Translation systems and experimental results of the EHR group for WAT2016 tasks

Terumasa EHARA

Ehara NLP Research Laboratory

http://www.ne.jp/asahi/eharate/eharate/

Participated tasks and used techniques

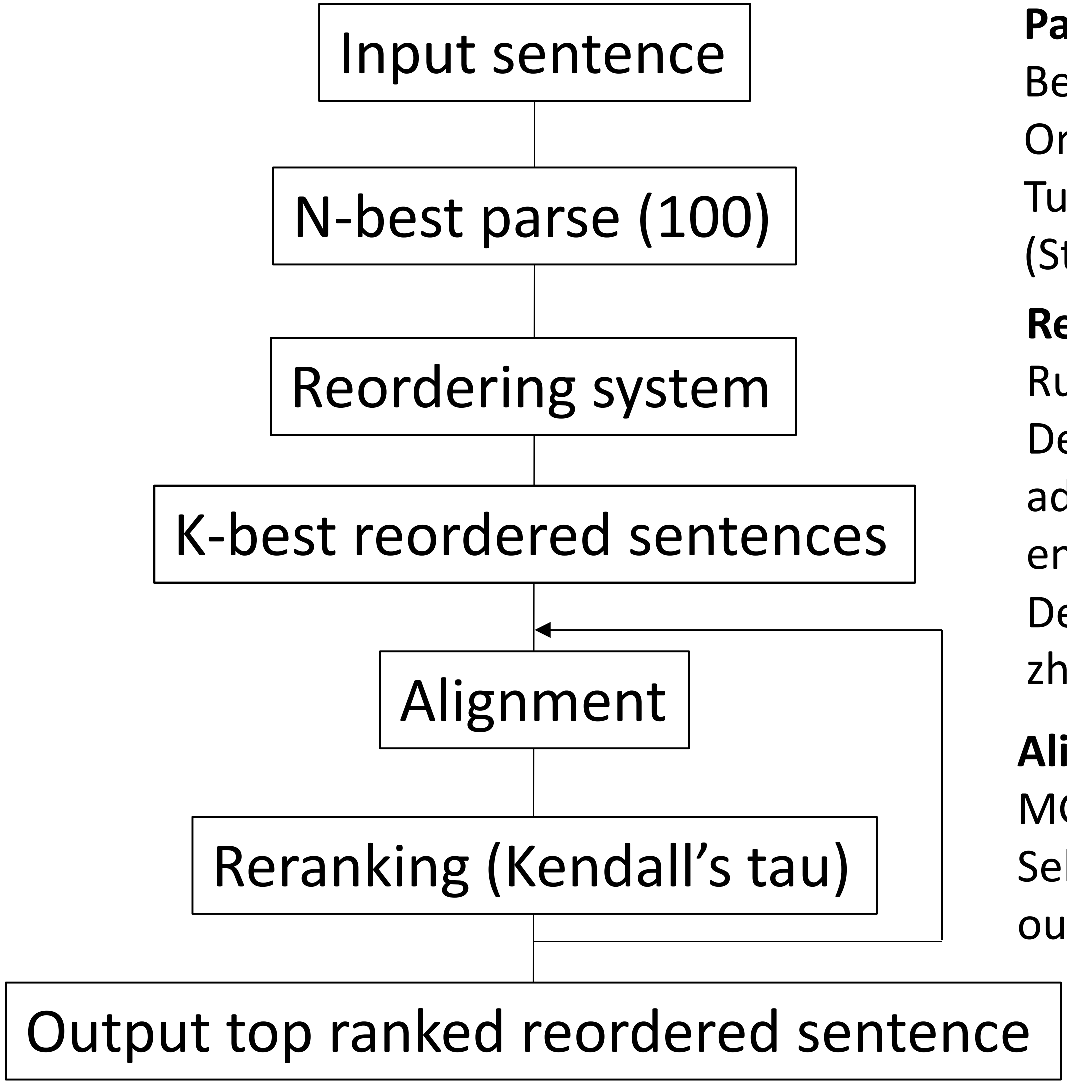
Task	Word-based PBSMT	Character-based PBSMT	RBMT+SPE	Reordering	Pivoting
en-ja	✔			✔	
zh-ja	✔	✔		✔	
JPCzh-ja	✔	✔	✔	✔	
JPCko-ja	✔	✔			
HINDENen-hi	✔			✔	
HINDENhi-ja	✔			✔	✔

PBSMT : Moses V.3, MGIZA++ v. 0.7.0
RBMT : Commercial system
SPE : Statistical post editing by Moses

