Word Segmentation and Transliteration in Chinese and Japanese

Masato Hagiwara

Rakuten Institute of Technology, New York

CUNY NLP Seminar 4/5/2013

楽天 Rakuten

Who am I?

HAGIWARA, Masato (萩原　正人)

Senior Scientist at Rakuten Institute of Technology, New York

- Ph.D. from Nagoya University (2009)
- Internship at Google and Microsoft Research (2005, 2008)
- R&D Engineer at Baidu, Japan (2009-2010)

Baidu, Inc. Beijing, China

Rakuten Institute of Technology, New York

楽天 Rakuten

Agenda

Word Segmentation

Transliteration

Integrated Models

Word Segmentation in Chinese and Japanese

Maximum Forward Match

Greedily match longest lexicon items from the beginning (or from the end)

日 文 章 鱼 怎 么 说

How do you say octopus in Japanese?

日 文	章 鱼	怎 么	说
Japanese	octopus	how	say

lexicon

日	(day)
日文	(Japanese)
文章	(article)
章鱼	(octopus)
鱼	(fish)
怎么	(how)
说	(say)

楽R天 RRakuten [Wong and Chan 1996]

Examples Where Maximum Match Fails

警 察 枪 杀 了 那 个 逃 犯

警 察	枪 杀	了	那	个	逃 犯
Police	gun-kill	(*perf*.)	that	(*mes*.)	escapee

警 察 用 枪 杀 了 那 个 逃 犯

Police with gun kill (*perf.*) that escapee

警 察	用	枪 杀
Police	with	gun-kill

？？？

- Heuristic rules
- “Word Binding” Scores

楽天 Rakuten

Heuristic Approaches – Minimum Bunsetsu Number

今 日 本 当 に 良 い 天 気 で す ね

今 日	本 当 に	良 い	天 気	で す	ね
Today	really	good	weather	is	(*part.*)

# of Bunsetsu = 4

今	日 本	当 に	良 い	天 気	で す	ね
Now	Japan	really	good	weather	is	(*part.*)

# of Bunsetsu = 5

今　　n. (now)
今日　　n (today)
...

lexicon

Optimizes for a whole sentence

楽R天　RRakuten　[Yoshimura et al. 1983]

What is Bunsetsu?

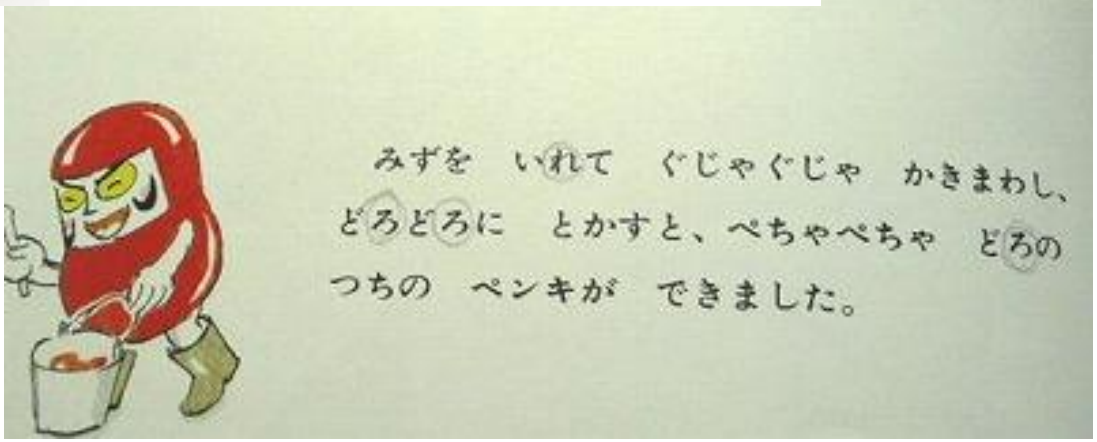
'용베어'는, 병실의 도어를 살그머니 열었습니다.
작은 아기가,
신체중에 관을 가득 붙여 자고 있습니다.

침대의 옆에는,
걱정일 것 같은 얼굴을 하고 있는 아버지와 엄마.
'마아훈'이 눈을 뜨는 것을
지금인가 지금일까하고 기다리고 있습니다.

