Exploiting Patent Information for the Evaluation of Machine Translation

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto,
Takehito Utsuro

Background

- Corpus-based machine translation is widely studied
- Competition-style MT workshops exist

However

- No large scale Japanese-English parallel corpus

NTCIR-7 Workshop

Workshop for evaluation of multi-lingual information access technologies

Patent translation task

- Intrinsic evaluation — BLEU, Human (Adequacy, Fluency)
- Extrinsic evaluation — Cross-lingual information retrieval

NTCIR-8 patent translation task will start from Sep 1 2009.

Intrinsic evaluation

- Training and test data
- Evaluation methods
- Multi-references
- Results

Training and test data

- Unexamined Japanese patent applications during 1993–2002. (3 500 000 documents)
- US patent grant data during 1993–2002 (1 300 000 documents)
- 85 000 patent families were extracted
- 1 800 000 Japanese–English sentence pairs for training
- 1381 sentence pairs for test (Avg. 29 English words)

Training and test data are publicly available now

Evaluation method

BLEU

Reference translations = Counterpart sentences and translations by human experts

Human judgment

- Adequacy and fluency (five-point rating)
- Randomly selected 100 test sentences

Producing multiple references (Problematic)

- **S600**: Three experts independently translated 600 sentences into English. However, they used a rule-based MT system.
- **S300**: Different three experts translated 300 sentences. One expert still used an RBMT system

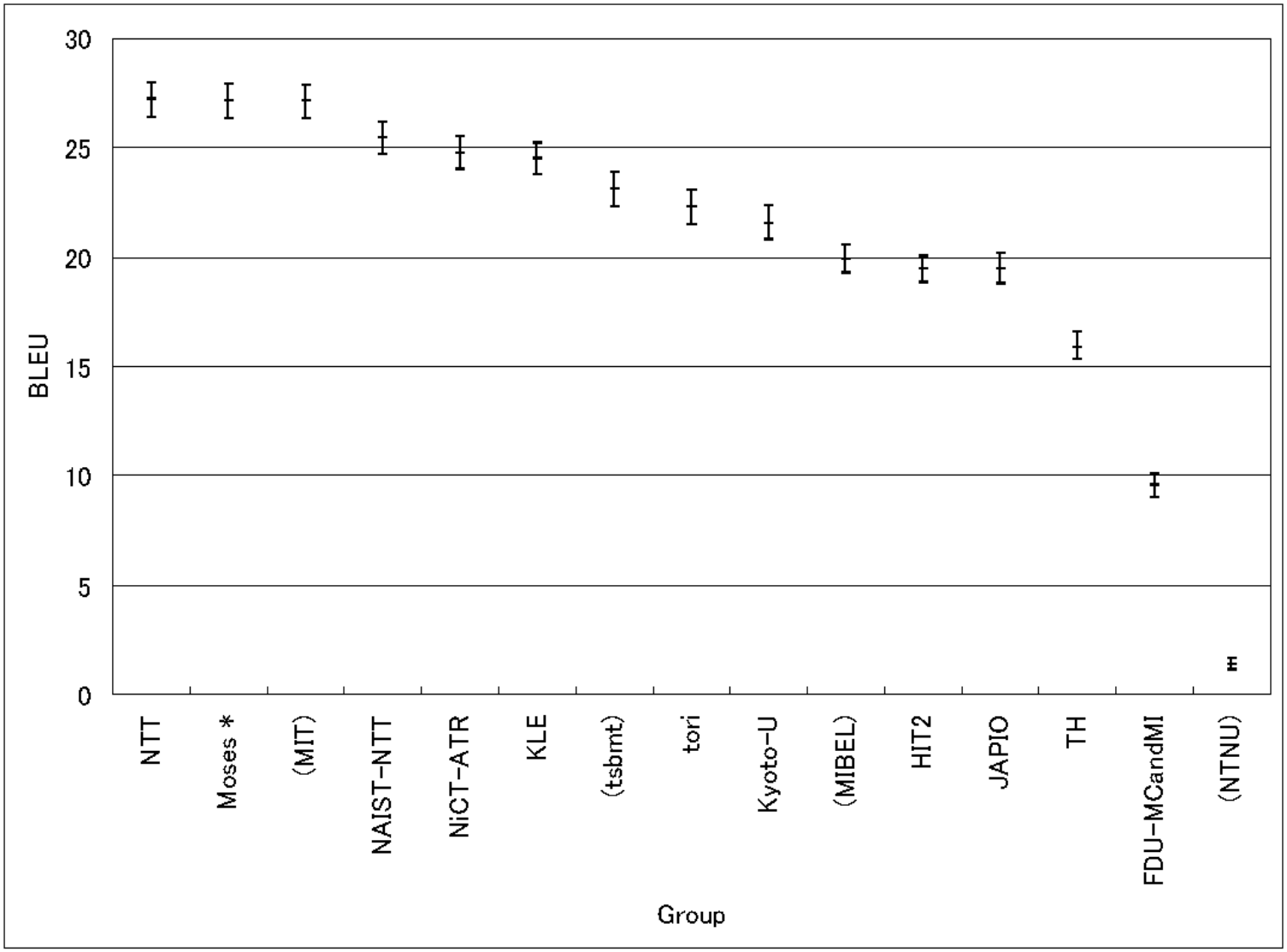
The counterpart English sentences for Japanese test sentences are also potentially influenced by RBMT systems, because it is often the case that a human expert edits a machine translated text.

