NUT-NTT
Statistical Machine Translation System
for IWSLT 2005

Kazuteru Ohashi and Kazuhide Yamamoto
Nagaoka University of Technology

Kuniko Saito and Masaaki Nagata
NTT Cyber Space Laboratory

Outline

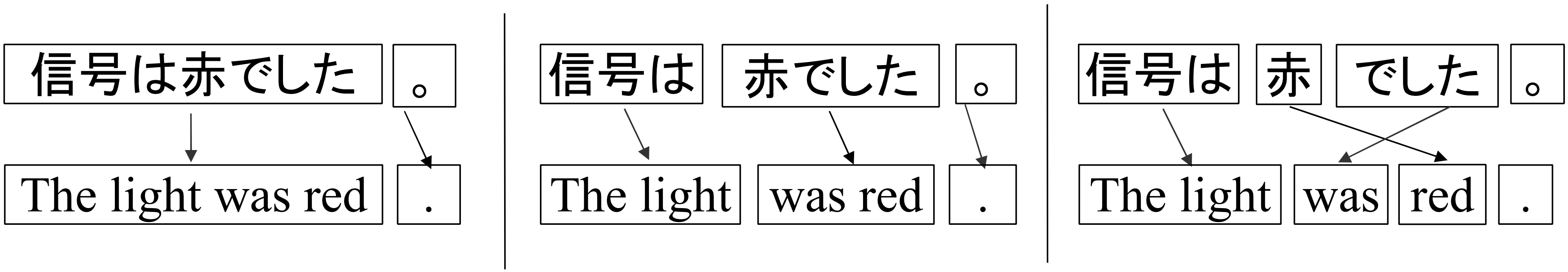
- We present
  - Novel distortion model for phrase-based SMT
  - Novel phrase alignment algorithm to compute the distortion model
- Out line of this talk
  - Motivation
  - Baseline system
  - Improvements
  - Experiments

Motivation

- Previous phrase-based translation models are not effective for global phrase reordering
  - Because they simply penalize non-monotonic alignments (Koehn et. al. 2003) (Och and Ney 2004)
  - It is difficult to handle complex reordering required for the translation between Japanese and English
- In order to compute phrase distortion model,
  - Phrase alignment for a pair of sentences is required
  - Method for accurate phrase alignment is not studied well, as far as we know.

Approach (1/2)
Phrase alignment

- We get N-best phrase alignments

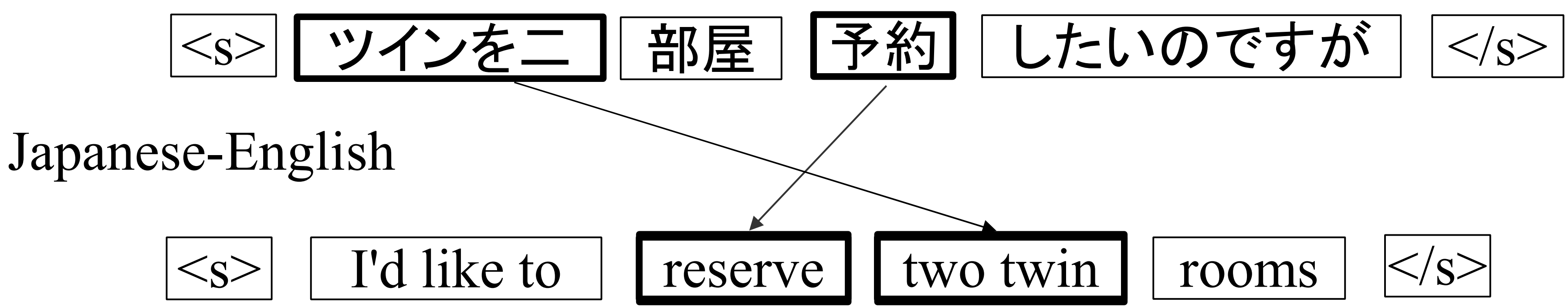


3-best alignments

- N-best phrase alignments are used for calculating phrase distortion probabilities and phrase translation probabilities

Approach (2/2)
Phrase distortion model

- We define phrase distortion model as
  - Probability of relative distance between two source language phrases that are aligned to two adjacent target language phrases