수술은 끝났지만,
의사는 어려운 얼굴을 하고 있었습니다.

앞으로도 쭉,
'마아훈'은 병과 함께
살아가지 않으면 안 됩니다.

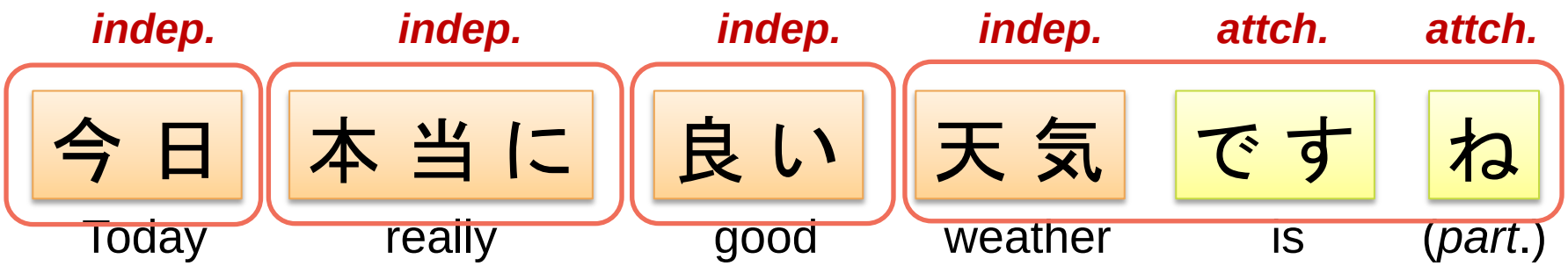
6



楽天 Rakuten

Minimum Bunsetsu Number

Bunsetsu (文節) = [indep. word] [attch. word]*



# of Bunsetsu = 4

$$\min \sum_{w} cost(w)$$

where

$$cost(w) = \begin{cases} 1 & w \text{ is an indep. word} \\ 0 & w \text{ is an attch. word} \end{cases}$$

A Special Case of Minimum Cost Methods

楽天 Rakuten [Yoshimura et al. 1983]

Word-based Models

lexicon

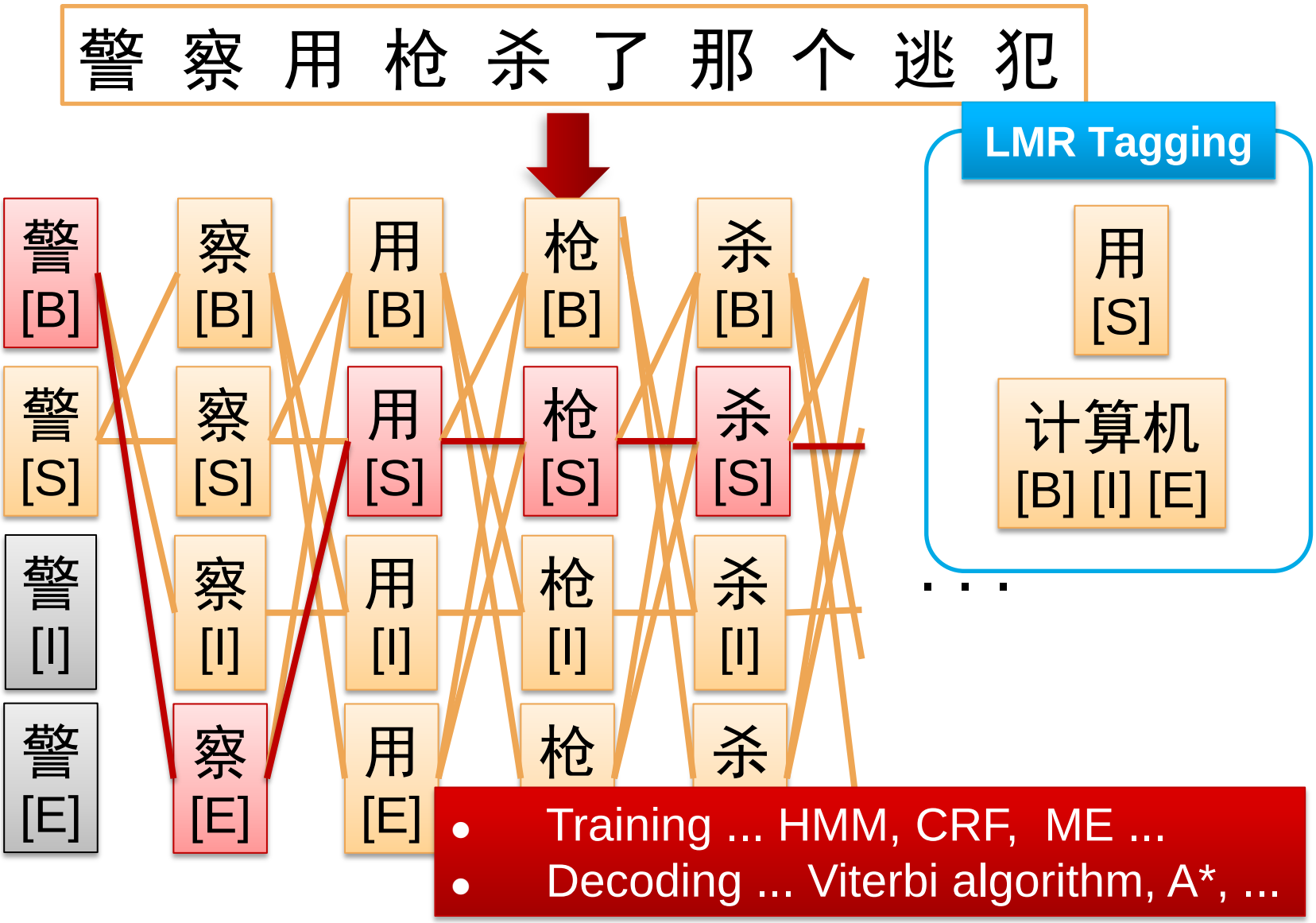
東 (east) pre.
東京 (Tokyo) n.
京 (capital) n.
京都 (Kyoto) n.
都 (Pref.) suf.
に (in) p.
住む (live) v.