PBSMT setting

- TM training data filtering out
  > 100 words
  ratio of word numbers is > 4 or < 0.25
- TM training and decoding
  Moses V.3, MGIZA++ v. 0.7.0
- LM training lmplz order=6
- Distortion limit
  0 (JPCko-ja task)
  6 (other tasks)

Reordering system (training data)

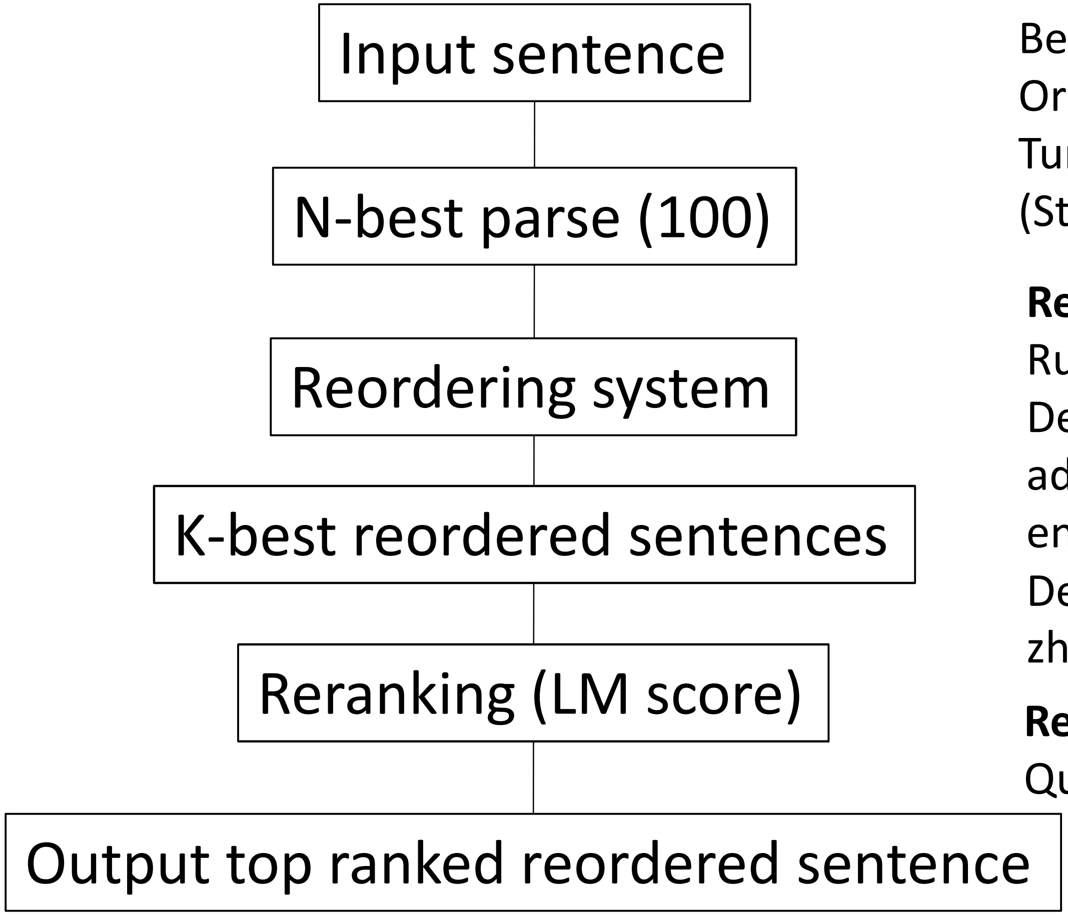

Input sentence
N-best parse (100)
Reordering system
K-best reordered sentences
Alignment
Reranking (Kendall’s tau)
Output top ranked reordered sentence

Parser
Berkeley parser
Original rule for en
Tuned rule for zh
(Stanford parser)

Reordering system
Rule based.
Deletion of articles (a, an and the) and adding case markers (subj and obj) for en-ja and en-hi tasks.
Deletion of case markers (が, は, を) for zh-ja task.

Aligner
MGIZA++
Self-built post processor for GIZA++ outputs (A3.finals, lex.f2e and lex.e2f)

Reordering system (dev devtest and test data)



Parser
Berkeley parser
Original rule for en
Tuned rule for zh
(Stanford parser)

Reordering system
Rule based.
Deletion of articles (a, an and the) and adding case markers (subj and obj) for en-ja and en-hi tasks.
Deletion of case markers (が, は, を) for zh-ja task.