- We classify relative distance into four states

Baseline system (1/4)
phrase-based translation model

- Model(in Foreign-English translation)

$$\hat{e} = argmax_e \, p(e|f) = argmax_e \, p(f|e) \, p(e)$$

Source sentence f is segmented into phrases $\bar{f}_1^I$

Target sentence e is segmented into phrases $\bar{e}_1^I$

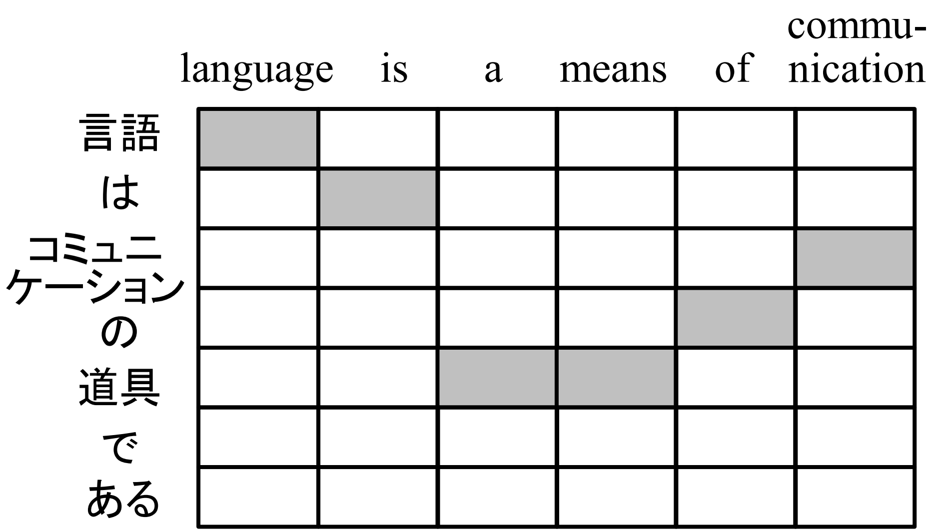
$$p(\bar{f}_1^I | \bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i | \bar{e}_i) \, d(a_i - b_{i-1})$$

$\phi(\bar{f}_i | \bar{e}_i)$ is phrase translation probability

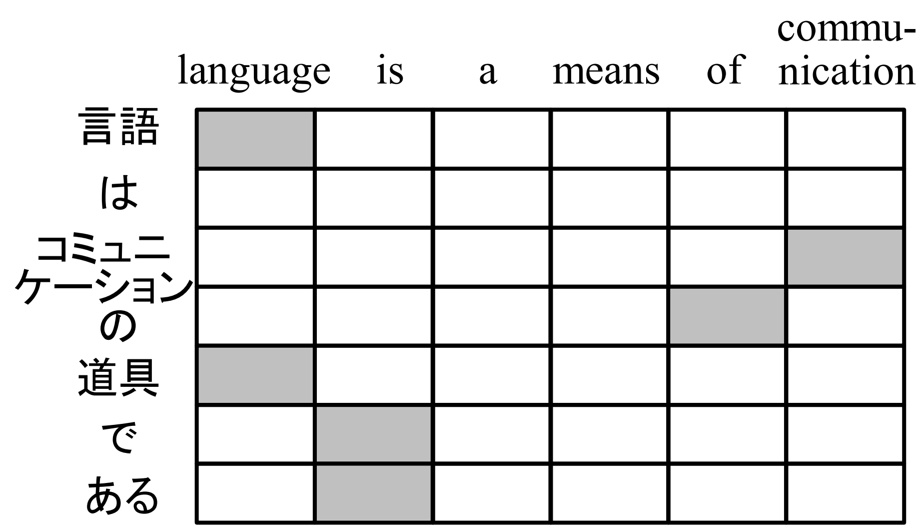
$d(a_i - b_{i-1})$ is phrase distortion probability