Three types of BLEU

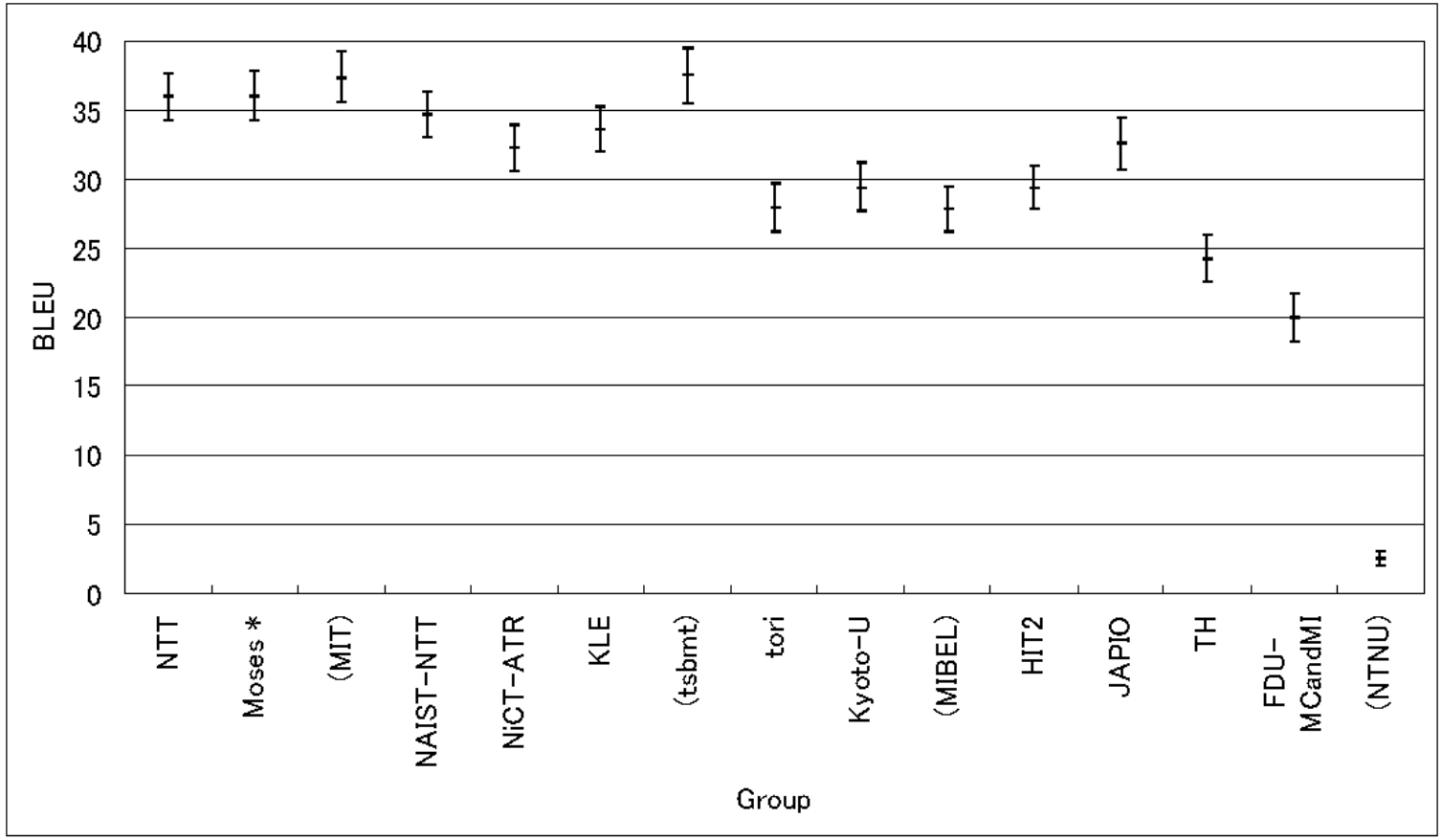
- **Single-Reference BLEU (SRB)** is calculated by the counterpart sentences for the 1381 test sentences. We can use all test sentences available. (1 ref)
- **Multi-Reference BLEU for S300 (MRB300)** is calculated by the reference translations produced by two experts w/o RBMT. We can use as many reference translations as possible, while avoiding the influence of RBMT systems. (2 ref)
- **Multi-Reference BLEU for S600 (MRB600)** is calculated by the S600 reference translations and the counterpart sentences. we can use as many reference translations and test sentences as possible. This value is potentially influenced by RBMT systems. (4 ref)

