Statistical Machine Translation Based on Hierarchical Phrase Alignment

Taro Watanabe †, Kenji Imamura and Eiichiro Sumita

taro.watanabe@atr.co.jp

ATR Spoken Language Translation Research Laboratories

Introduction to SMT (refer to TMI Tutorial)

$$\mathbf{e} \longleftarrow \boxed{\text{Translator}} \longleftarrow \mathbf{f}$$

$$\mathbf{e} = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f})$$

Introduction to SMT (refer to TMI Tutorial)

$$\mathbf{e} \longleftarrow \boxed{\text{Translator}} \longleftarrow \mathbf{f}$$

$$\mathbf{e} = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f})$$

Apply the Bayes Rule:

$$\boxed{\text{Source Model}} \longrightarrow \mathbf{e} \longrightarrow \boxed{\text{Channel Model}} \longrightarrow \mathbf{f}$$

$$\begin{aligned} \mathbf{e} &= \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) \\ &= \arg\max_{\mathbf{e}} P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e}) \end{aligned}$$

Introduction to SMT (refer to TMI Tutorial)

$$\mathbf{e} \longleftarrow \boxed{\text{Translator}} \longleftarrow \mathbf{f}$$

$$\mathbf{e} = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f})$$

Apply the Bayes Rule:

$$\boxed{\text{Source Model}} \longrightarrow \mathbf{e} \longrightarrow \boxed{\text{Channel Model}} \longrightarrow \mathbf{f}$$

$$\begin{aligned} \mathbf{e} &= \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) \\ &= \arg\max_{\mathbf{e}} P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e}) \end{aligned}$$

- $P(\mathbf{e})$ — Language Model
- $P(\mathbf{f}|\mathbf{e})$ — Translation Model

Translation Model

- How to represent $P(\mathbf{f}|\mathbf{e})$? (a correspondence between $\mathbf{e}$ and $\mathbf{f}$)

Translation Model

- How to represent $P(\mathbf{f}|\mathbf{e})$? (a correspondence between $\mathbf{e}$ and $\mathbf{f}$)
- Introduction of $\mathbf{a}$: alignment

$$P(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

Translation Model

- How to represent $P(\mathbf{f}|\mathbf{e})$? (a correspondence between $\mathbf{e}$ and $\mathbf{f}$)
- Introduction of $\mathbf{a}$: alignment

$$P(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

- An example of alignments

NULL $\quad$ The_1 $\quad$ $poor_2$ $\quad$ $don't_3$ $\quad$ $have_4$ $\quad$ any_5 $\quad$ $money_6$

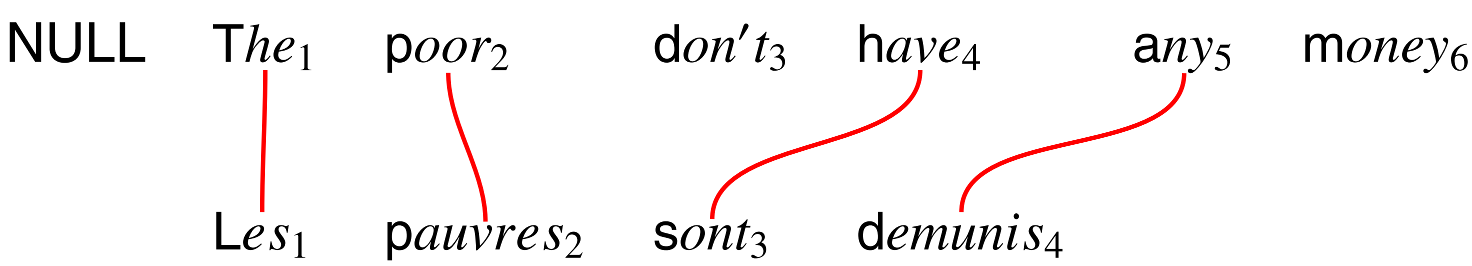
Les_1 $\quad$ $pauvres_2$ $\quad$ $sont_3$ $\quad$ $demunis_4$

Translation Model

- How to represent $P(\mathbf{f}|\mathbf{e})$? (a correspondence between $\mathbf{e}$ and $\mathbf{f}$)
- Introduction of $\mathbf{a}$: alignment

$$P(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

- An example of alignments



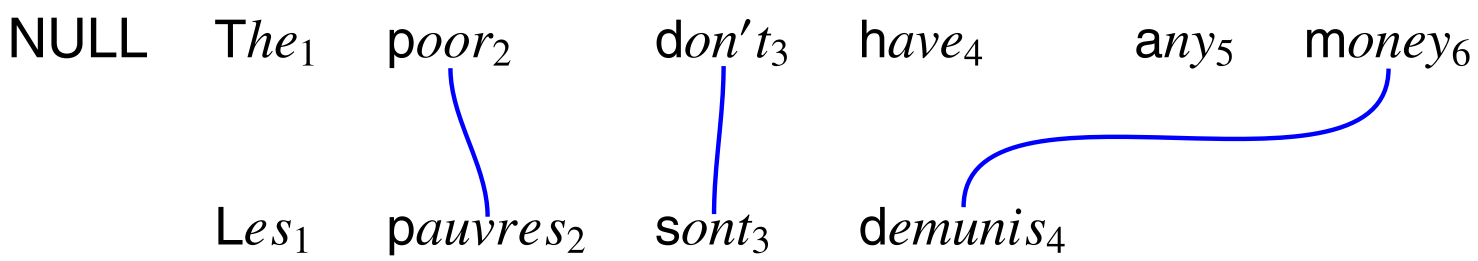
$$\mathbf{a} = (1, 2, 4, 5)$$

Translation Model

- How to represent $P(\mathbf{f}|\mathbf{e})$? (a correspondence between $\mathbf{e}$ and $\mathbf{f}$)
- Introduction of $\mathbf{a}$: alignment

$$P(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

- An example of alignments



$$\mathbf{a} = (0, 2, 3, 6)$$

Structure of TM (IBM Model 4)

Translation Model

Lexical Model

$\prod t(f_j|e_i)$

Fertility Model

$\prod n(\phi_i|e_i)$

Distortion Model

Head $\prod d_1(j-k|\mathcal{A}(e)\mathcal{B}(f))$

Non-Head $\prod d_{1>}(j-j'|\mathcal{B}(f))$

NULL Translation Model

$\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0}$

Structure of TM (IBM Model 4)

Translation Model

Lexical Model	Fertility Model
$\prod t(f_j \mid e_i)$	$\prod n(\phi_i \mid e_i)$

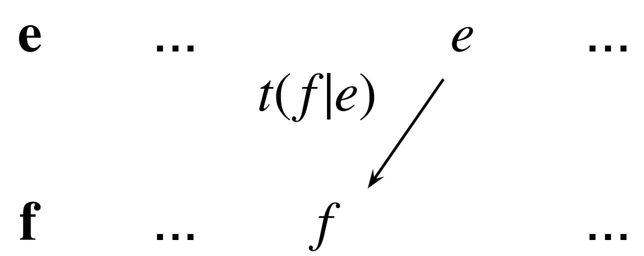
Distortion Model

Head	$\prod d_1(j - k \mid \mathcal{A}(e)\mathcal{B}(f))$
Non-Head	$\prod d_{1>}(j - j' \mid \mathcal{B}(f))$

