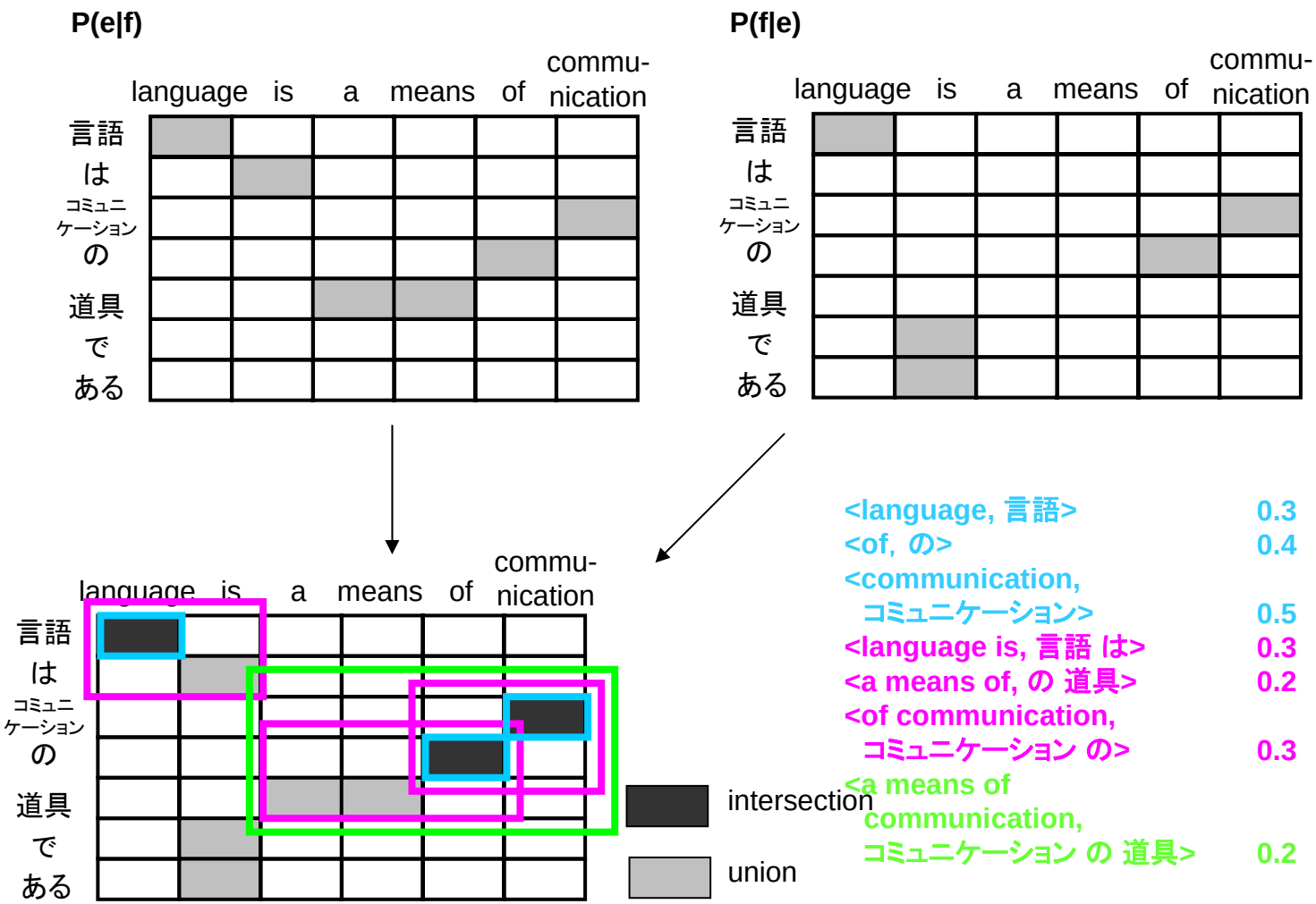
NTT SMT System 2008 at NTCIR-7

Taro Watanabe, Hajime Tsukada,
and Hideki Isozaki
NTT Communication Science Laboratories

