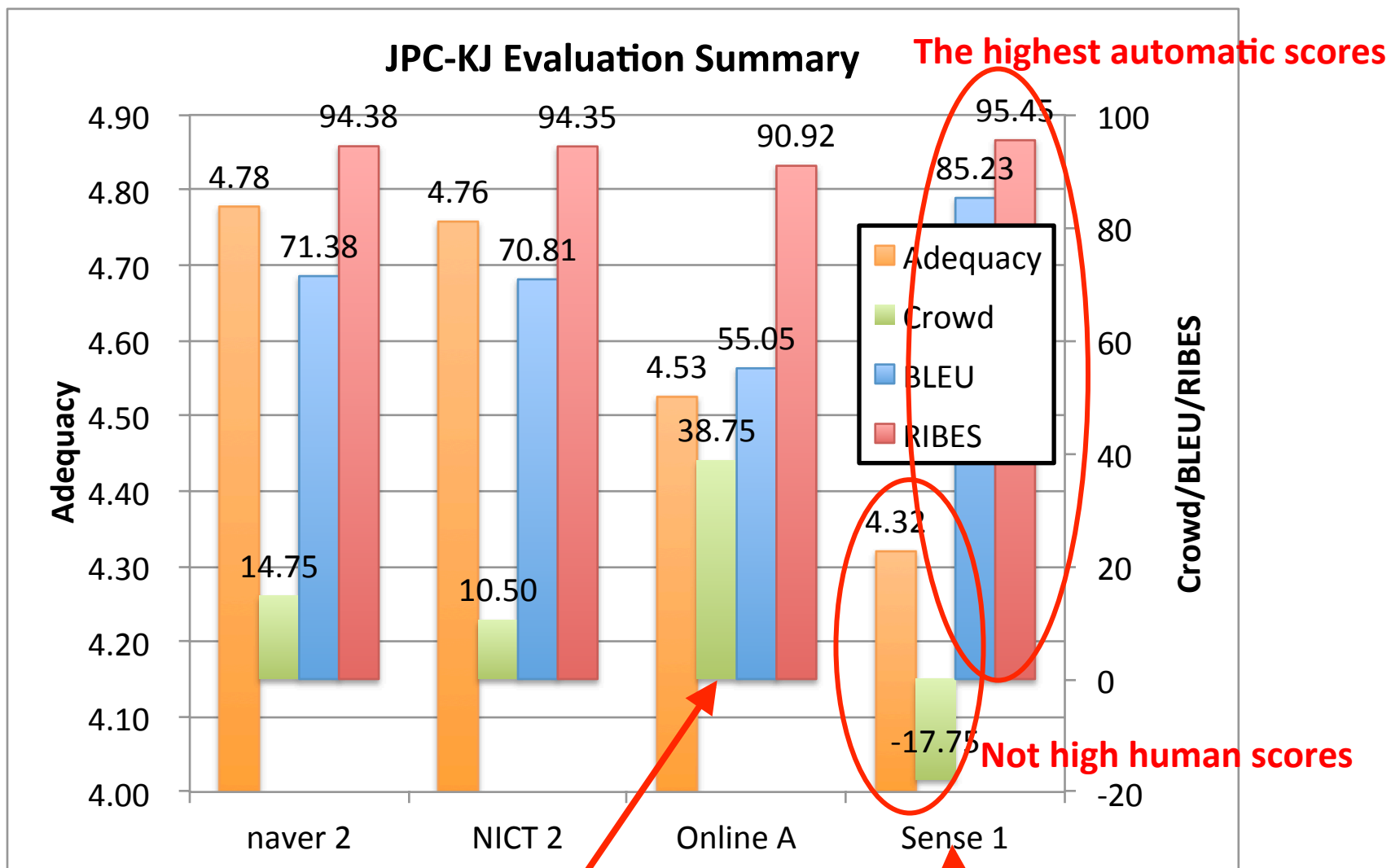


The ranking of Crowd was partially consistent.

The ranking of BLEU was partially consistent.

The ranking of **RIBES** was **consistent**.

Comparison with JPO Adequacy (JPC-KJ)