NULL Translation Model

$$\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0}$$

Lexical Model



Structure of TM (IBM Model 4)

Translation Model

Lexical Model

$\prod t(f_j|e_i)$

Fertility Model

$\prod n(\phi_i|e_i)$

Distortion Model

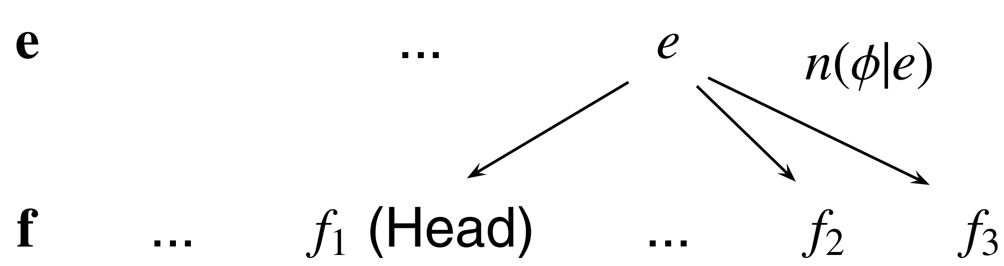
Head $\prod d_1(j-k|\mathcal{A}(e)\mathcal{B}(f))$

Non-Head $\prod d_{1>}(j-j'|\mathcal{B}(f))$

NULL Translation Model

$\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0}$

Fertility Model



Structure of TM (IBM Model 4)

Translation Model

Lexical Model

$$\prod t(f_j|e_i)$$

Fertility Model

$$\prod n(\phi_i|e_i)$$

Distortion Model

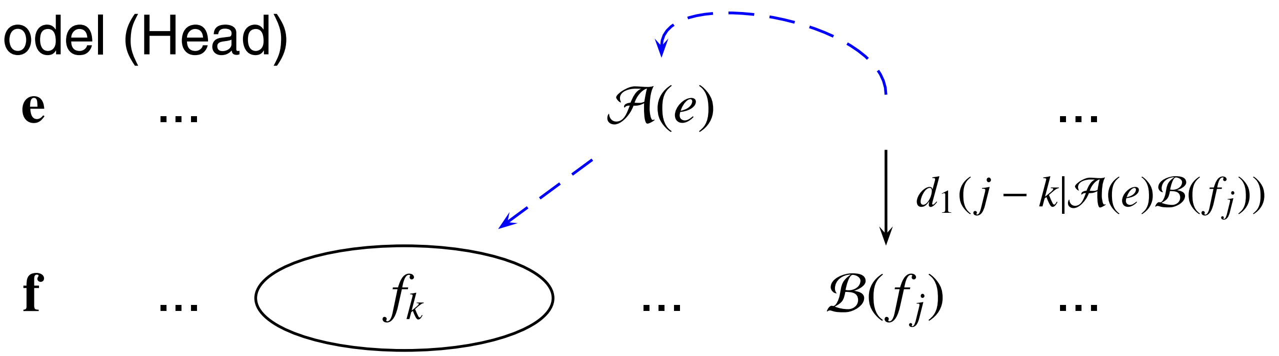
Head $\prod d_1(j-k|\mathcal{A}(e)\mathcal{B}(f))$

Non-Head $\prod d_{1>}(j-j'|\mathcal{B}(f))$

NULL Translation Model

$$\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0}$$

Distortion Model (Head)



Structure of TM (IBM Model 4)

Translation Model

Lexical Model

$\prod t(f_j|e_i)$

Fertility Model

$\prod n(\phi_i|e_i)$

Distortion Model

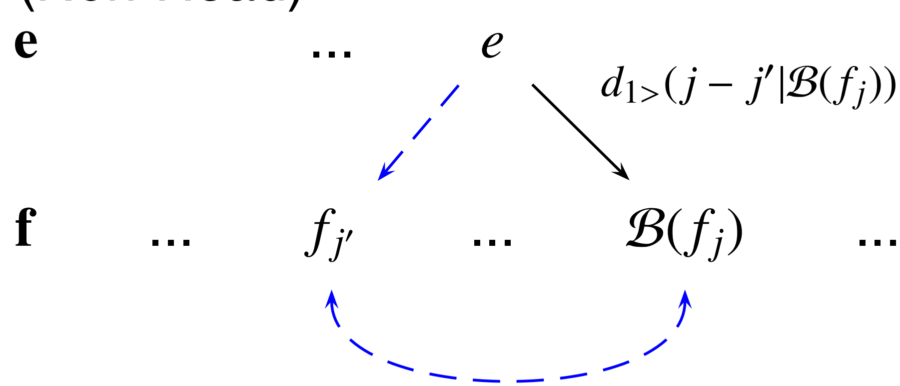
Head $\prod d_1(j - k|\mathcal{A}(e)\mathcal{B}(f))$

Non-Head $\prod d_{1>}(j - j'|\mathcal{B}(f))$

NULL Translation Model

$\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0}$

Distortion Model (Non-Head)



Structure of TM (IBM Model 4)

Translation Model

Lexical Model

$\prod t(f_j|e_i)$

Fertility Model

$\prod n(\phi_i|e_i)$

Distortion Model

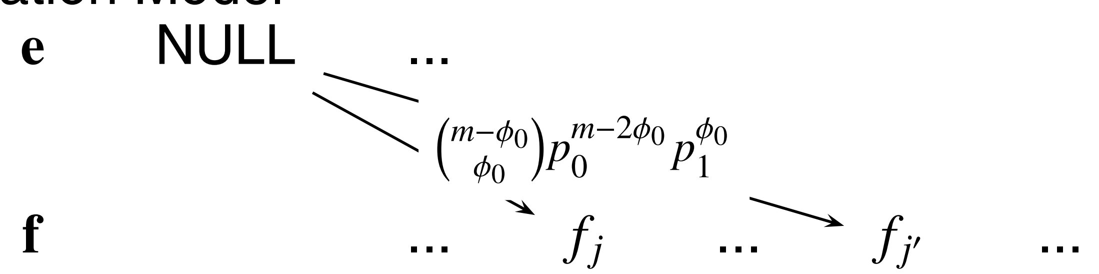
Head $\prod d_1(j-k|\mathcal{A}(e)\mathcal{B}(f))$

Non-Head $\prod d_{1>}(j-j'|\mathcal{B}(f))$

NULL Translation Model

$\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0}$

NULL Translation Model



Problems of SMT — Modeling

- Good statistical translation model?