Reranking
Query command in Moses

en-ja task and HINDENen-hi task

- Moses tokenizer for en
- Indic NLP normalizer and tokenizer for hi
- JUMAN for ja

- Training corpus size
  TM 1,502,767 (en-ja) alignment score≧0.08
  1,450,896 (en-hi) filter out 21,637 data
  LM 3,824,408 (en-ja) from en-ja and hi-ja tasks
  1,599,708 (en-hi) from en-hi and ja-hi tasks

zh-ja task and JPCzh-ja task

- Stanford Chinese segmenter
  plus self-built post processor for zh
- JUMAN plus self-built post processor for ja
- RBMT+SPE for JPCzh-ja task
- Character base only for zh side
- Merging of three outputs (word based SMT, character based SMT and RBMT+SPE) by LM score

zh-ja task and JPCzh-ja task

- Training corpus size
  TM 667,922 (zh-ja) zh-ja task data
  995,385 (JPCzh-ja) JPCzh-ja task data
  LM 3,680,815 (zh-ja) from zh-ja and
  en-ja task data
  4,186,284 (JPCzh-ja) from JPCzh-ja task and
  NTCIR-10's en-ja task data

JPCko-ja task

- Mecab-ko for ko tokenizer
- JUMAN for ja segmenter
- Character base both for ko and ja side
- Merging of two outputs (word based SMT and character based SMT) by LM score
- Handling of parentheses surrounding a number :
  (1) delete paren. to ko side
  (2) add paren. to ja side
  (3) add paren. to ja side and delete them after decoding

JPCko-ja task

- Training corpus size
  TM 996,339 JPCko-ja task data
  LM 5,186,284 from JPCko-ja, JPCzh-ja
  and NTCIR-10’s en-ja task data

HINDENhi-ja task

- Four methods is conducted :
  (1) Simple PBSMT (direct translation)
  (2) Sentence level pivoting without reordering
  (3) Sentence level pivoting with reordering
  (4) Table level pivoting with reordering
- User dictionary : 931 words (OOV in dev and test)
- Training corpus size for the method (1)
  TM 149,743 hi-ja task data plus user dictionary
  LM 406,766 hi-ja task data plus TED corpus