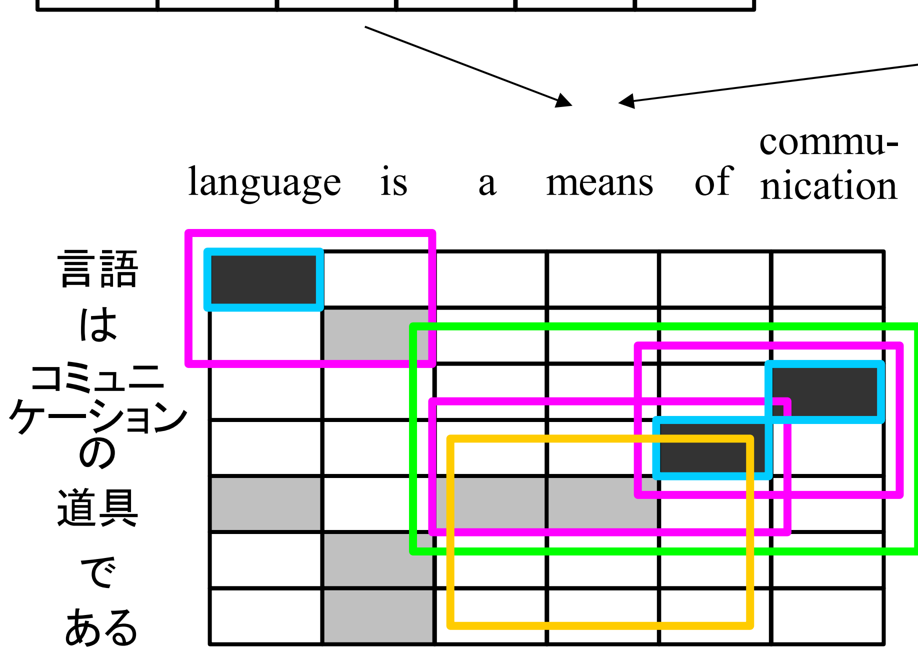
Baseline system (2/4)
phrase extraction (not phrase alignment)

Japanese to English alignment (IBM Model4)



English to Japanese alignment (IBM Model4)





(言語,language)
(の,of)
(コミュニケーション,communication)

(言語は, language is)
(の道具, a means of)
(コミュニケーションの, of communication)

(コミュニケーションの道具,
a means of communication)

(の道具である, a means of)

intersection union

Baseline system (3/4)
phrase translation probability

- translation probability
  - relative frequency

$$\phi(\bar{f}|\bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}} count(\bar{f}, \bar{e})}$$

Baseline system (4/4)
Phrase distortion model

- Penalty consider two features
  - a_i : the start position of the source phrase for target phrase
  - b_{i-1} : the end position of the source phrase for previous target

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|}$$

current previous

0 1 2 3 4 5 6

| <s> | ツインを二 | 部屋 | 予約 | したいのですが | </s> |

Japanese-English

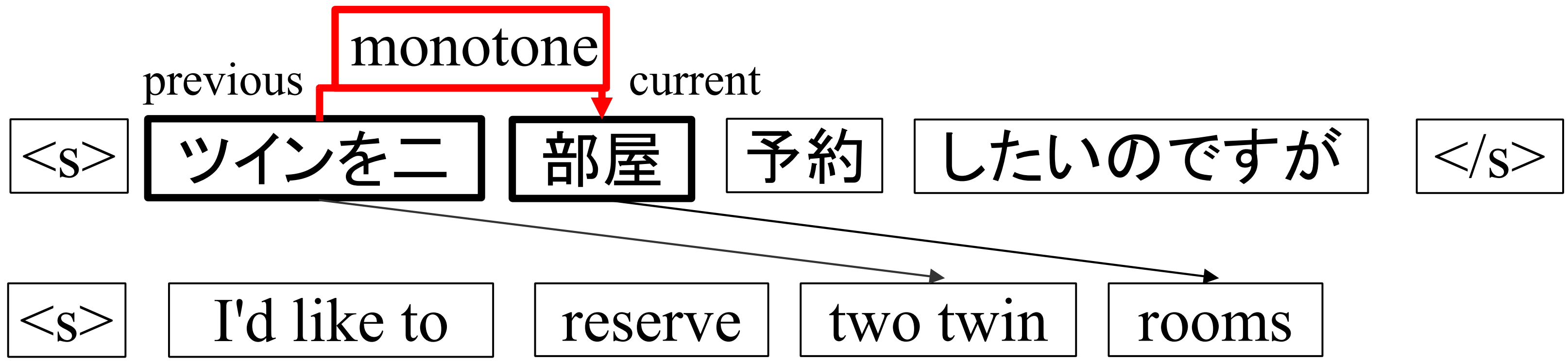
| <s> | I'd like to | reserve | d=\|3-5\| = 2