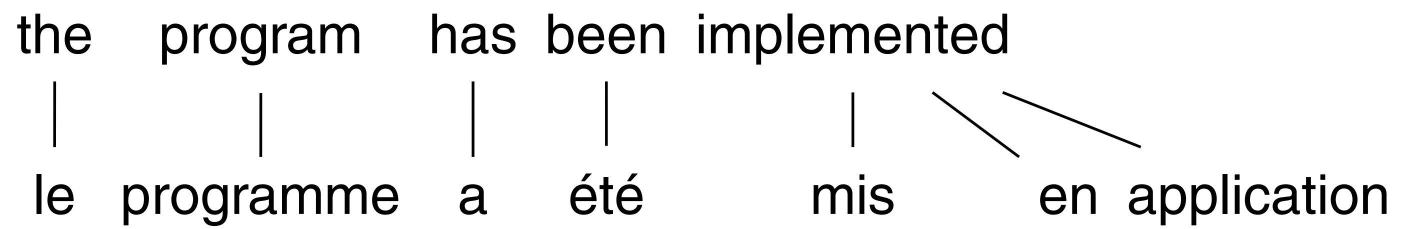
Problems of SMT — Modeling

- Good statistical translation model?
- No syntactical knowledge
- Basically, word-for-word translation considering reordering
  - Phrasal constraints implicit in IBM Model 4 and 5
  - Very good for similar language pairs
  - What about Japanese and English or others?

Problems of SMT — Modeling

- Good statistical translation model?
- No syntactical knowledge
- Basically, word-for-word translation considering reordering
  - Phrasal constraints implicit in IBM Model 4 and 5
  - Very good for similar language pairs
  - What about Japanese and English or others?

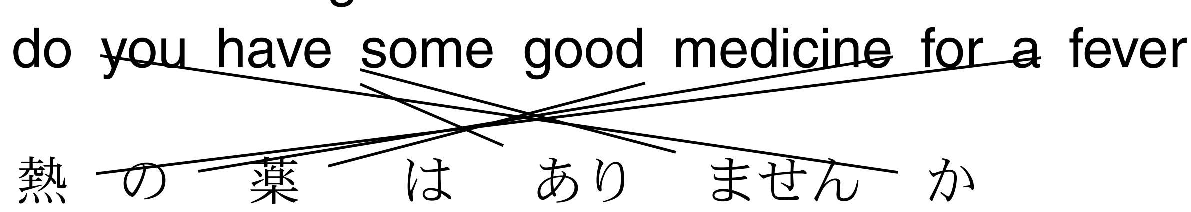
An example of viterbi alignment for F-E (from Mathematics of SMT)



Problems of SMT — Modeling

- Good statistical translation model?
- No syntactical knowledge
- Basically, word-for-word translation considering reordering
  - Phrasal constraints implicit in IBM Model 4 and 5
  - Very good for similar language pairs
  - What about Japanese and English or others?

An example of viterbi alignment for J-E



Problems of SMT — Training

- Possible to estimate good parameters?

Problems of SMT — Training

- Possible to estimate good parameters?
- EM-algorithm with bootstrapping
  - start with simpler models, such as
    - IBM Model 1 or 2 — word-for-word translation model
    - HMM Model — alignment with 1st order dependency

    to determine initial parameters
- Impossible to enumerate all the possible alignments
  (inevitable for IBM Model 3 – 5)
  Pegging
  - $\sum$ over *neighbours* of probable alignments
  - *probable alignments* derived from IBM Models 1 or 2

Problems of SMT — Search

- Given an input, can we translate it?

Problems of SMT — Search

- Given an input, can we translate it?
- input length = 10, output length = 11 and 20,000 vocabulary
  - $20,000^{11}$ possible translations
  - $(11 + 1)^{10}$ possible alignments
- NP-complete problem — Traveling Salesman Problem
  - visit all the cities (input words)
  - visit some of the hotels in a city (output words)
- (Almost) linear alignment (with local reordering) for G-E, F-E etc.
  - What about J-E? — drastical reordering

Introduction to HPA

- Align biligual text phrase-by-phrase

Introduction to HPA

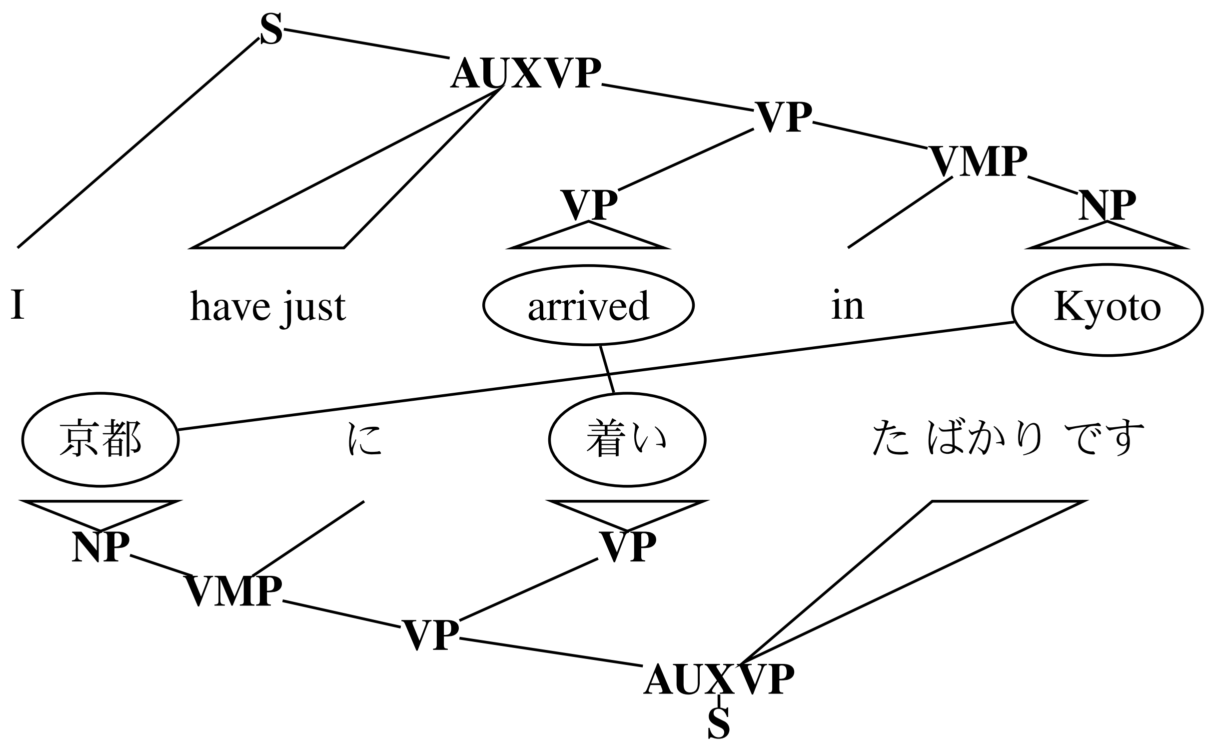
- Align biligual text phrase-by-phrase
- An example