Sentence level pivoting

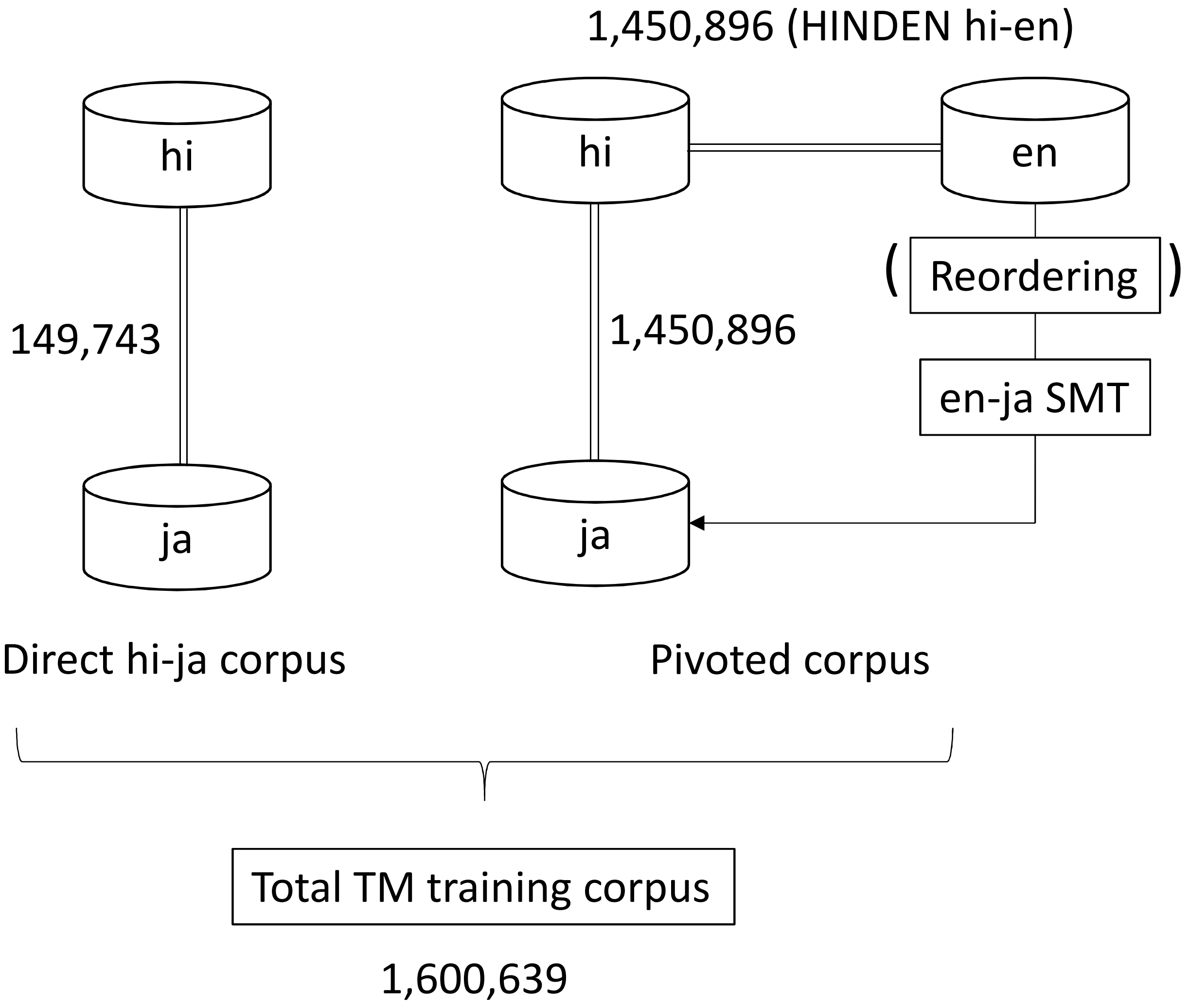
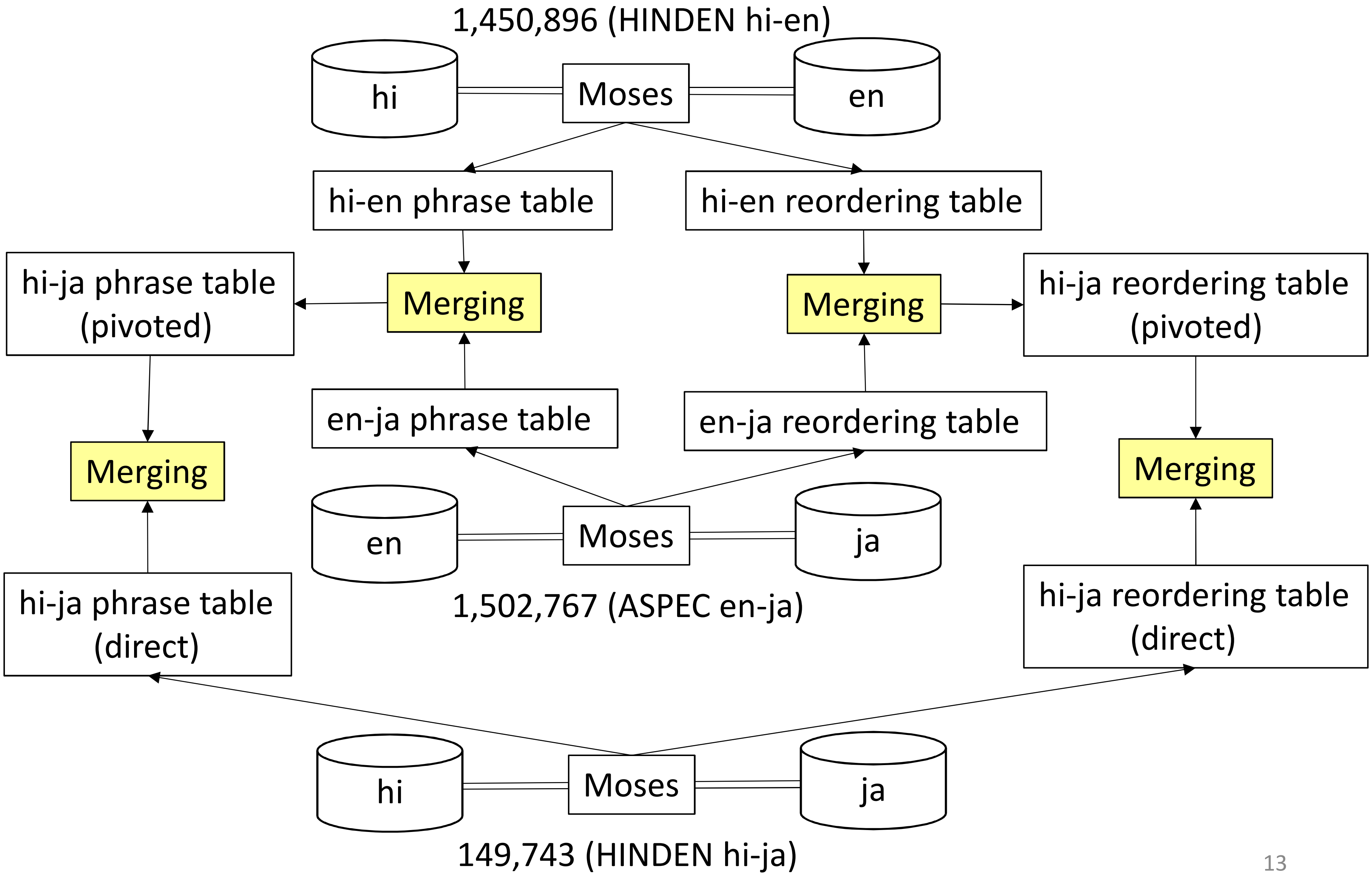