- Considering relative position between phrases only

Proposed phrase distortion model

- We define phrase distortion model as
  - $p\left(d \middle| \bar{e}_{i-1}, \bar{e}_i, \bar{f}_{i-1}, \bar{f}_i\right)$
  - $\bar{e}_{i-1}$ and $\bar{e}_i$ are adjacent two target phrases
  - $\bar{f}_{i-1}$ and $\bar{f}_i$ are source phrases aligned to $\bar{e}_{i-1}$ and $\bar{e}_i$
  - d is relative distance between $\bar{f}_{i-1}$ and $\bar{f}_i$
- We classify d into 4 states
  - monotone, monotone-gap, reverse, reverse-gap

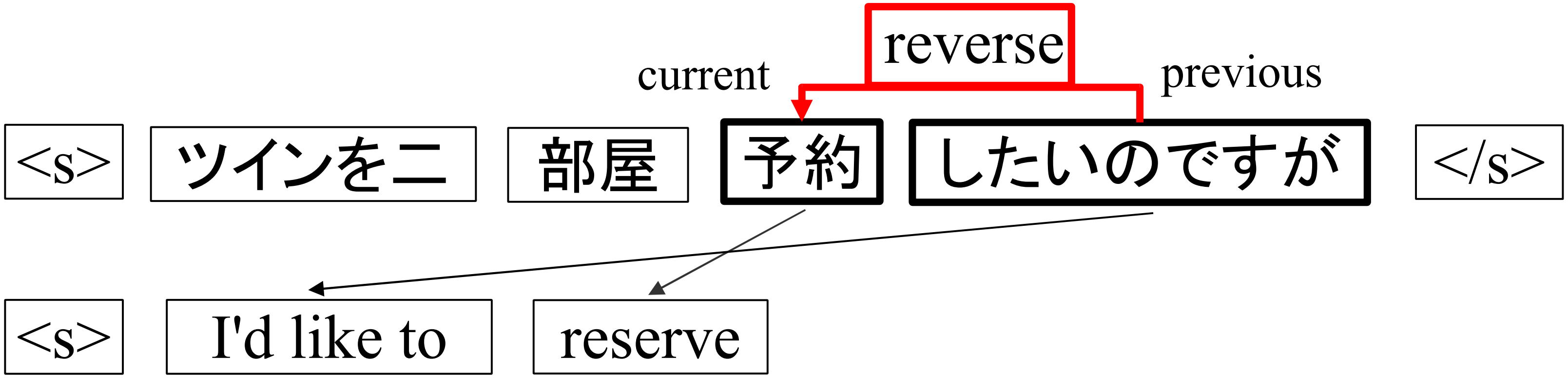
Monotone and monotone-gap



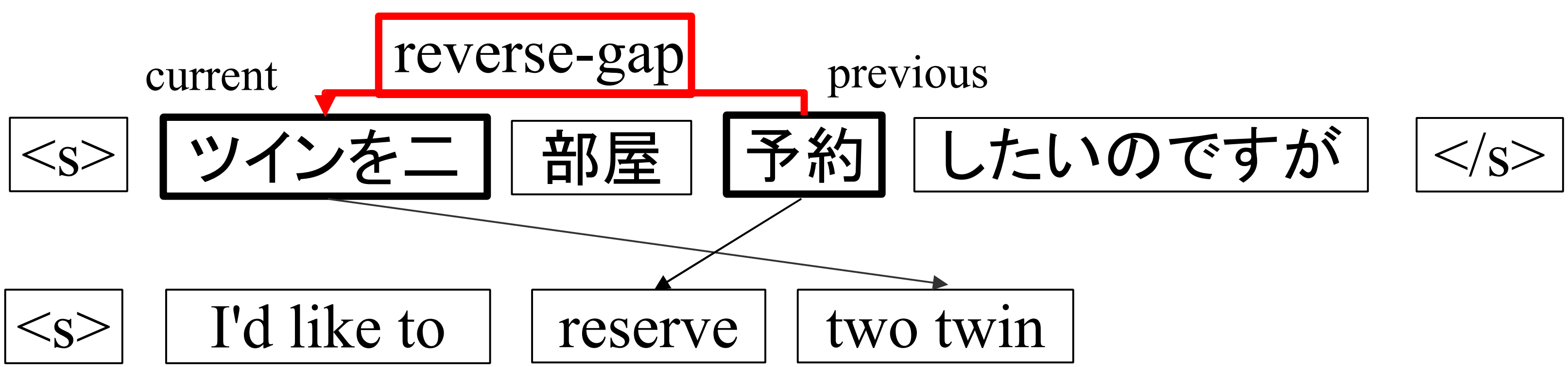
- Two source language phrases for the adjacent two target phrases, “two twin” and “rooms”, are
  - Same order (monotone) and adjacent (without gap)