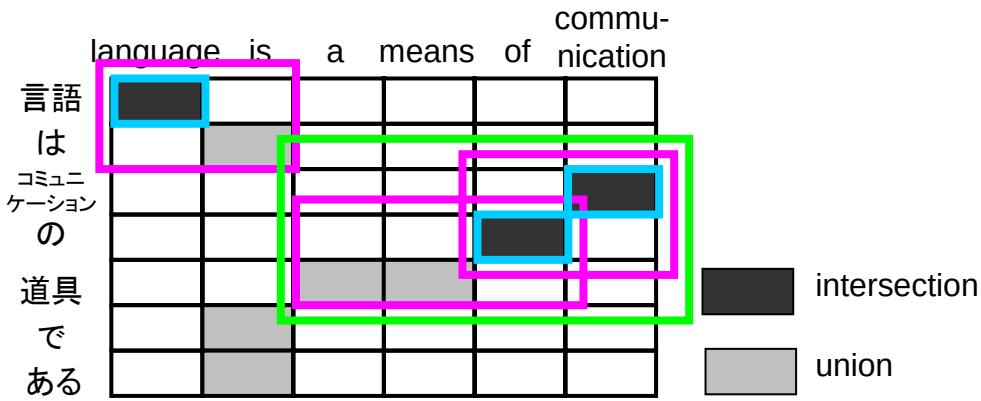
Summary

- Hierarchical phrase-based statistical machine translation system
- The algorithm of Chiang (2007) is faithfully followed as a baseline of our research
- Performance is very competitive with top ranked systems

Phrase Model



Hierarchical Phrase Model (Chiang 2007)



X→<language, 言語> 0.3
X→<of, の> 0.4
X→<communication, コミュニケーション> 0.5
X→<language is, 言語 は> 0.3
X→<a means of, の 道具> 0.2
X→<of communication, コミュニケーション の> 0.3
X→<a means of communication, コミュニケーション の 道具> 0.2

Additional rules

X→<X is, X は> 0.4
X→<of X, X の> 0.3
X→<X communication, コミュニケーション X> 0.3
X→<a means X, X 道具> 0.3

- Weighted synchronous CFGs are obtained from bilingual corpora
- Each internal phrase is generalized by nonterminal symbol X
- CKY-based algorithm with cube-pruning is used for decoding

Experimental Conditions

- Basically, identical conditions are used in J-to-E and E-to-J

- Primary run:
  - PSD-1(J-to-E) for TM and LM

- Contrastive run:
  - PSD-1(J-to-E) + PSD-2(E-to-J) for TM
  - PPD-1(J-to-E) + PPD-2(E-to-J) for LM
  - English Web-1T 5-gram / Japanese Web-1T 7-gram

Experimental Conditions (cont'd)

- Cases are preserved
- Normalized based on NFKC, an Unicode standard for encoding normalization
- Japanese tokenization: MeCab
- English tokenization follows English Web 1T data
- Old style English symbol notations are converted into new styles

Experimental Conditions (cont'd)

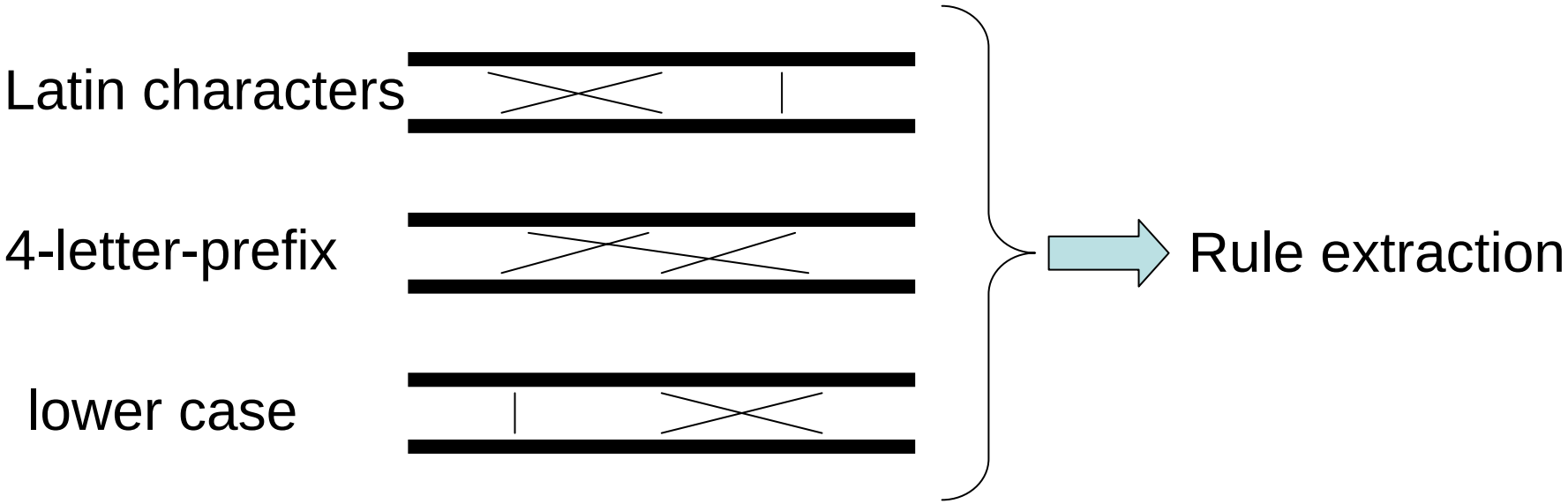
- Word alignment based on HMM (Vogel et al. 1996)
- Extracted rules contain 5 terminals at most
- Devset sentences at most 40 words long (2000 sentences) are used for MERT
- Various token normalizations for word alignment

