

Overview of WAT2014

Toshiaki Nakazawa
nakazawa@pa.jst.jp

Hideya Mino
hideya.mino@nict.go.jp

Isao Goto
goto.i-es@nhk.or.jp

Sadao Kurohashi
kuro@i.kyoto-u.ac.jp

Eiichiro Sumita
eiichiro.sumita@nict.go.jp

Participants List

Team ID	Organization	J->E	E->J	J->C	C->J
NAIST	Nara Institute of Science and Technology	✓	✓	✓	✓
EIWA	Yamanashi Eiwa College	✓			✓
Kyoto-U	Kyoto University	✓	✓	✓	✓
WEBLIO-EJ1	Weblio, Inc.			✓	
TMU	Tokyo Metropolitan University	✓			
BJTUNLP	Beijing Jiaotong University			✓	
NII	National Institute of Informatics	✓			
SAS_MT	SAS Research and Development Co., Ltd		✓	✓	✓
Sense	Saarland University & Nanyang Technological University	✓	✓	✓	✓
NICT	National Institute of Information and Communication Technology			✓	
TOSHIBA	Toshiba Corporation		✓	✓	
WASUIPS	Waseda University			✓*	✓*
company					
outside Japan					

* Only submitted to the automatic evaluations

HUMAN score

<http://www.lancers.jp>

* Pairwise evaluation compared to the baseline using crowdsourcing
 - reduce the cost, time, number of sentences to be evaluated
 - enable the evaluation of new translation results after the workshop
 * Reference translations are not shown

2つの機械翻訳結果の優劣判断

機械翻訳結果の差異をもとに各スコアをもとに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いいかを判断してください。
 判定方法：評議会で投票する場合、投票をしてください。

入力文：Fugen Power Plant can be calculated by using DERS software.
 訳文1：ふげん発電所の蓄電池の性能を用いて計算できる。DERSソフトウェアである。
 訳文2：「ふげん」発電所の蓄電池の性能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つの翻訳の方が良い ◎2つの翻訳の方が良い ◎同程度

292

* 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

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	B	Tie	B	Tie	B	B	B
Decision	A	A	A	A	Tie	B	Tie	B	B	B

* 400 sentences are randomly selected
 (paragraph-based) from the test set

BASELINE (Phrase-based SMT)	IN-MT-1	SYSTEM1-MT-1
	BASELINE-MT-2	SYSTEM1-MT-2
	BASELINE-MT-3	SYSTEM1-MT-3
	BASELINE-MT-4	SYSTEM1-MT-4
	BASELINE-E-MT-5	SYSTEM1-MT-5
	BASELINE-MT-6	SYSTEM1-MT-6
	BASELINE-E-MT-7	SYSTEM1-MT-7
	BASELINE-IN-MT-8	SYSTEM1-MT-8
	BASELINE-MT-9	SYSTEM1-MT-9
	BASELINE-E-MT-10	SYSTEM1-MT-10

VS.

$$\text{HUMAN} = 100 \times \frac{\text{Wins} - \text{Losses}}{\text{Wins} + \text{Losses} + \text{Ties}}$$