Table level pivoting



Phrase table pivoting

$$\emptyset(f|e) = \sum_{p} \emptyset(f|p)\ \emptyset(p|e)$$

$$lex(f|e) = \sum_{p} lex(f|p)\ lex(p|e)$$

$$\emptyset(e|f) = \sum_{p} \emptyset(e|p)\ \emptyset(p|f)$$

$$lex(e|f) = \sum_{p} lex(e|p)\ lex(p|f)$$

filter out for
$\emptyset(f|e)\emptyset(e|f) < 0.000001$

f : source (hi)

e : target (ja)

p : pivot (en)

Phrase table merging

$$\emptyset(f|e) = \frac{\emptyset_p(f|e)\,\mathrm{F}_p(f) + \emptyset_d(f|e)\,\mathrm{F}_d(f)}{\mathrm{F}_p(f) + \ \mathrm{F}_d(f)}.$$

p : pivoted
d : direct
Fp : frequency in the pivoted corpus
Fd : frequency in the direct corpus

Pivoted reordering table orientation

fp＼pe	m	s	d
m	m	s	d
s	s	m	s
d	d	s	m

fp : source (hi) to pivot (en) orientation

pe : pivot (en) to target (ja) orientation

m : monotone

s : swap

d : discontinuous

Reordering table pivoting

$$\mathrm{m}\ (\mathrm{f} \rightarrow \mathrm{e}) = \sum\nolimits_{p} \{\mathrm{m}(\mathrm{f} \rightarrow \mathrm{p})\mathrm{m}(\mathrm{p} \rightarrow \mathrm{e}) + \mathrm{s}(\mathrm{f} \rightarrow \mathrm{p})\mathrm{s}(\mathrm{p} \rightarrow \mathrm{e})$$

$$+\ \mathrm{d}(\mathrm{f} \rightarrow \mathrm{p})\mathrm{d}(\mathrm{p} \rightarrow \mathrm{e})\}/D$$

$$\mathrm{s}(\mathrm{f} \rightarrow \mathrm{e}) = \sum\nolimits_{p} \{\mathrm{m}(\mathrm{f} \rightarrow \mathrm{p})\mathrm{s}(\mathrm{p} \rightarrow \mathrm{e}) + \mathrm{s}(\mathrm{f} \rightarrow \mathrm{p})\mathrm{m}(\mathrm{p} \rightarrow \mathrm{e})$$

$$+\ \mathrm{d}(\mathrm{f} \rightarrow \mathrm{p})\mathrm{s}(\mathrm{p} \rightarrow \mathrm{e}) + \mathrm{s}(\mathrm{f} \rightarrow \mathrm{p})\mathrm{d}(\mathrm{p} \rightarrow \mathrm{e})\}\ /D$$

$$\mathrm{d}(\mathrm{f} \rightarrow \mathrm{e}) = \sum\nolimits_{p} \{\mathrm{m}(\mathrm{f} \rightarrow \mathrm{p})\mathrm{d}(\mathrm{p} \rightarrow \mathrm{e}) + \mathrm{d}(\mathrm{f} \rightarrow \mathrm{p})\mathrm{m}(\mathrm{p} \rightarrow \mathrm{e})\}\ /\mathrm{D}$$