Token Normalization

	Transistors	トランジスタ
Latin characters	transistors ⟷	toranjisuta
4-letter-prefix	Tran+ ⟷	トランジ+
lower case	transistors ⟷	トランジスタ

- Normalized tokens are used for word alignment

Rule Extraction

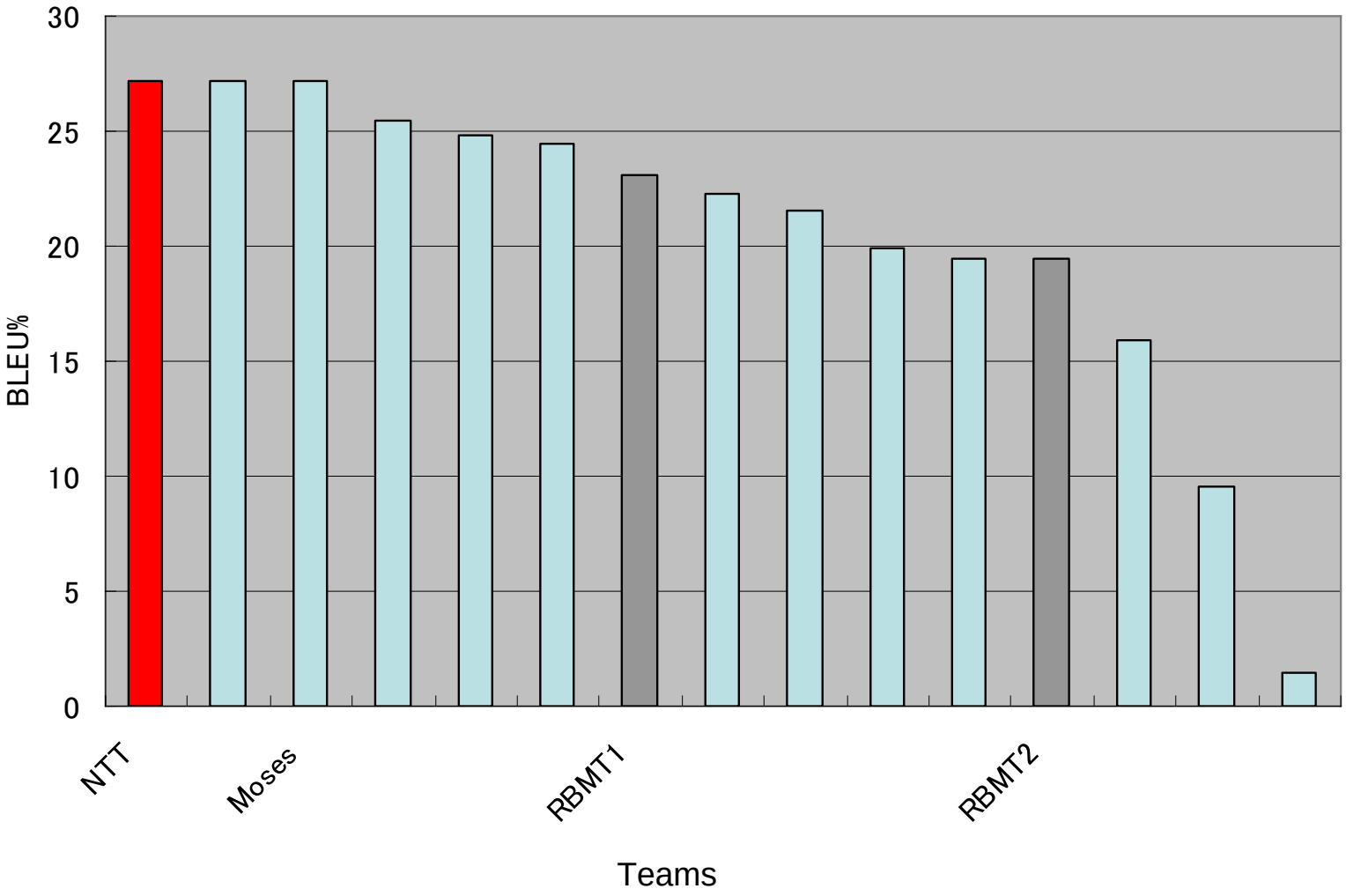


- Word alignment may differ for each token normalization
- Rules are extracted by considering various word alignments