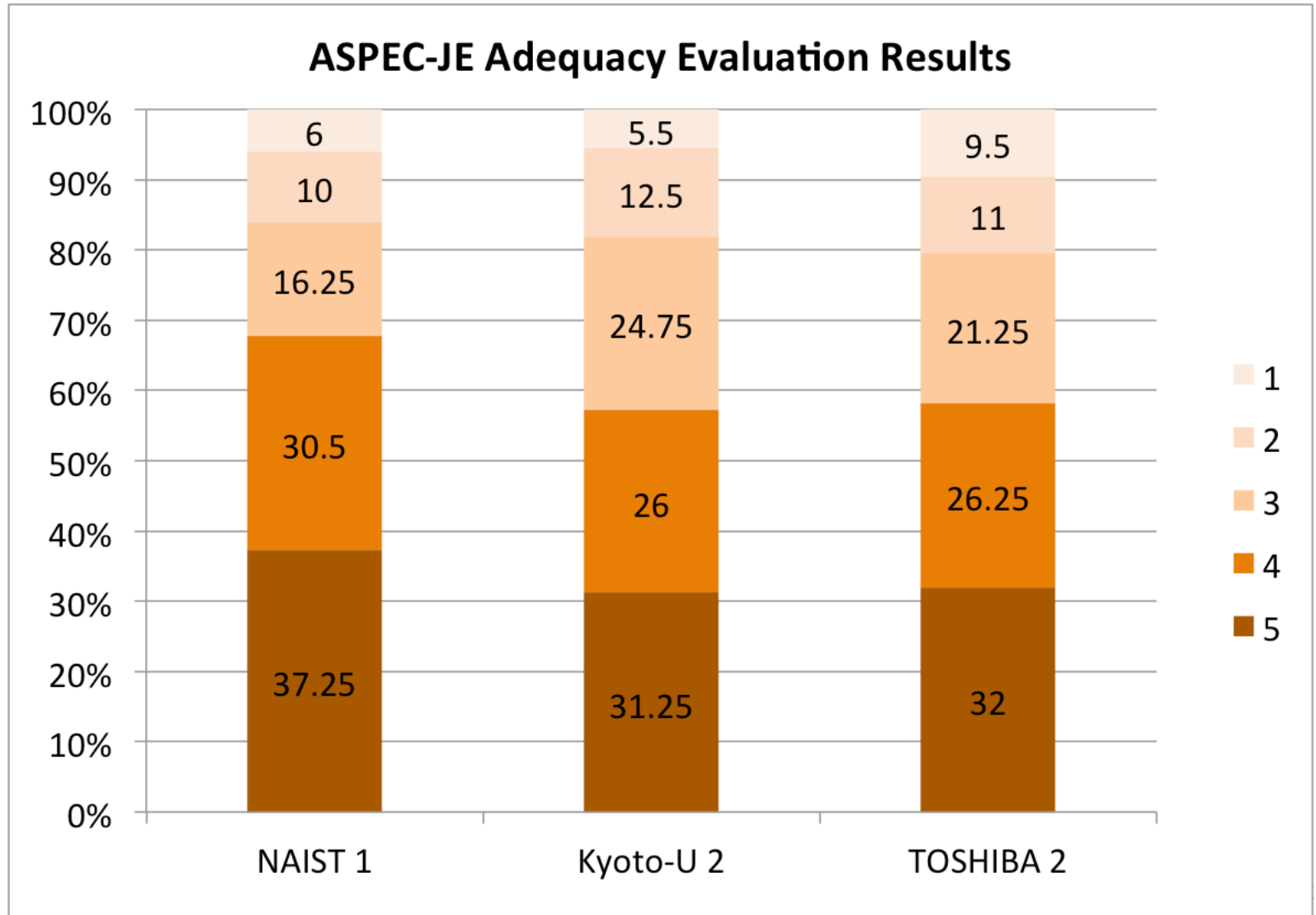
The Crowd score of Online A is very high.
JPO adequacy of Online A was not top-ranked.

The best BLEU and RIBES scores.
However, human evaluation was not high.

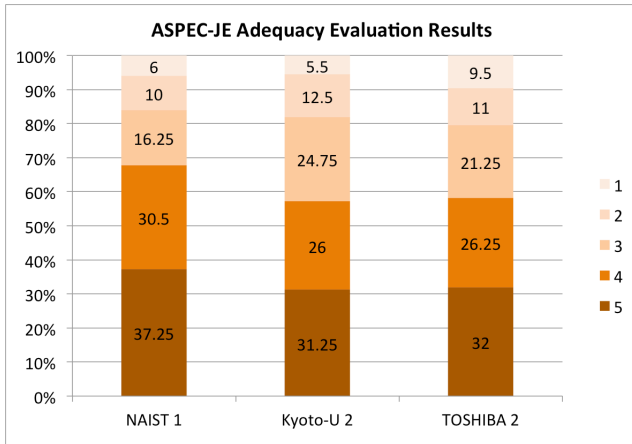
There might be a remaining problem in automatic evaluation.