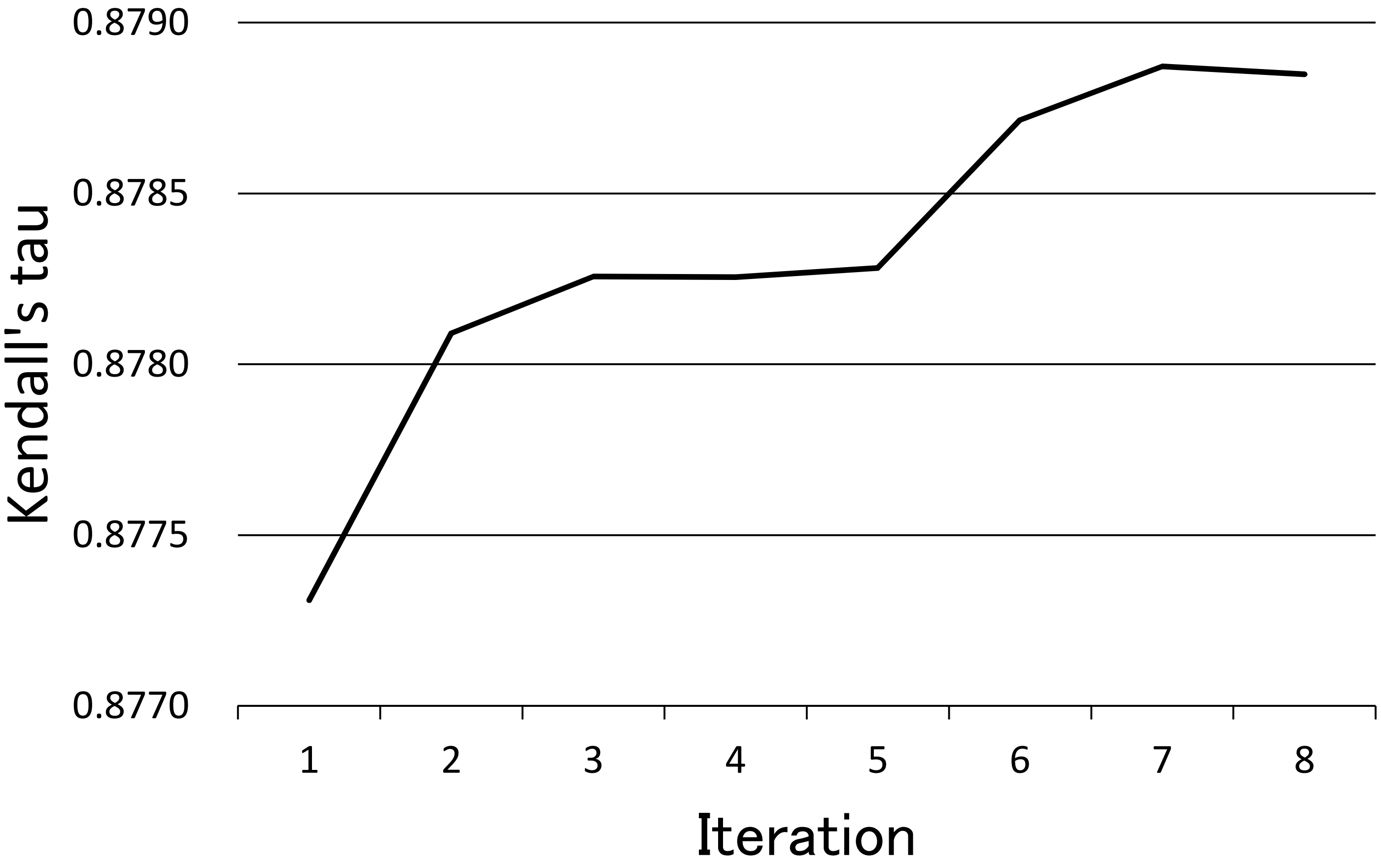
D : normalizer such that $\mathrm{m}(\mathrm{f} \rightarrow \mathrm{e}) + \ \mathrm{s}(\mathrm{f} \rightarrow \mathrm{e}) + \mathrm{d}(\mathrm{f} \rightarrow \mathrm{e}) = 1$

Reordering table merging

$$\mathrm{a}(\mathrm{f} \rightarrow \mathrm{e}) = \frac{\mathrm{a}_{p}(\mathrm{f} \rightarrow \mathrm{e})\ \mathrm{F}_{p}(\mathrm{f}) + \mathrm{a}_{d}(\mathrm{f} \rightarrow \mathrm{e})\ \mathrm{F}_{d}(\mathrm{f})}{\mathrm{F}_{p}(\mathrm{f}) + \ \mathrm{F}_{d}(\mathrm{f})}$$

a : {m|s|d}
p : pivoted
d : direct
Fp : frequency in the pivoted corpus
Fd : frequency in the direct corpus