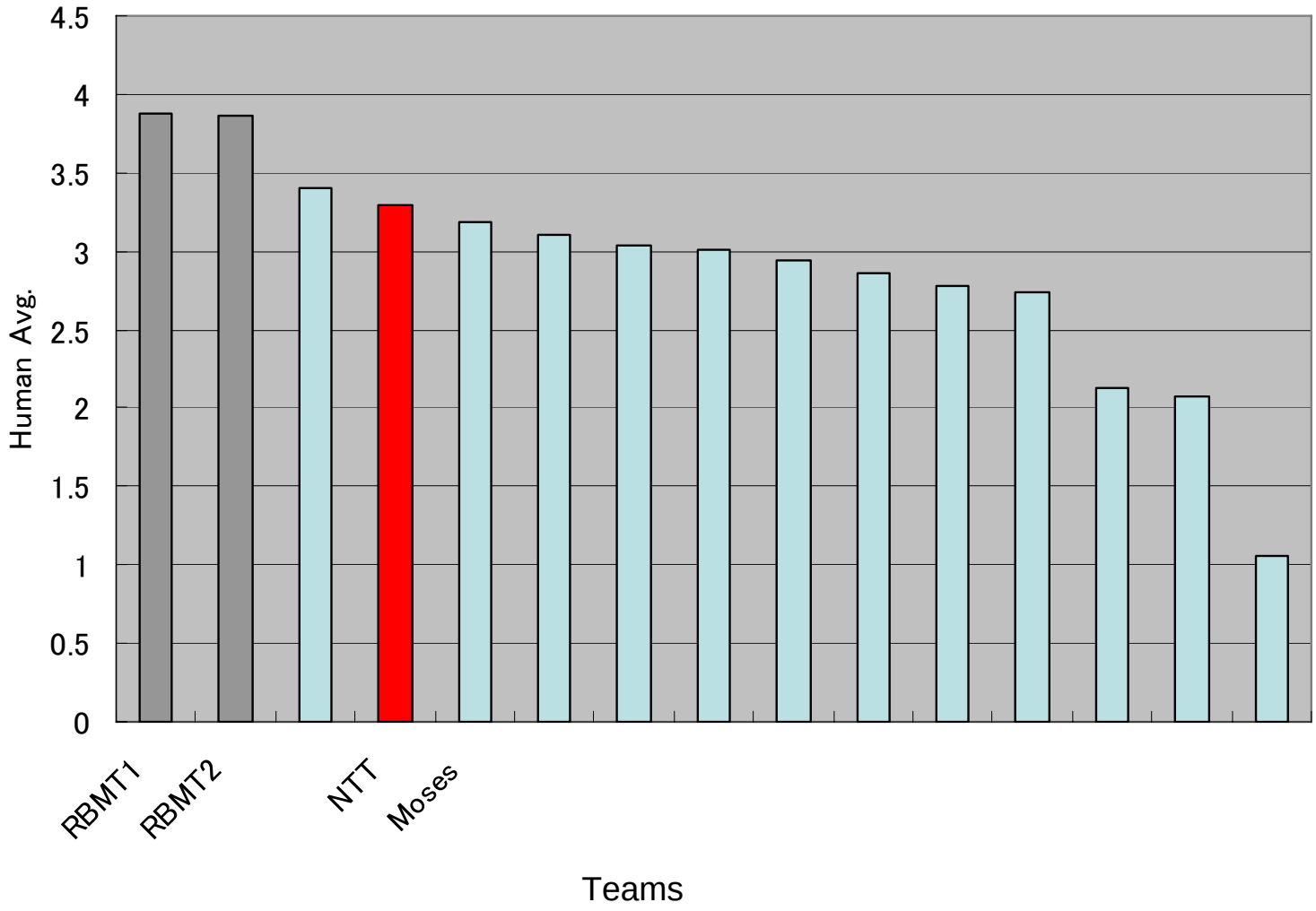
Results

	sBLEU	mBLEU (m300-DE)
J-to-E primary	**27.20**	35.93
+ Web 1T/PSD-2/PPD	26.88	**36.05**
E-to-J primary	**28.07**	
+Web 1T/PSD-2/PPD	27.20	

BLEU-JE-all



Human Evaluation (JE)



Conclusion

- Hierarchical phrase-based system (Chiang 2007) was employed
- Performance was very competitive with top ranked systems
- Additional corpora (PSD-2, PPD-1,2) and Web-1T n-gram were less helpful than expected