Group	Method	JE BLEU			Human
		SRB	MRB300	MRB600	
NTT	SMT	27.20	35.93	43.72	3.30
Moses *	SMT	27.14	36.02	43.40	3.18
(MIT)	SMT	27.14	37.31	44.69	3.40
NAIST-NTT	SMT	25.48	34.66	41.89	3.04
NiCT-ATR	SMT	24.79	32.29	39.40	2.78
KLE	SMT	24.49	33.59	40.20	2.94
(tsbmt)	RBMT	23.10	37.51	48.02	3.88
tori	SMT	22.29	27.92	35.02	3.01
Kyoto-U	EBMT	21.57	29.35	35.49	3.10
(MIBEL)	SMT	19.93	27.84	32.99	2.74
HIT2	SMT	19.48	29.33	33.60	2.86
JAPIO	RBMT	19.46	32.62	41.77	3.86
TH	SMT	15.90	24.20	28.72	2.13
FDU-MCandWI	SMT	9.55	19.94	20.27	2.08
(NTNU)	SMT	1.41	2.48	2.63	1.06

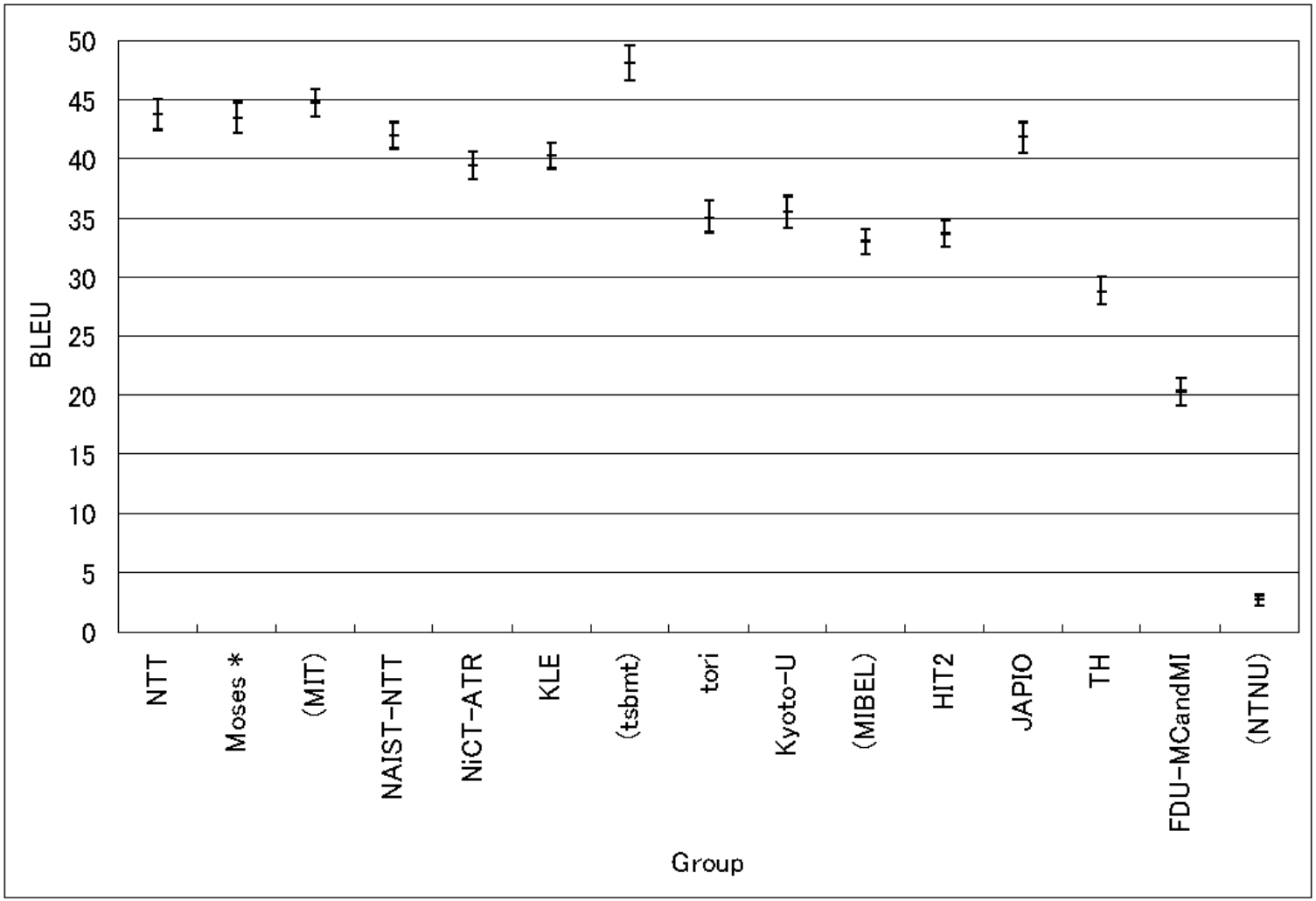
SRB with a 95% confidence interval (1 ref)



MRB300 with a 95% confidence interval (2 ref)



MRB600 with a 95% confidence interval (4 ref)