I have just arrived in Kyoto

京都 に 着い た ばかり です

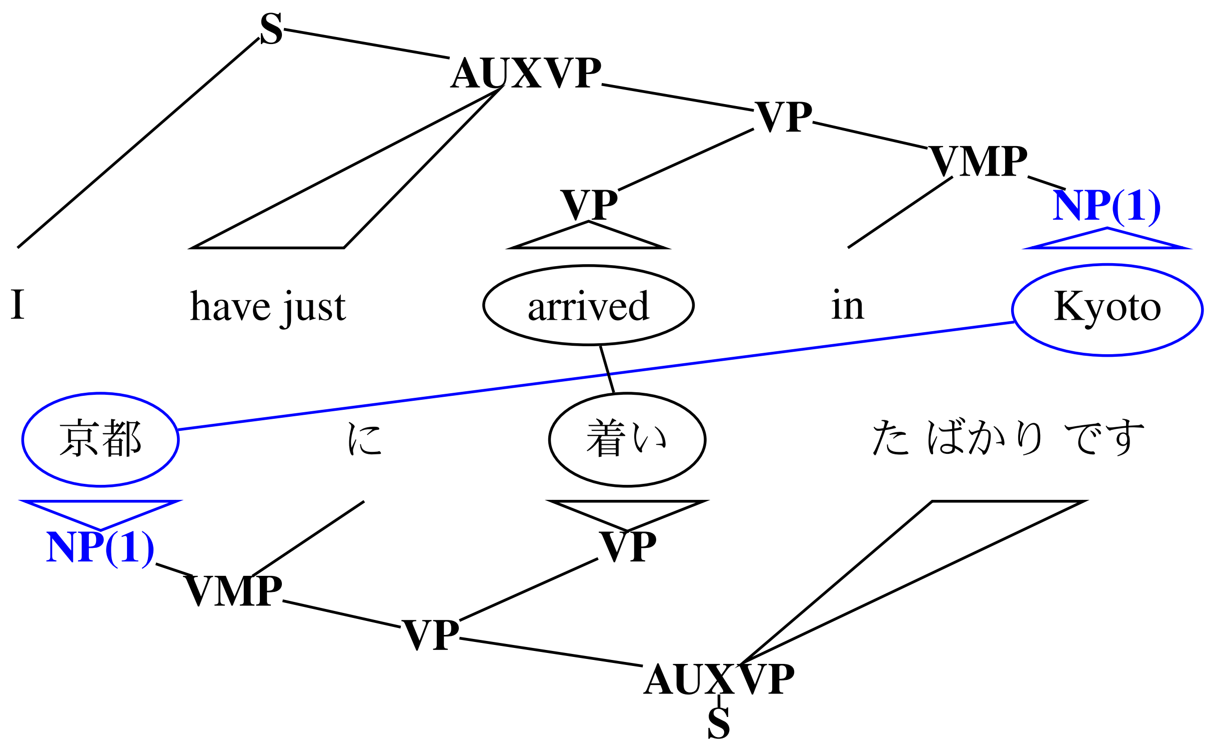
in Kyoto	—	京都 に
arrived in Kyoto	—	京都 に 着い
have just arrived in Kyoto	—	京都 に 着い た ばかり です

An Example of HPA



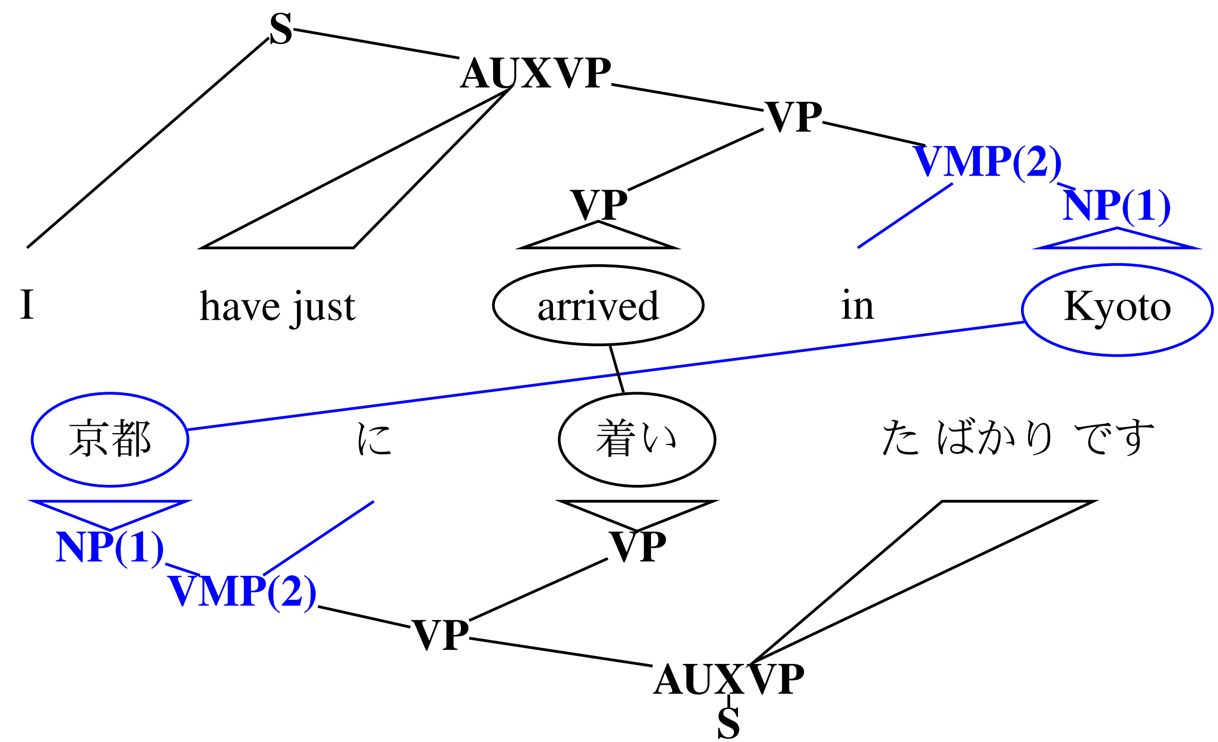
- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximumize the number of aligned phrases