$$\min \sum_{i=1}^{N} [cost_1(w_i) + cost_2(w_{i-1}, w_i)]$$

- Training ... HMM, Perceptron, CRF, ...
- Decoding ... Viterbi algorithm, A*, ...

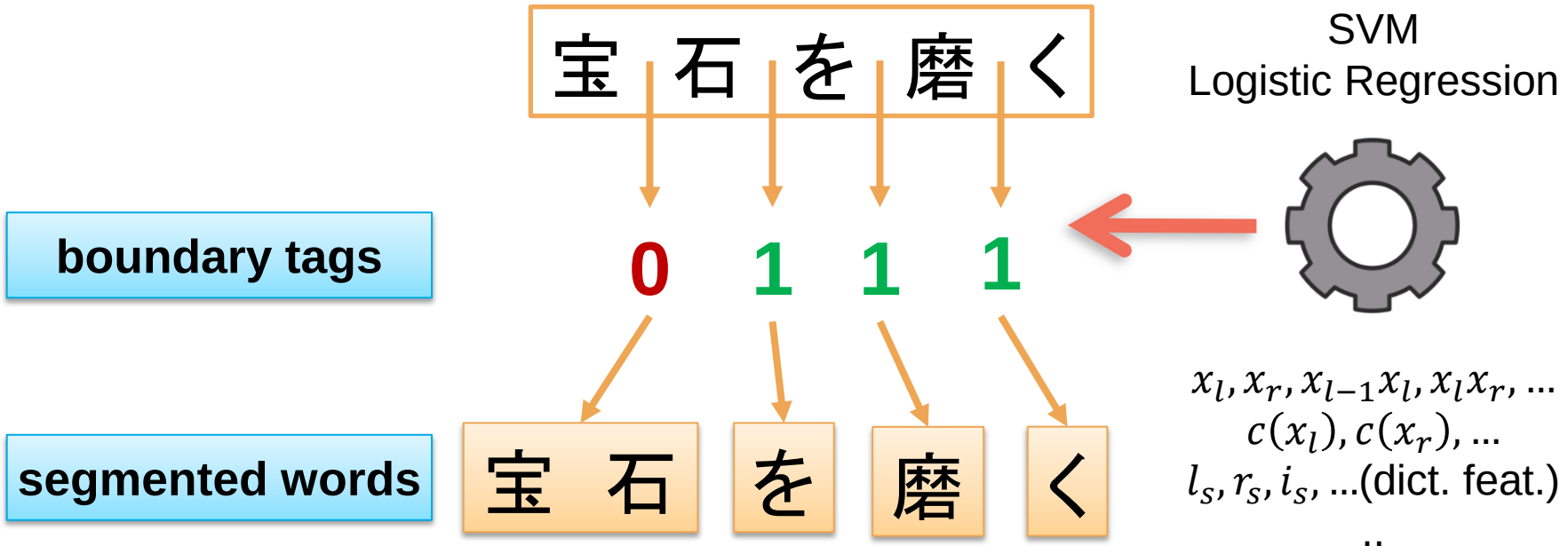
楽天 Rakuten [Kudo et al. 2004]