  - Same order (monotone) and not adjacent (with gap)

Reverse and reverse-gap



- Two source phrases for the adjacent two target phrases, “I’d like to” and “reserve”, are
  - Not same order (reverse) and adjacent (without gap)



  - Not same order (reverse) and not adjacent (with gap)

Proposed phrase distortion model

- We classify each phrase by the part of speech
  - Single POS
    - English and Chinese ... first word of each phrase
    - Japanese ... last word of each phrase

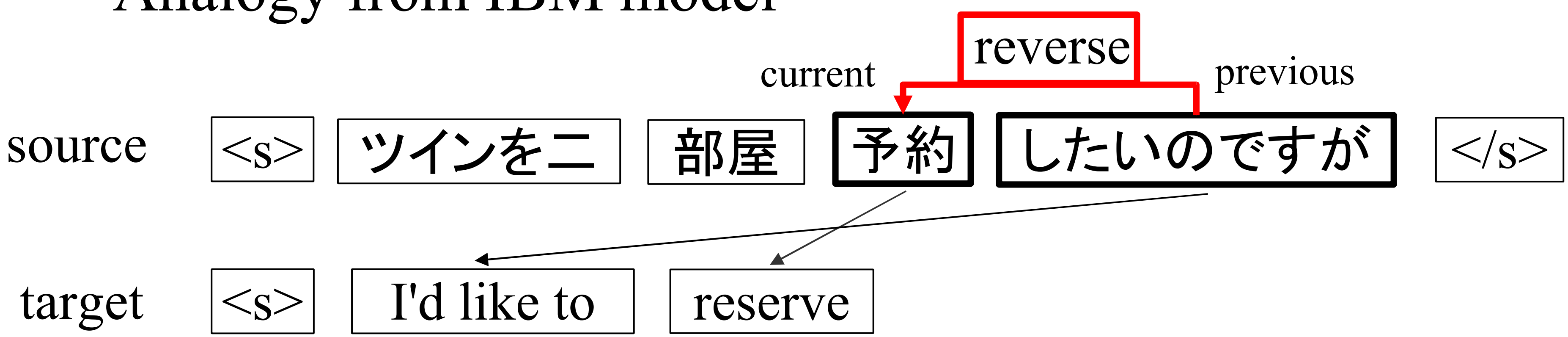
ex)

信号 は	particle
赤 でし た	auxiliary verb
the light	article
was red	verb

  - Double POS
    - First and last word of each phrase for any languages

Proposed phrase distortion model

- We consider a series of distortion models that have increasingly complex dependencies
  - Analogy from IBM model



Type1: $p(d)$

Type2: $p(d|class(\bar{f}_i))$

Type3: $p(d|class(\bar{e}_{i-1}),class(\bar{f}_i))$

Type4: $p(d|class(\bar{e}_{i-1}),class(\bar{f}_{i-1}),class(\bar{f}_i))$

Type5: $p(d|class(\bar{e}_{i-1}),class(\bar{e}_i),class(\bar{f}_{i-1}),class(\bar{f}_i))$

Phrase alignment

- We search for the segmentation of bilingual sentences that maximizes the product of lexical translation probabilities

$$\hat{\bar{f}}_{1,}^{I}\hat{\bar{e}}_{1}^{I}=argmax_{\bar{f}_{1,}^{I}\bar{e}_{1}^{I}}\prod_{i=1}^{I}p(\bar{f}_{i}|\bar{e}_{i})$$

- Lexical translation probability (Phrase translation probability) is defined in (Vogel et. al. 2003)

$$p(\bar{f}|\bar{e})=\prod_{j}\sum_{i}p(f_{j}|e_{i})$$

Phrase alignment

- Search steps
    1. Consider all combinations of phrase from each language
    2. Delete candidates by threshold of lexical translation probability
    3. Search for consistent phrase alignment among all combinations of the above phrase translation candidates
- We can obtain the N-best phrase alignment by using A* search (Ueffing and et. al. 2002)

Phrase alignment

- 1. Consider all combinations of phrase

ex) 部屋 を 予約 し たい の です が

I'd like to reserve a room

部屋	I	1e-10	部屋を	I	1e-17
部屋	I'd	1e-15	部屋を	I'd	1e-23
…	…	…	…	…	…
部屋	room	0.5	部屋を	room	0.1
…	…	…	…	…	…