An Example of HPA



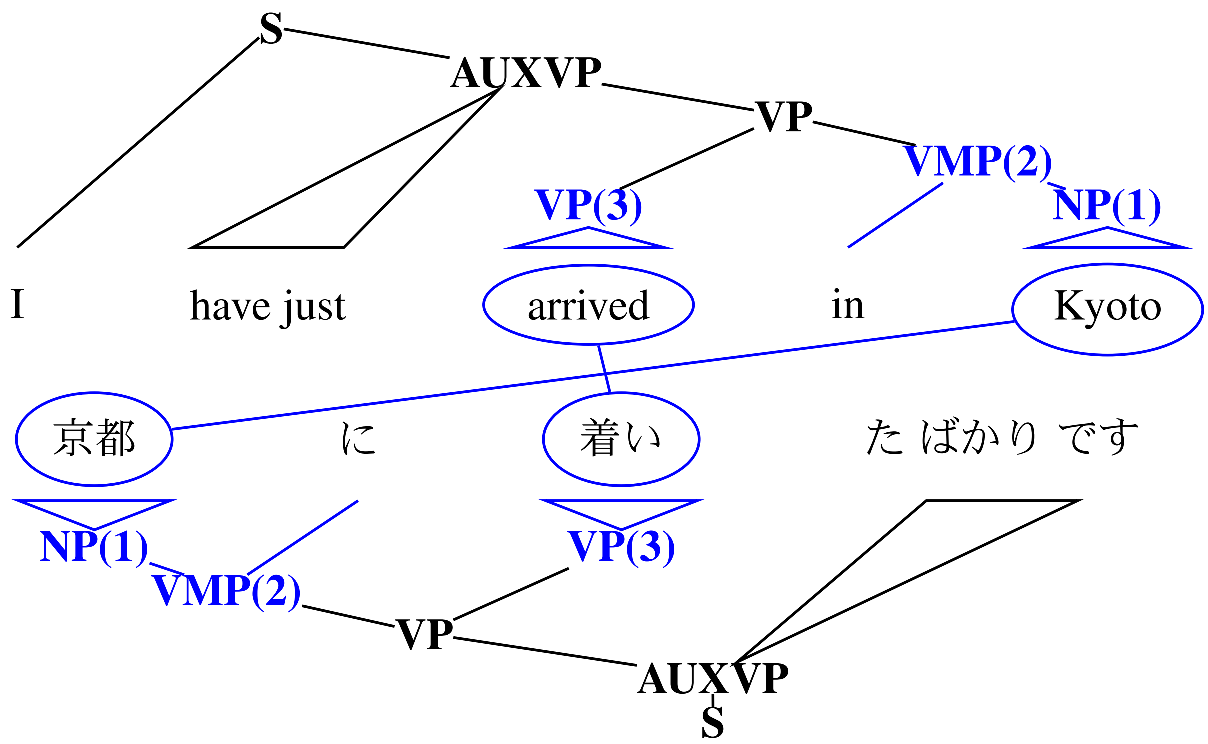
- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximumize the number of aligned phrases

An Example of HPA



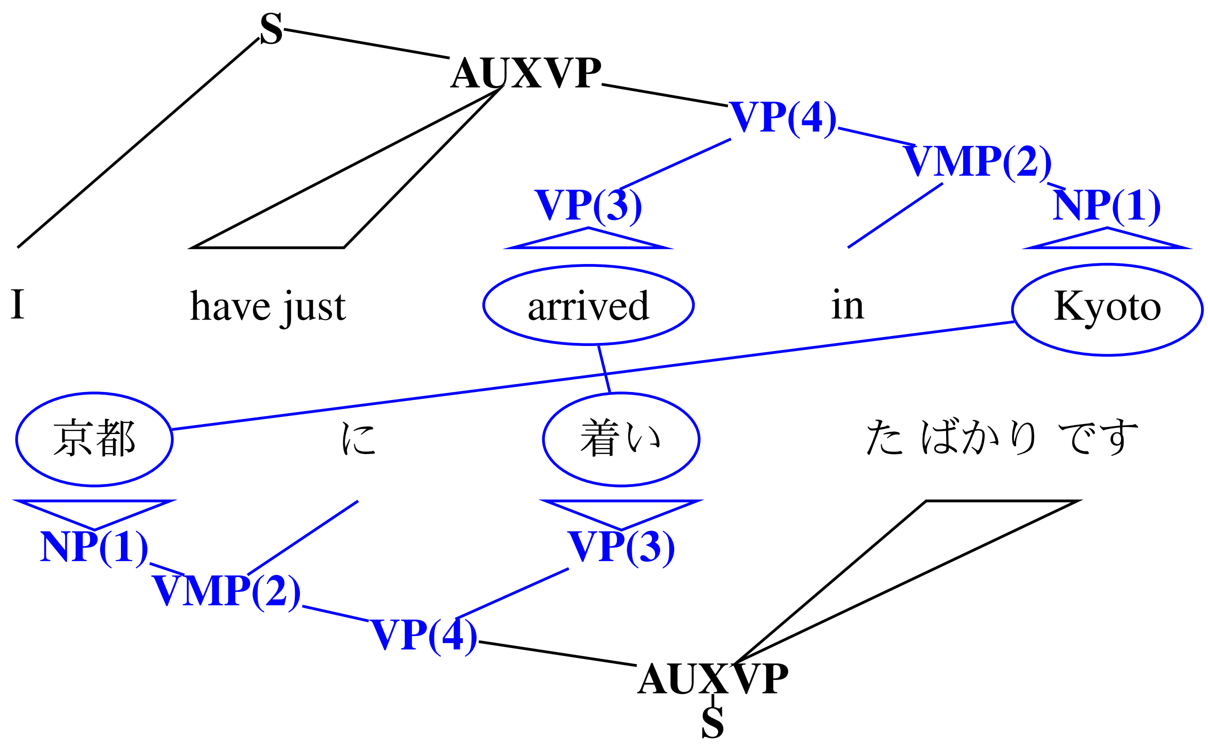
- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximumize the number of aligned phrases

An Example of HPA



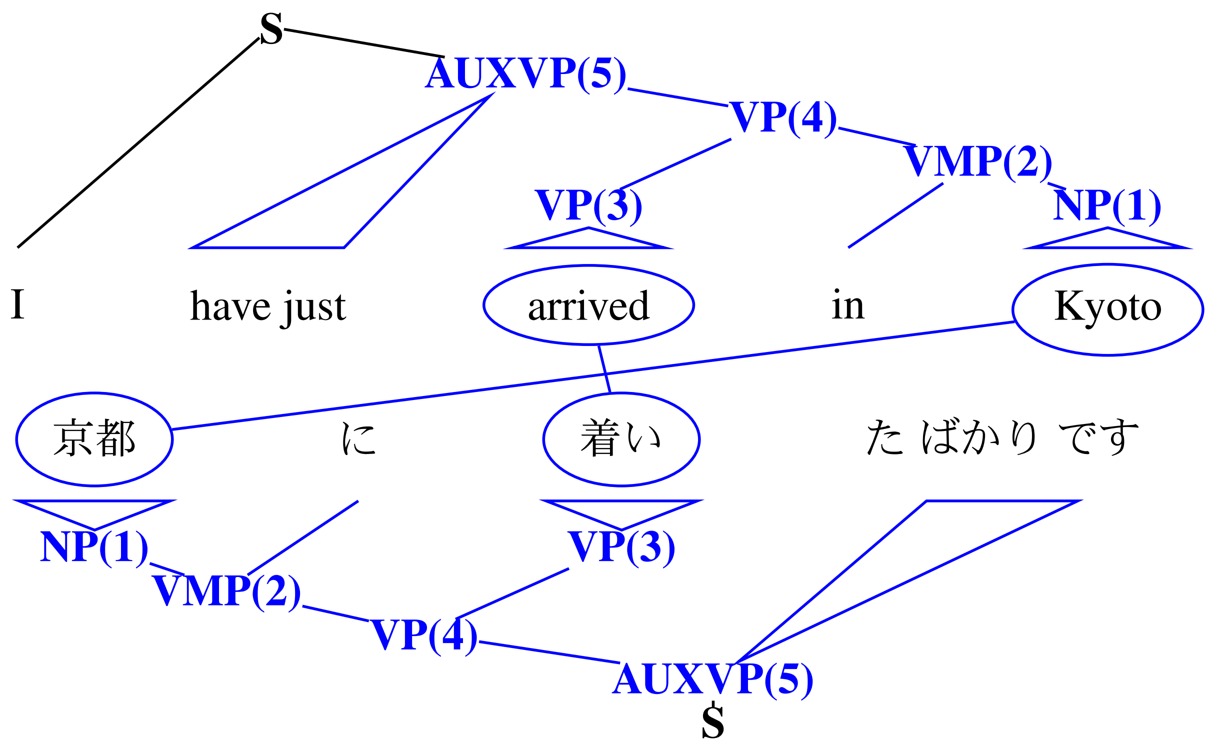
- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximumize the number of aligned phrases