Character-based Models 1 – Character Tagging



[Xue and Shen 2003] [Peng et al. 2004]

Character-based Models 2 – “Boundary” Tagging

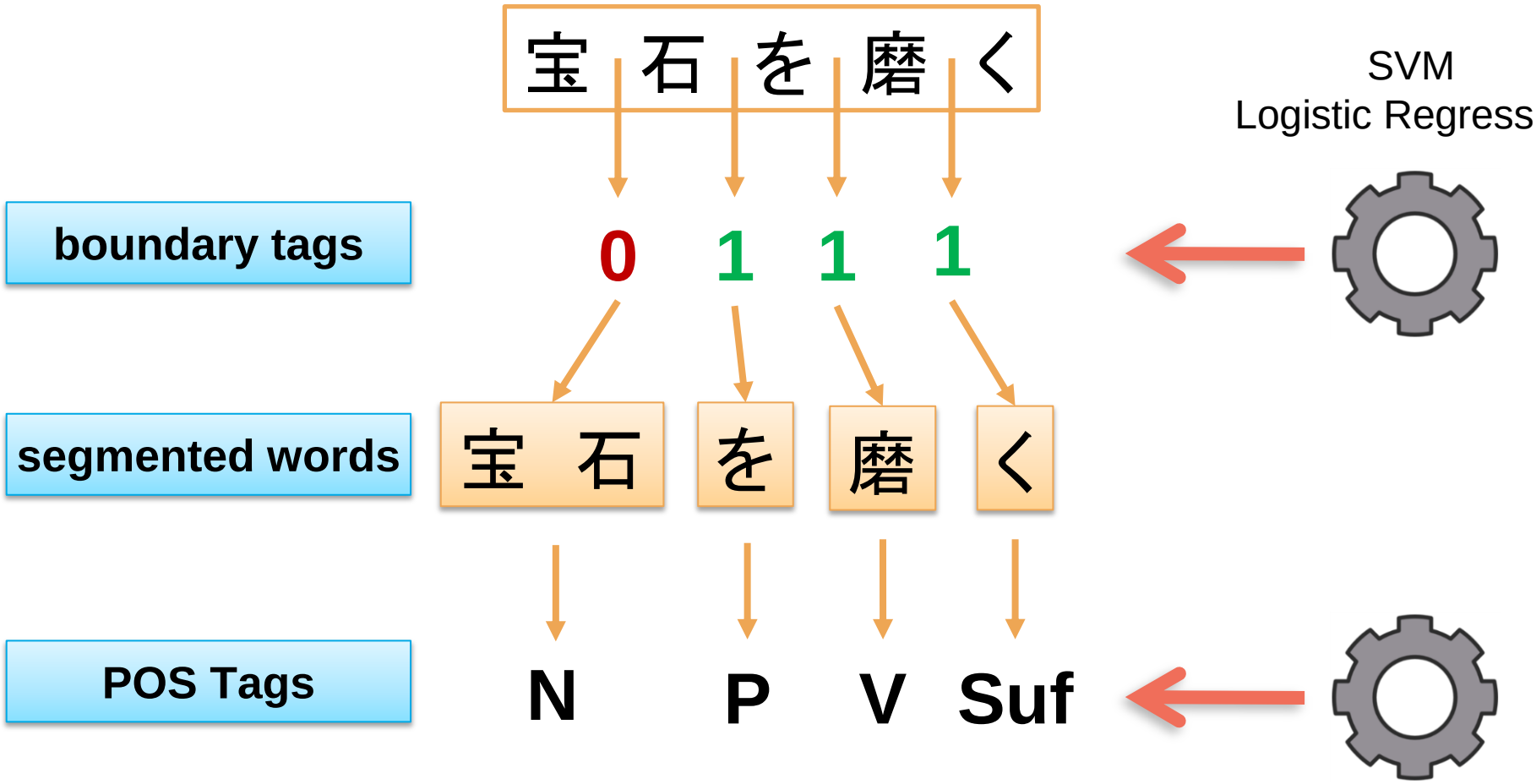


Boundary Decisions ... Independent from each other

“Gains provided by structured prediction