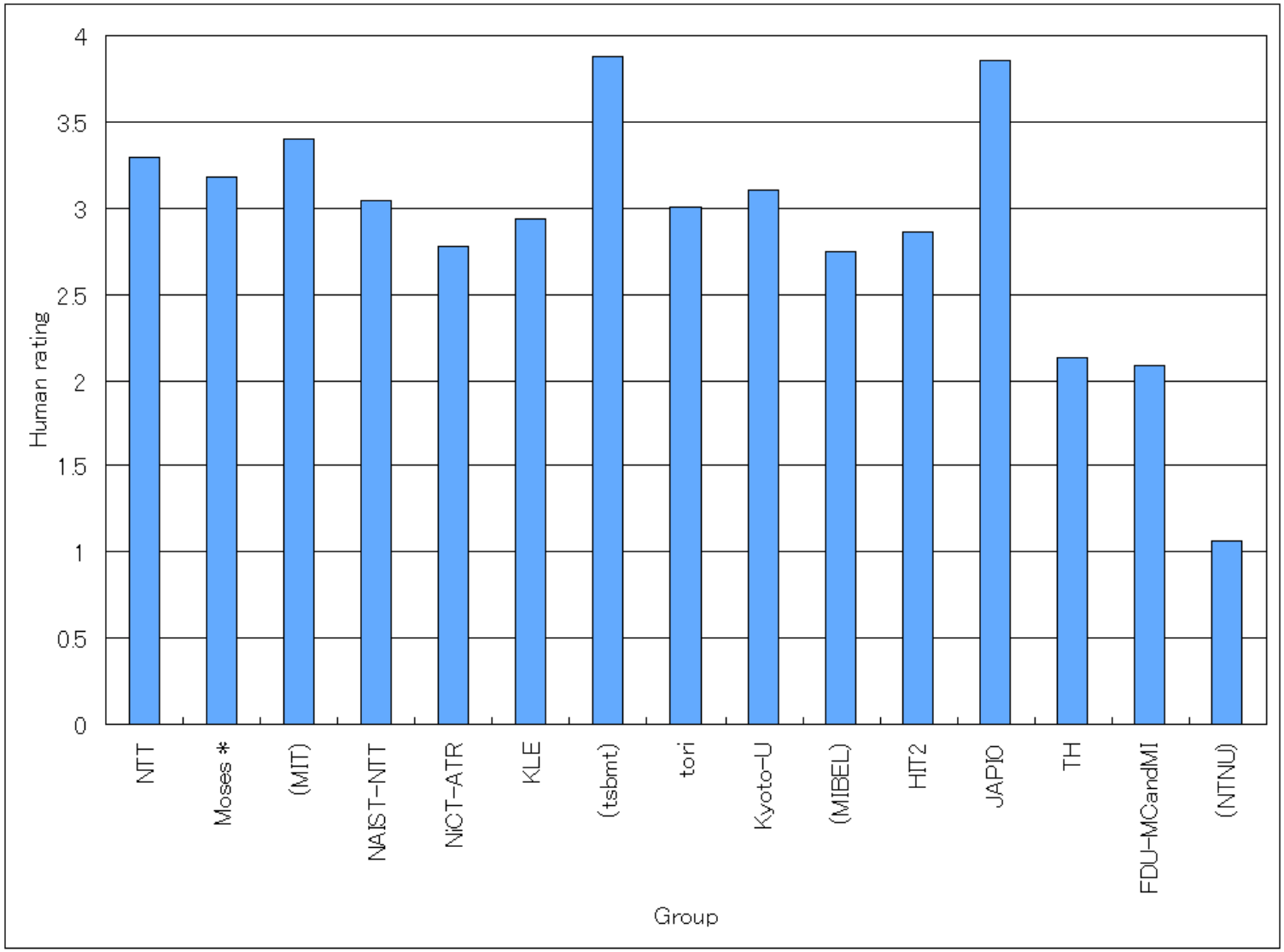
Observations for JE

- SMT obtained large values for SRB
- Increases of "tsbmt" and "JAPIO" in MRB300 and MRB600 are noticeable