An Example of HPA



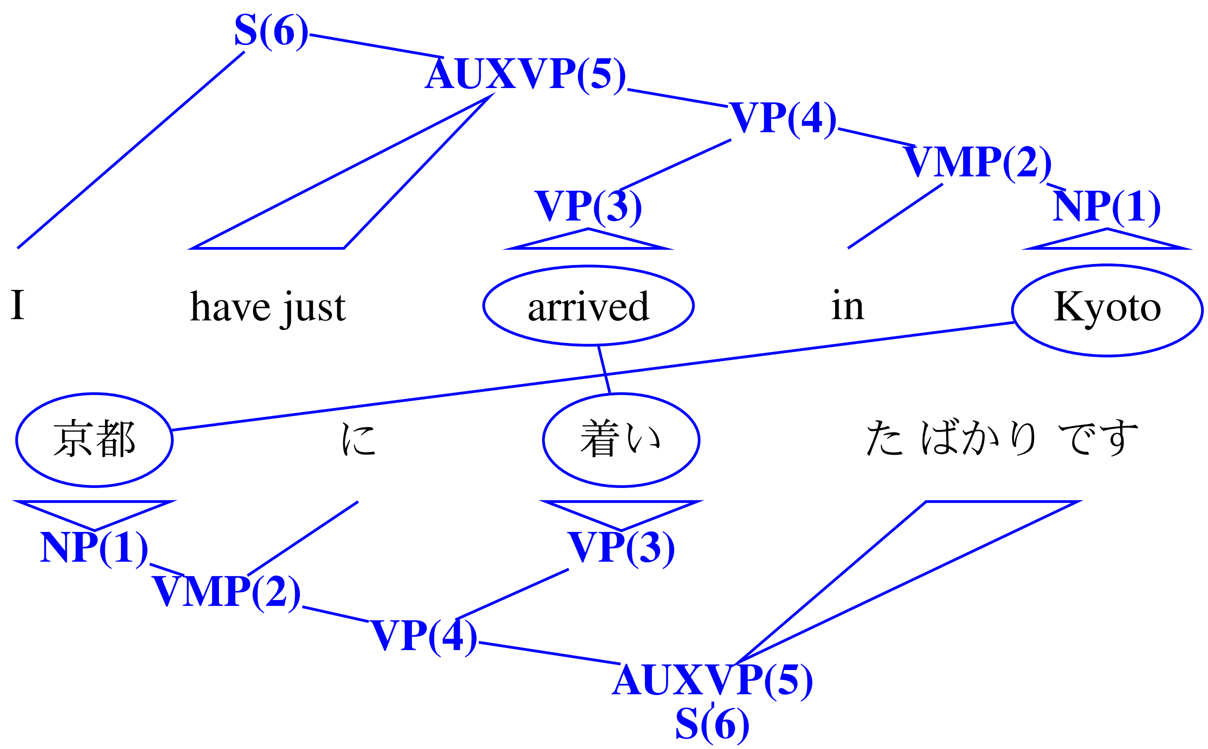
- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximumize the number of aligned phrases

An Example of HPA



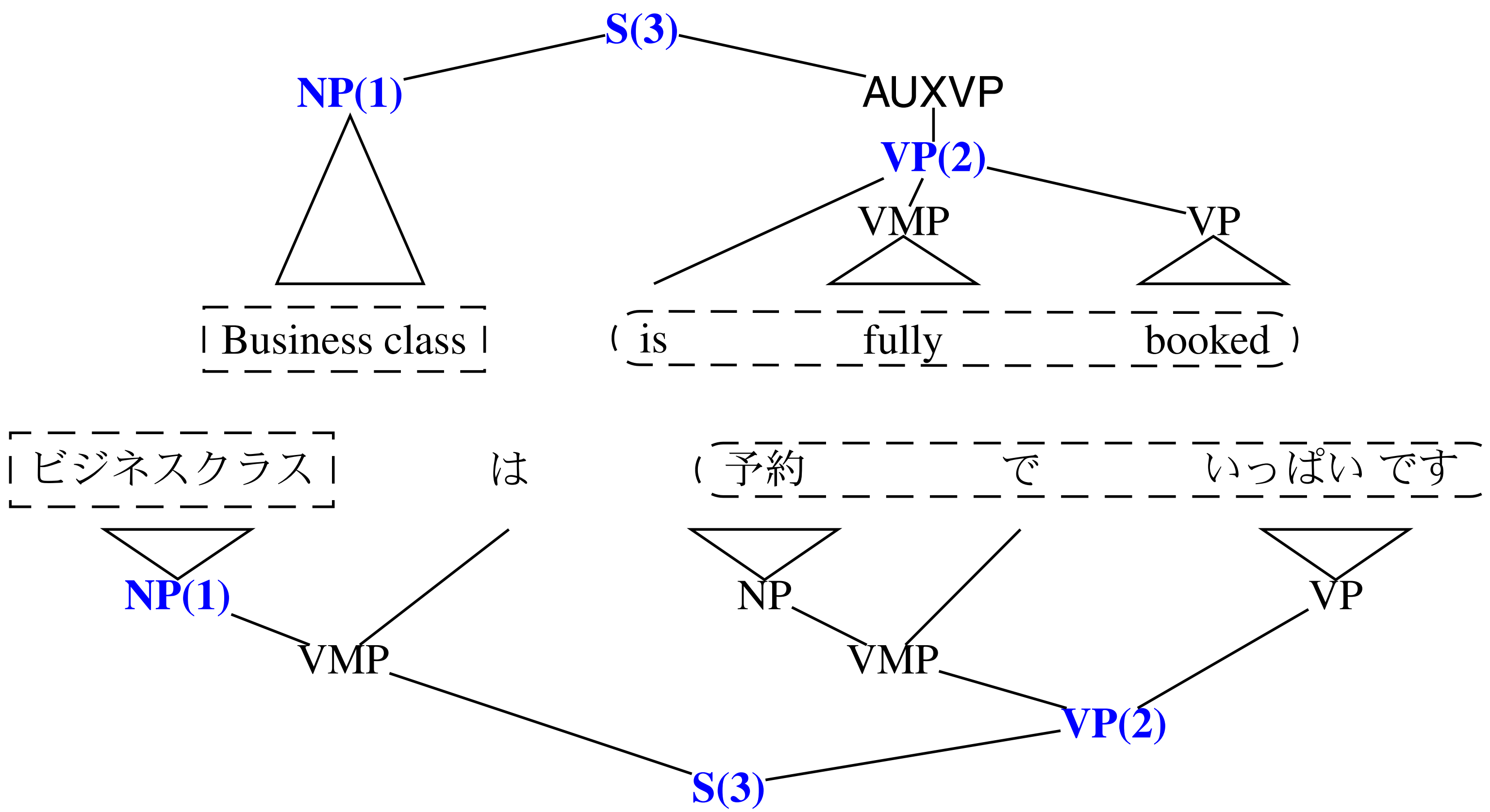
- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximumize the number of aligned phrases