can be largely recovered by using a richer feature set.” [Liang et al. 2008]

楽天 Rakuten [Neubig et al. 2011]

One-at-a-Time PoS Tagging Models



Enables Domain Adaptation through Partial Annotation

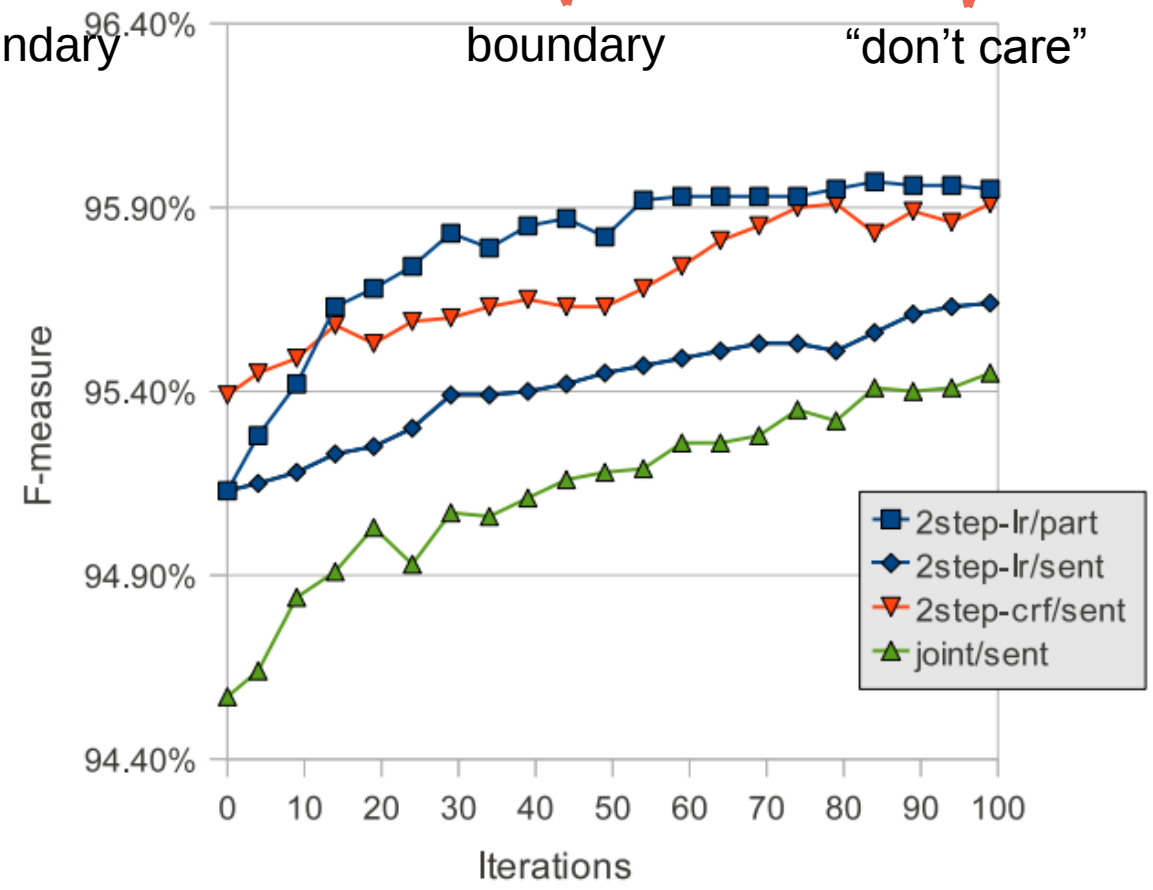
楽天 Rakuten [Neubig et al. 2011]