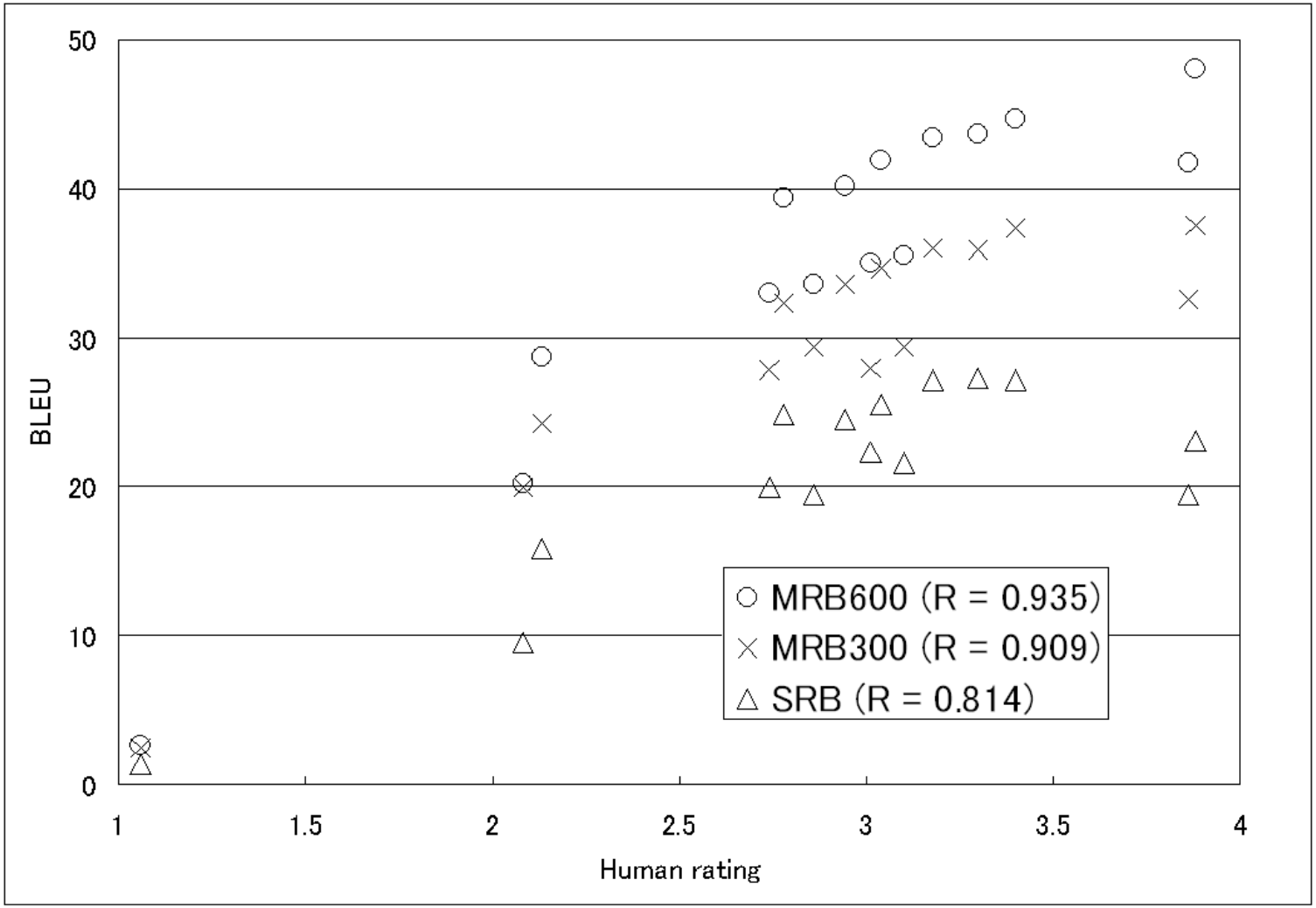
Possible reasons

- MRB600 were potentially influenced by RBMT systems.
- RBMT matched well for counterpart and human translated sentences
- SMT didn't match well for human translated sentences

Human rating for J–E intrinsic evaluation.



Relationship between BLEU and human rating



Familiar results confirmed for JE

- RBMT (tsbmt, JAPIO) outperformed other systems in human ratings
- Evaluation by BLEU became similar to that by human rating with multiple references

Extrinsic Evaluation

- Machine translate search topics from English into Japanese
- Translated search topics was used to search patent documents in Japanese
- Invalidity search was performed
- 124 claims were translated (Avg. 115.4 words, very long!)
- Mean average precision (MAP) was used

Results of E–J int/ext evaluation

Group	Method	Intrinsic		Extrinsic	
		BLEU	Human	BLEU	MAP
Moses *	SMT	30.58	3.30	20.70	.3140