An Example of HPA



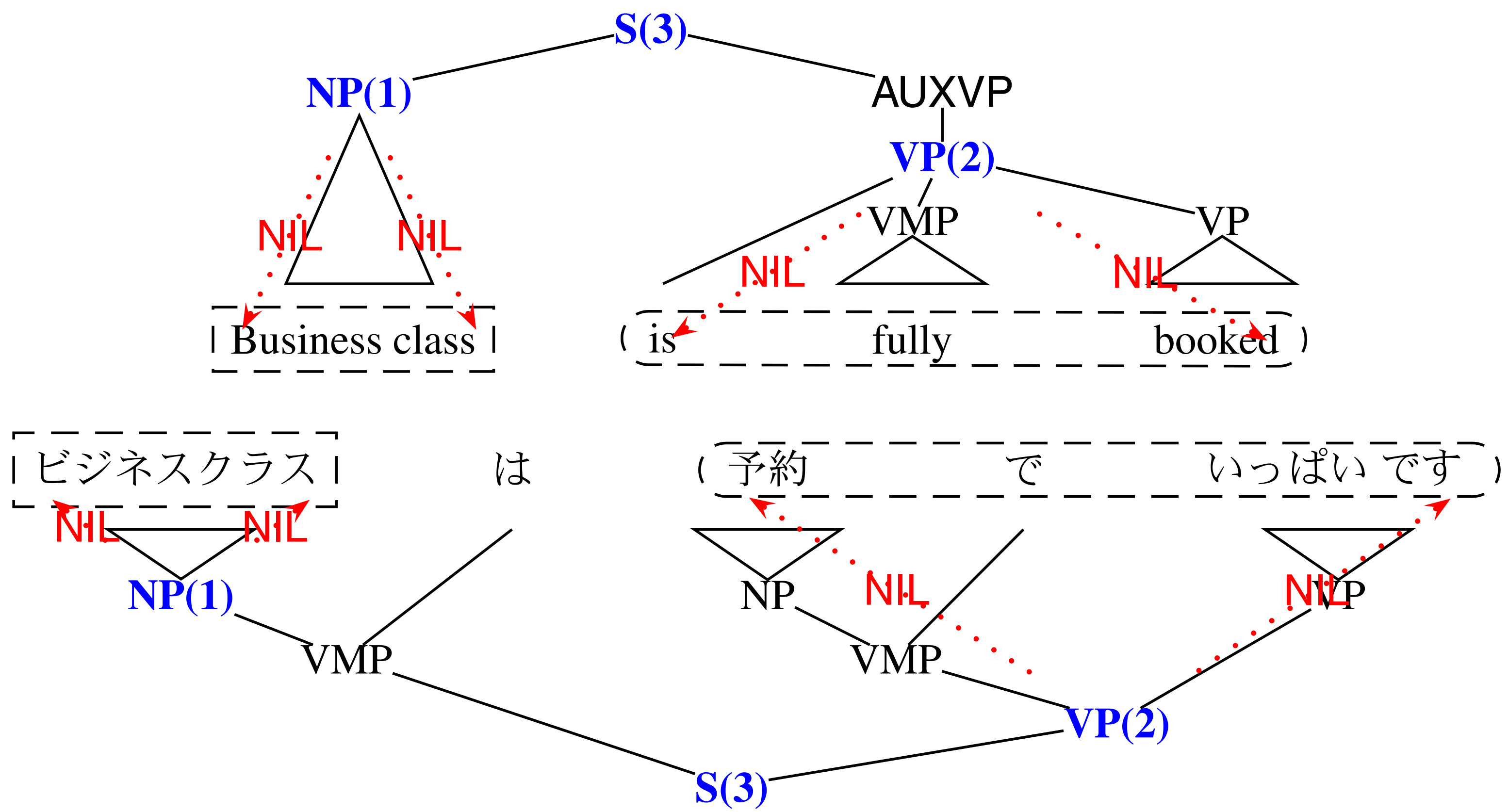
- Pairing of nodes by syntactic categories starting from word-linkage
- Phrase alignments which maximumize the number of aligned phrases

Chunking by HPA



- Chunking by extracting low-level phrases

Chunking by HPA



- Chunking by extracting low-level phrases

Chunking Model

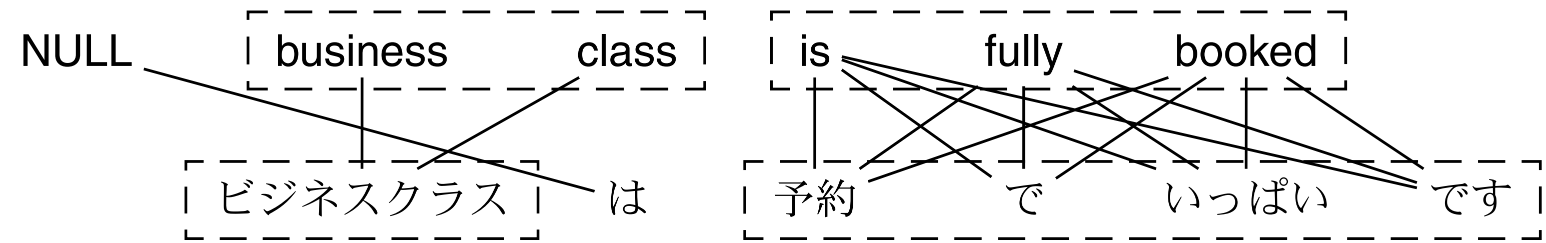
- Create a model by treating each chunk as a token

business class	$\longrightarrow$	business:class
is fully booked	$\longrightarrow$	is:fully:booked
予約 で いっぱい です	$\longrightarrow$	予約:で:いっぱい:です

- Bootstrapping from IBM Model 1 and create IBM Model 4

HPA Model

- A set of alignments hypothesized by HPA



$$\mathbf{A}_{\mathbf{HPA}} = (\{1,2\}, \quad \{0\}, \quad \{3,4,5\}, \quad \{3,4,5\}, \quad \{3,4,5\}, \quad \{3,4,5\})$$