Pointwise Approaches and Active Learning

partial annotation

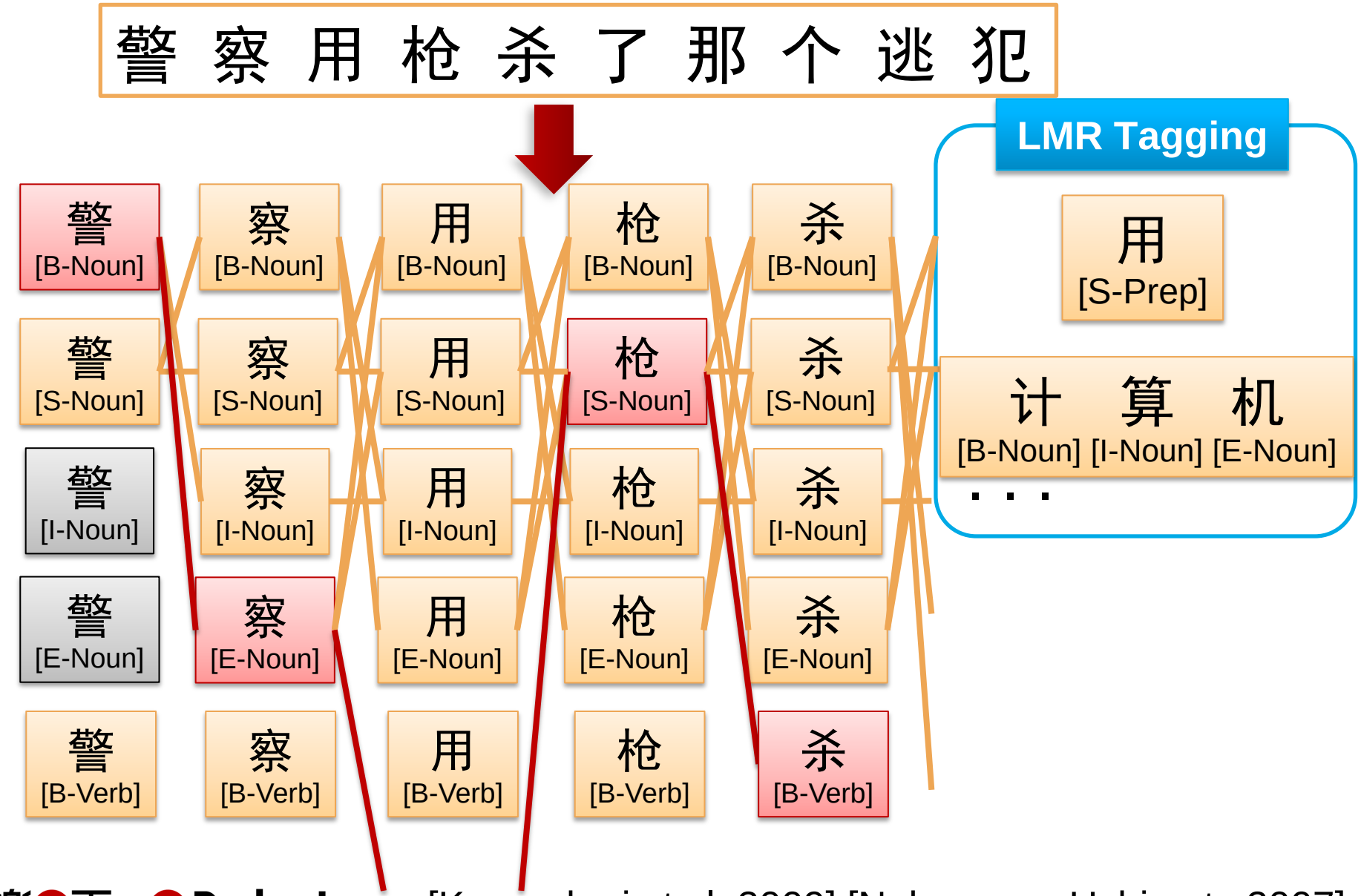
| ア-ク-チ-ン|フ-ィ-ラ-メ-ン-ト|は|細 胞 内 小 器 官|の|1|つ|だ|

non-boundary　　boundary　　“don’t care”