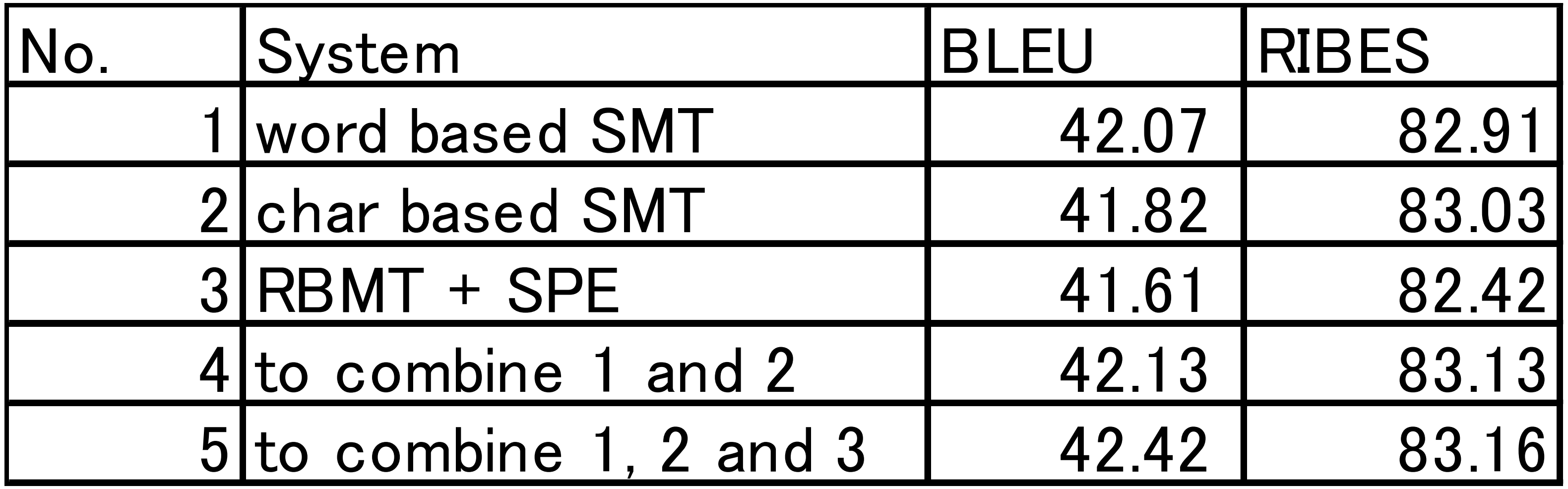
Results of iterative reordering (JPCzh-ja)



Results of iterative reordering

Task	Iteration	Kendall's tau
en-ja	4	0.7655
zh-ja	4	0.9083
JPCzh-ja	8	0.8788
HINDENen-hi	4	0.8398

Evaluation result of system combination (JPCzh-ja)



No.	System	BLEU	RIBES
1	word based SMT	42.07	82.91
2	char based SMT	41.82	83.03
3	RBMT + SPE	41.61	82.42
4	to combine 1 and 2	42.13	83.13
5	to combine 1, 2 and 3	42.42	83.16