Phrase alignment

- 2. Delete candidates by threshold of lexical translation probability

ex) 部屋 を 予約 し たい の です が

I 'd like to reserve two twin rooms

~~部屋~~	~~I~~	~~1e-10~~	~~部屋を~~	~~I~~	~~1e-17~~
~~部屋~~	~~I'd~~	~~1e-15~~	~~部屋を~~	~~I'd~~	~~1e-23~~
…	...	...	…	...	...
部屋	room	0.5	部屋を	room	0.1
…	…	…	…	…	…

Phrase alignment

- 3. Search for consistent phrase alignment
  - All words are to be included in a single phrase for each languages
  - Forward beam search and backward A* search(Ueffing et. Al.)
  - We get N-best phrase alignment

Corpus and Tools

- Supplied Data + Tools Track
  - Additional corpus is not used
- Japanese-English and Chinese-English
- Tokenization(segmentation) and tagging
  - English: tokenizer.sed and MXPOST
  - Japanese：ChaSen
  - Chinese: a tool developed by NTT
- English are lowercased

Corpus and Tools

- Word translation probability
  - GIZA++: IBM Model4
- Language model
  - Palmkit: back-off ngram
- Minimum error rate training
  - Tool provided by CMU (A. Venugopal 2005)

Experiments
Phrase extraction method

- Parameters of phrase alignment
  - N-best of phrase alignment : 20
  - Phrase candidate threshold : 1e-15
  - Beam width : 1000
- Translation accuracy for development set 2 of Japanese-English with different phrase extraction methods

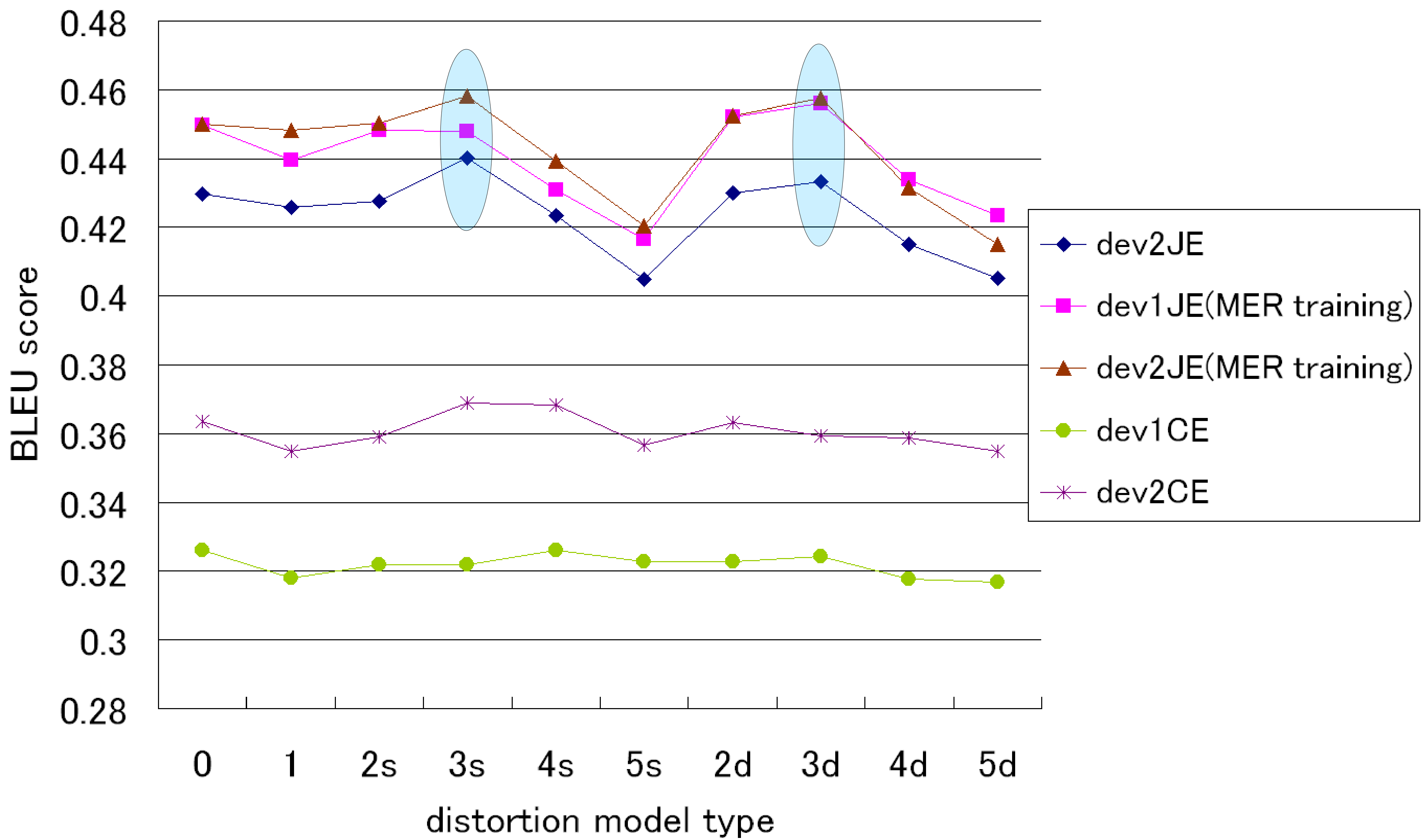
phrase extraction	NIST score	BLEU score
conventional	7.6162	0.3375
our method	**8.8159**	**0.4471**