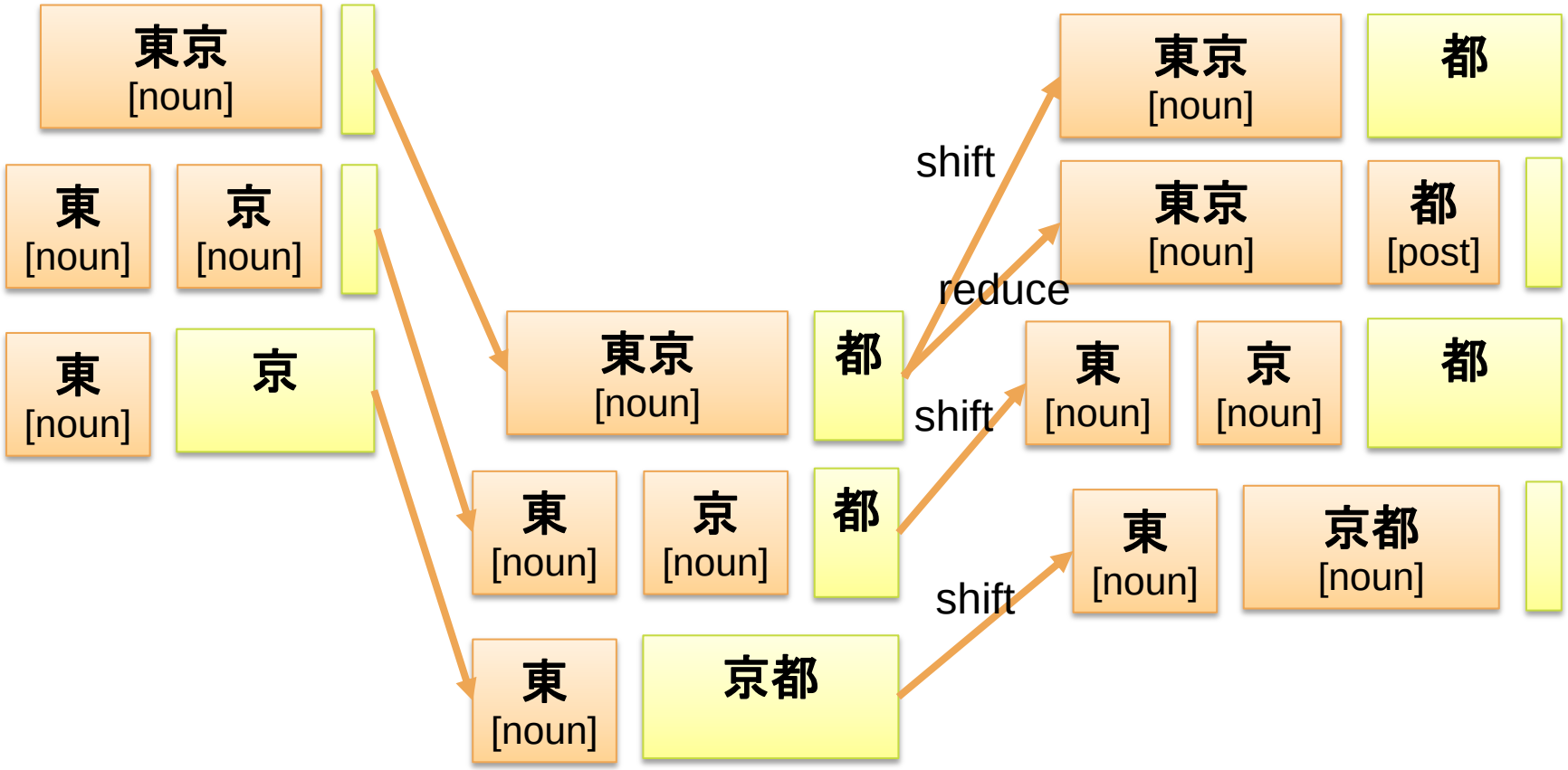
楽天 Rakuten

Character-based Joint Models



楽天 Rakuten [Kruengkrai et al. 2009] [Nakagawa+Uchimoto 2007]