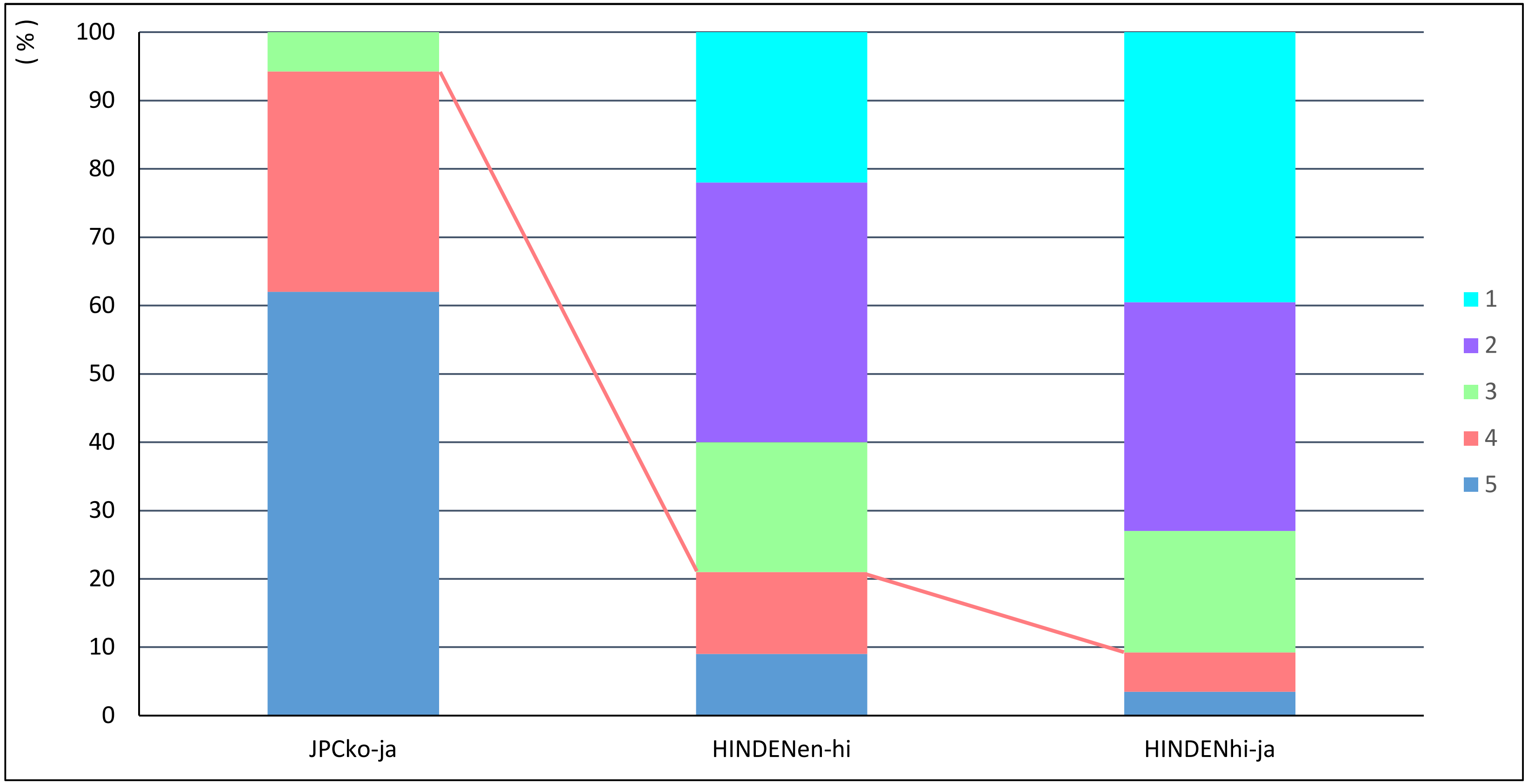
Evaluated systems

Task	System No.	Word-based PBSMT	Character-based PBSMT	RBMT+SPE	Reordering	Sentence level pivoting	Table level pivoting	Parenthes handling
en-ja	1	✔			✔			
zh-ja	1	✔	✔		✔			
JPCzh-ja	1	✔	✔	✔	✔			
	2	✔	✔		✔			
JPCko-ja	1	✔	✔					del
	2	✔	✔					add & del
	3	✔	✔					add
HINDENen-hi	1	✔			✔			
HINDENhi-ja	1	✔			✔		✔	
	2	✔			✔	✔		
	3	✔				✔		
	4	✔						

Evaluation results

Task	System No.	BLEU	RIBES	AMFM	HUMAN	HUMAN (top team)	JPO adq.	JPO adq. (top team)
en-ja	1	31.32	0.7599	0.7467	39.000	55.250	---	4.02
zh-ja	1	39.75	0.8437	0.7695	32.500	63.750	---	3.94
JPCzh-ja	1	41.05	0.8270	0.7350	35.500	46.500	---	3.44
	2	40.95	0.8280	0.7451	39.000		---	
JPCko-ja	1	71.51	0.9447	0.8664	-3.000	21.750	---	4.62
	2	68.78	0.9411	0.8517	---		---	
	3	62.33	0.9271	0.8180	21.750		4.56	
HINDENen	1	11.75	0.6719	0.6508	0.000	57.250	2.48	2.55
HINDENhi	1	7.81	0.5793	0.4681	13.750	39.750	2.00	2.13
	2	7.66	0.5860	0.4731	10.000		---	
	3	7.47	0.5823	0.4549	---		---	
	4	2.36	0.4402	0.3628	---		---	

Evaluation results by JPO adequacy



Conclusion

- Our translation techniques are effective
  Iterative reordering
  System combination
  Pivoting with reordering
- Remaining issues
  To improve parsing accuracy
  To improve hi-ja and en-hi accuracy
  To challenge MT for other Asian languages
  (Indonesian, Thai, Vietnamese, Mongolian, etc.)