Experiments
Phrase distortion model

- Phrase distortion models are named “Type [0-5][sd]” such as “Type 2s” and “Type 3d”
  - [0-5] represents the type of distortion model
    - 0 is baseline distortion model (aka. Pharaoh)
  - “s” (single) means each phrase is classified by the POS of one word (either the first or last word in the phrase)
  - “d” (double) means each phrase is classifed by the POS of two words (both the first and last words in the phrase)
  - We tested 11 phrase distortion model types
    - 0, 1, 2s, 3s, 4s, 5s, 2d, 3d, 4d, 5d

Experiments
Phrase distortion model

- Features for Minimum error rate training
  - Phrase translation probability(both direction)
  - Lexical translation probability(both direction)
  - Word penalty
  - Phrase distortion probability



Type 3s and 3d are slightly better than others

$$3\text{s}: p\left(d \,\middle|\, class\left(e_{i-1}^{-}\right), class\left(\bar{f}_{i}\right)\right)$$

Discussion

- We could not get phrase alignment 1095 of the 20000 training sentences(5.5%)
  - If the training parallel sentence is too long, we cannot get phrase alignment because of the large search space.
- Some countermeasure is needed
  - Limiting the search space for those long sentences by using the distortion model obtained from relatively short tentences.

Discussion

- Is the current phrase segmentation appropriate ?
  - Phrase segmentation is decided by the lexical translation probability
  - It might be better to consider not only lexical translation probability but also other probabilities such as word penalty
  - By using linguistic phrase boundaries provided by syntactic parsers, we might be able to improve the translation accuracy
  - Improvement of phrase segmentation will improve phrase classification

Conclusion

- We present
  - A novel phrase distortion model
  - A novel phrase alignment method
- The phrase distortion model described herein offers improved translation accuracy over the baseline method.

Thank you

- References

[1] P. Koehn, F.J. Och, and D. Marcu, "Statistical phrase-based translation," in HLT-NAACL 2003

[2] F.J. Och and H. Ney, "The alignment template approach to statistical machine translation," Computational Linguistics, vol. 30, no. 4, pp. 417-449, 2004.

[3] S. Vogel, Y. Zhang, F.Huang, A. Tribble, A. Venugopal, B. Zhao, and A. Waibel, "The CMU statistical machine translation system," in MT Summit IX, New Orleans, USA, 23-27, 2003.

[4] N. Ueffing, F.J. Och, and H. Ney, "Generation of word graphs in statistical machine translation," in Proceedings of the Conference on EMNLP. 2002, pp.156-163.

Examples of phrase distortion model

- Model type 2 and classified POS of last word in phrase(Japanese-English)
  - -1 名詞-副詞可能|0.380
  - -1 連体詞-連体詞|0.0595
  - -2 フィラー-フィラー|0.578
- Model type 3 and classified POS of first and last words in phrase (Japanese-English)
  - -1 名詞-非自立 名詞-副詞可能 PRP PRP|0.75
  - -1 名詞-非自立 連体詞-連体詞 DT NNS|1
  - -1 名詞-副詞可能 記号-句点 NNP NNP|0.0526

Discussion

- In distortion model type 4d and 5d, BLEU score were generally low
- This is probably caused by data sparseness
- In model type 4d, consider 8 POSs
- In model type 5d, consider 10 POSs