- Directly compute IBM Model 4 parameters w/o pegging

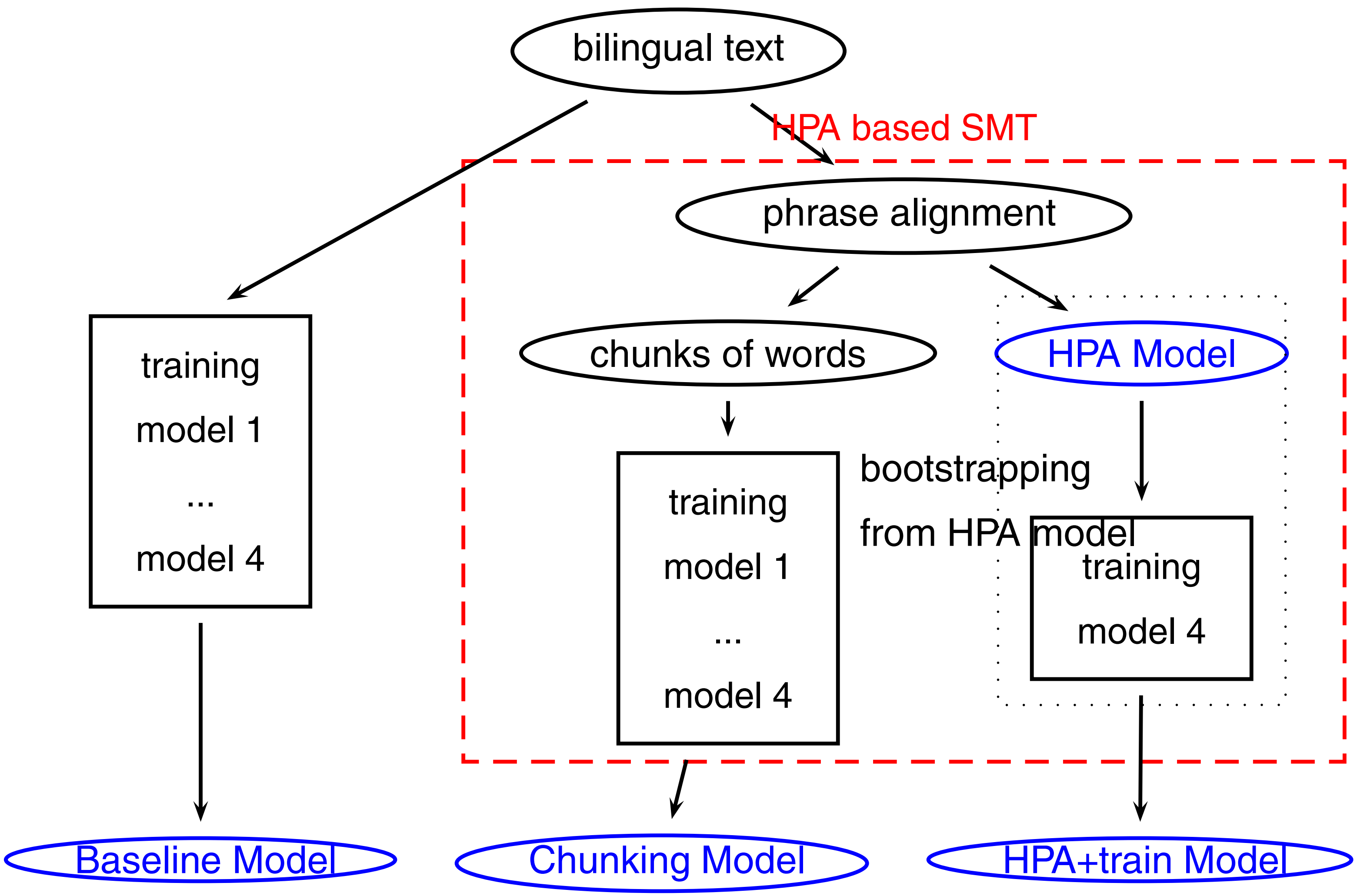
$$tc(f|e; \mathbf{f}, \mathbf{e}, \mathbf{A}_{\mathbf{HPA}}) = \sum_{\mathbf{a} \in \mathbf{A}_{\mathbf{HPA}}} P(\mathbf{a}|\mathbf{f}, \mathbf{e}) \sum_{i,j} \delta(f, f_j)\delta(e, e_{a_j})$$

$$t(f|e) \leftarrow \sum_{s \in train} tc(f|e; \mathbf{f}_s, \mathbf{e}_s, \mathbf{A}_s)$$

HPA+train Model

- Use the HPA model (= IBM Model 4) as initial parameters for further training of IBM Model 4
- Use pegged alignments

Overview of Models



Experimental Results — Settings

- Corpus

	English	Japanese
number of sentences	145,432	
number of words	835,048	896,302
vocabulary size	13,162	20,348
average sentence length	5.74	6.16
trigram perplexity	36.03	32.93

- Chunking

	English	Japanese
number of chunks	7,604	6,750
vocabulary size (of chunks)	2,166	1,624
average number of chunks per sentence	0.759	0.673
average number of words per chunk	2.21	2.52
trigram perplexity	72.36	72.07

J-E Translation Results (1)

- Tested on 150 inputs

Model	WER	PER	SE			
			A	B	C	D
baseline	70.2	59.2	12.7	33.3	14.7	38.7
chunking	64.0	53.1	21.3	28.0	16.7	34.0
HPA	64.5	58.1	17.3	32.0	15.3	35.3
HPA+train	71.0	59.3	16.0	32.0	22.0	30.0

WER: word error rate

PER: position independent word error rate

SE: subjective evaluation (A: perfect, B: fair, C: acceptable, D: nonsense)

J-E Translation Results (2)

Model	WER			PER			SE(A+B+C)		
length	6	8	10	6	8	10	6	8	10
baseline	66.6	67.5	76.6	56.8	60.7	60.0	66.0	64.0	52.0
chunking	54.5	57.0	80.6	48.4	48.9	62.0	78.0	72.0	48.0
HPA	59.5	65.7	68.4	55.3	60.7	58.4	72.0	66.0	56.0
HPA+train	64.3	72.6	76.2	55.8	62.5	59.7	78.0	72.0	60.0

ATR

Sample Translations

ステーキ の:焼き 具合 は どう さ れ ます か

baseline:	(D)	can you steak
chunking:	(A)	how do you like your:steak
HPA+train:	(A)	how do you like your steak

ゴルフ場 の:予約 でき:ます:か

baseline:	(C)	can i make-a-reservation
chunking:	(A)	can:i make-a-reservation:for golf
HPA+train:	(A)	could you make-a-reservation for the golf course

シカゴ から シアトル まで どのくらい:時間:が かかり:ます:か

baseline:	(A)	how-long does it take to seattle from chicago
chunking:	(A)	how-long will:it:take to seattle from chicago
HPA+train:	(B)	do you how-long will it take to seattle from chicago

Sample Translations (Contd.)

席 の 確保 では くれぐれも 最高 の 所 を お願いします

(please be sure to secure the best available seats for us)

baseline:	(B)	i would like a seat in a great place please
chunking:	(D)	what 's the maximum area for sends providing seats
HPA+train:	(D)	my best regards to your seat find a place please

初心者 な:の:だ けど 参加し ても:いい:です:か

(i am a beginner may i join)

baseline:	(D)	do you have may but take beginner
chunking:	(D)	can:i join beginners ring
HPA+train:	(D)	it is but i am a beginner