Comparison of Translation Quality between Languages

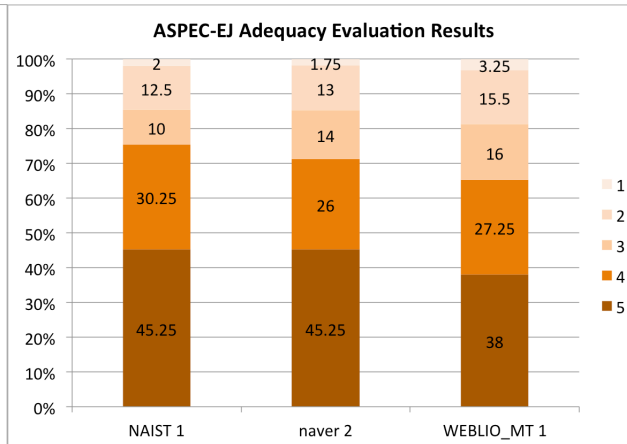
Graph Format of JPO Adequacy



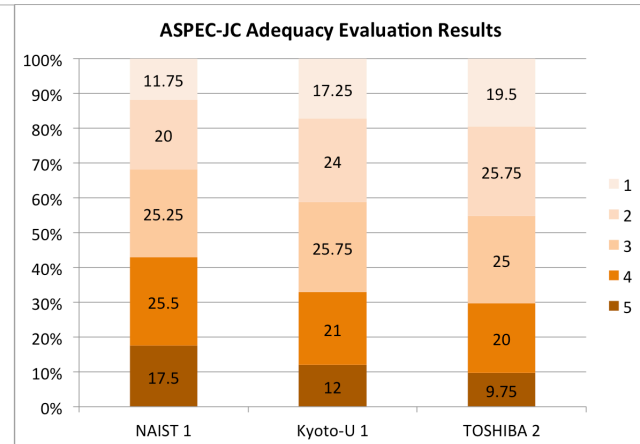
ASPEC-JE



ASPEC-EJ



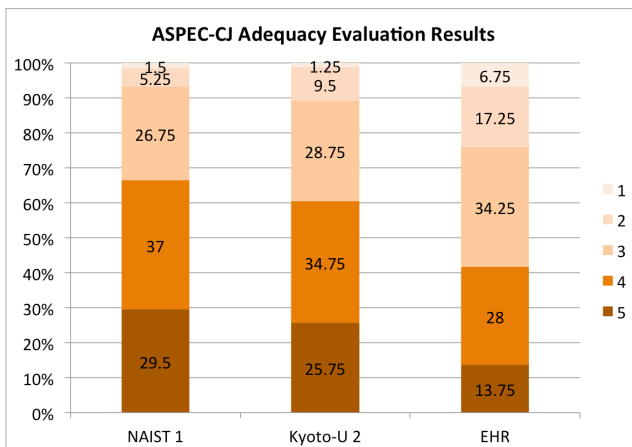
ASPEC-JC



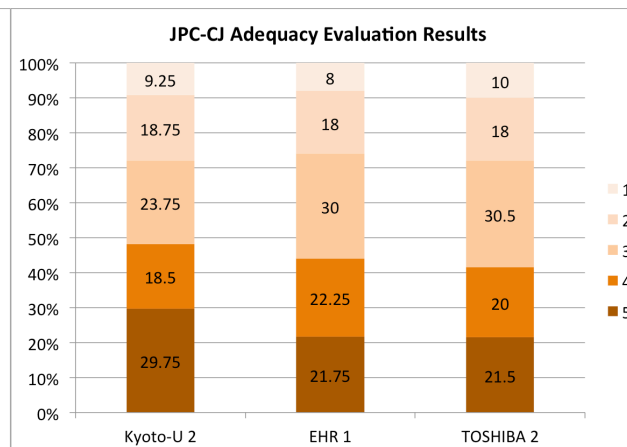
The second high quality

Low quality
(Difficult language pair)

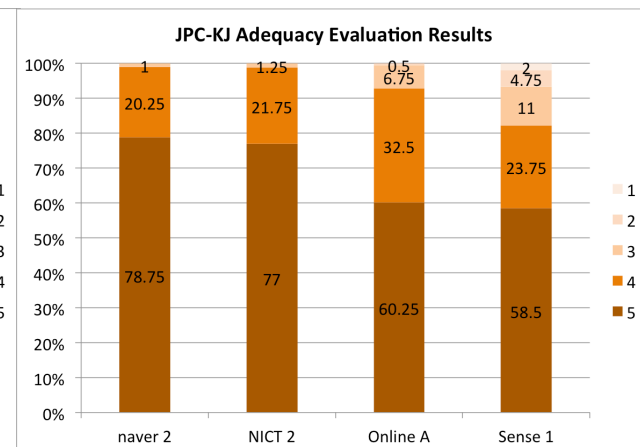
ASPEC-CJ



JPC-CJ



JPC-KJ



High quality was achieved!

Summary of WAT2015

- **12** participants for the evaluation tasks
 - Including **4 companies** and **3 teams outside Japan**
- 2 domains (scientific paper and patent) and 4 languages (Japanese, Chinese, Korean, and English)
- Human evaluations were performed
 - Pairwise Evaluation by crowdsourcing
 - JPO Adequacy by professional translators
- Empirically confirmed that MT systems achieved high quality Korean-Japanese patent translation.
- Each idea used for the submissions will be presented by the participants.

Future Perspective

- Papers submitted to WAT will appear on the ACL Anthology.
- 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...
- Need more investigation to acquire reliable human evaluation results at low cost

Thank you very much
for attending WAT2015