HCRL	SMT	29.97	—	21.10	.3536
NiCT-ATR	SMT	29.15	2.89	19.40	.3494
NTT	SMT	28.07	3.14	18.69	.3456
NAIST-NTT	SMT	27.19	—	20.46	.3248
KLE	SMT	26.93	—	19.07	.2925
tori	SMT	25.33	—	17.54	.3187
(MIBEL)	SMT	23.72	—	18.67	.2873
HIT2	SMT	22.84	—	17.71	.2777
(Kyoto-U)	EBMT	22.65	2.48	13.75	.2817
(tsbmt)	RBMT	17.46	3.60	12.39	.2264
FDU-MCandWI	SMT	10.52	—	11.10	.2562
TH	SMT	2.23	—	1.39	.1000
Mono	—	—	—	—	.4797

Observations

- Best MAP by HCRL was 74% of that by Mono
- Correlation between BLEU and MAP was 0.967
- Correlation between Human ratings and MAP was not large

Good translation =? Good CLIR

Conclusion

- NTCIR-7 Patent translation task
- Familiar results observed for intrinsic evaluation
- Correlation between BLEU and MAP was very high
- Correlation between Human ratings and MAP was not large

Advertisement

- Training and test data are publicly available now
- NTCIR-8 patent translation task will start from Sep 1 2009