Cost

one judgment by a worker = 5 JPY

x 3 judgments for one decision

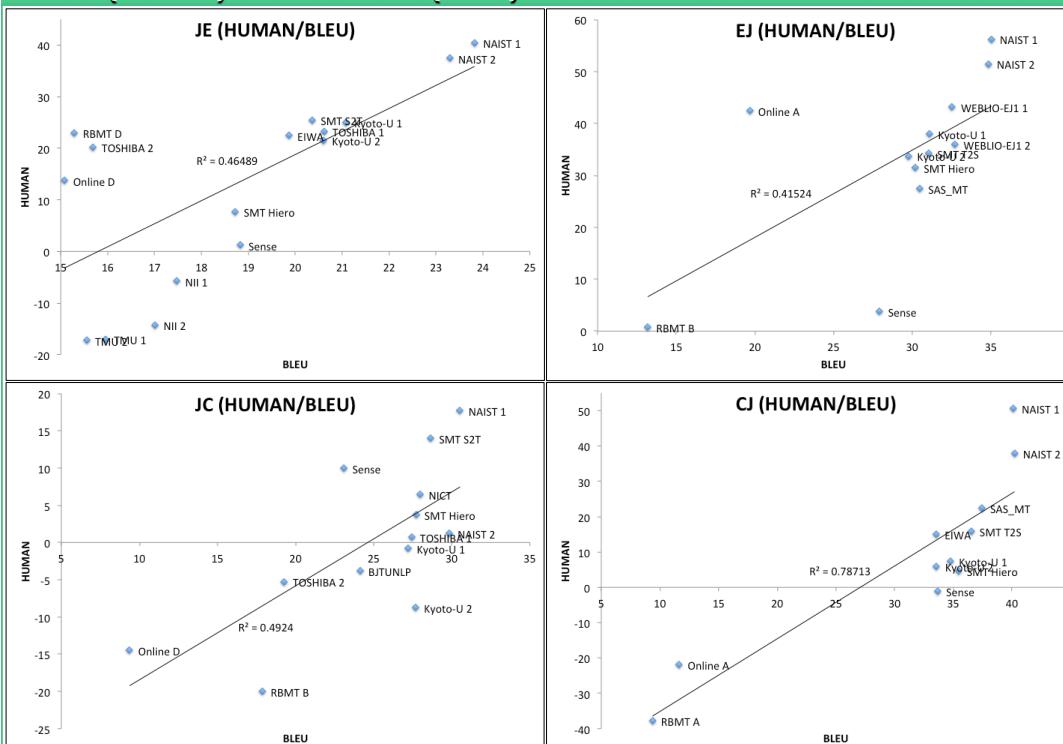
x 400 sentences

|| 6000 JPY / submission

Time

Depending on the translation direction
 On the average, one evaluation finished
 in a couple of days

Manual (HUMAN) and Automatic (BLEU) Evaluation Results and Their Correlations



Inter Annotator Agreement

JE	EJ	JC	CJ
System ID	Kappa	System ID	Kappa
NAIST 1	0.162	NAIST 1	0.077
NAIST 2	0.047	NAIST 2	0.280
SMT S2T	0.099	WEBLIO-EJ1 1	0.230
Kyoto-U 1	0.070	Online A	0.219
TOSHIBA 1	0.091	Kyoto-U 1	0.216
RBMT D	0.075	WEBLIO-EJ1 2	0.240
EIWA	0.083	SMT T2S	0.240
Kyoto-U 2	0.139	Kyoto-U 2	0.229
TOSHIBA 2	0.071	BITUNLP	0.198
Online D	0.055	SAS_MT	0.066
SMT Hiero	0.119	TOSHIBA 2	0.066
Sense	0.245	Kyoto-U 2	0.163
RBMT B	0.217	Online A	0.239
NII 1	0.119	RBMT A	0.130
NII 2	0.086	Ave. 0.254	Ave. 0.200
TMU 1	0.091	RBMT B	0.083
TMU 2	0.136	Ave. 0.106	Ave. 0.101

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

Case Study

Submission ID	BLEU	RIBES	HUMAN	Forest Input
WEBLIO-EJ1 1	32.53	0.782	43.25	No
WEBLIO-EJ1 2	32.69	0.785	36.00	Yes

VS.

WEBLIO-EJ1 1 **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 and Future Perspective

- 12 participants for the evaluation task
 - including 3 companies and 3 outside Japan
- NAIST team achieved the best results for all the subtasks
- Shared the findings of MT for scientific papers
- 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
- Automatic evaluation server will keep running even after the workshop to promote continuous evolution of MT research show: <http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/> submit: <http://lotus.kuee.kyoto-u.ac.jp/WAT/submit/> **KEEP ON UPDATING THE STATUS OF YOUR MT ENGINE!!**
- WAT will be held annually
- The importance of both sub/super sentential context
- Let's share your resource monolingual/bilingual corpora, dictionaries and so on

**Thank you very much for attending WAT2014
 Stay tuned for the next WAT workshop!**