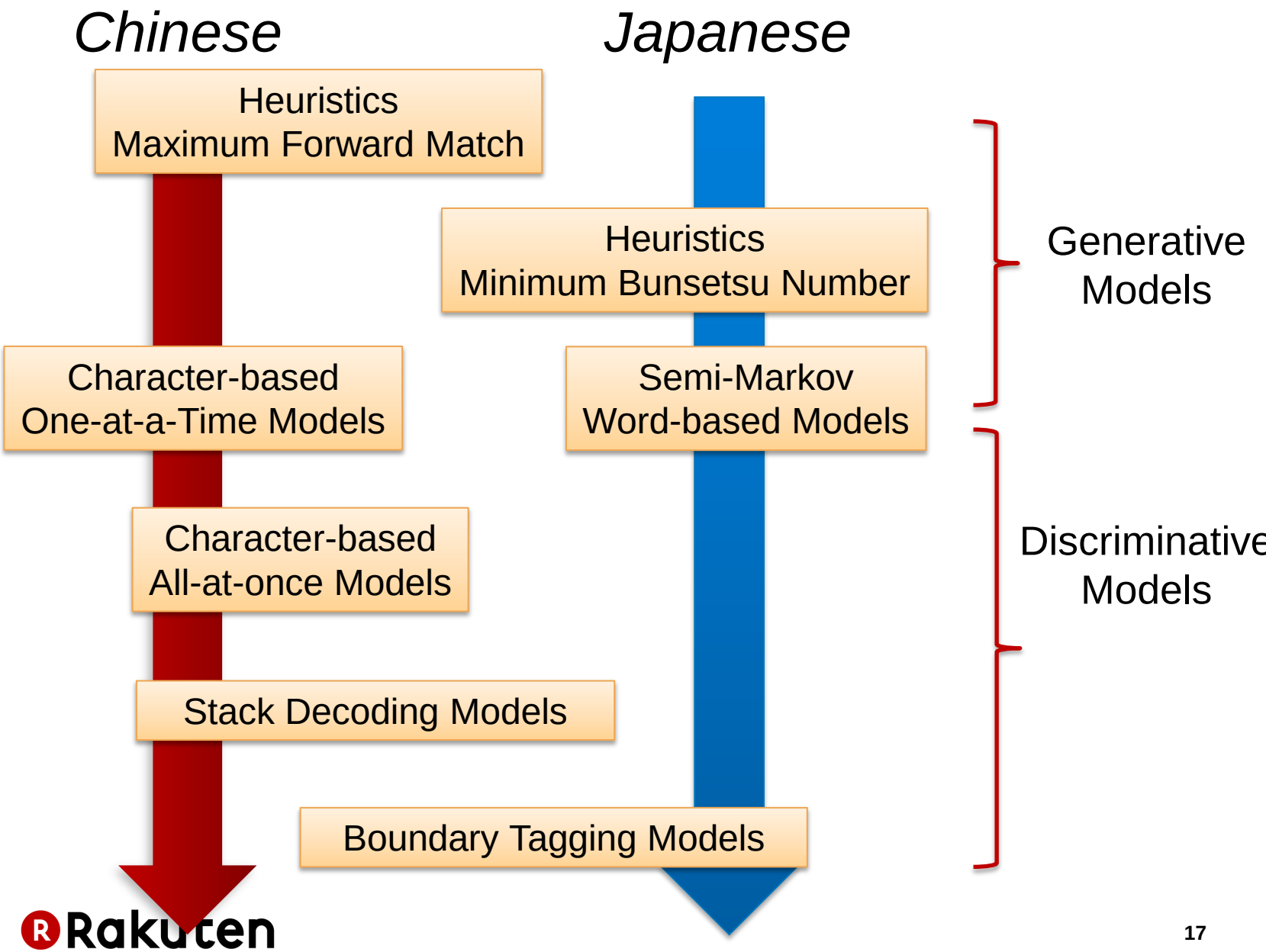
Stack Decoding Models



- No distinction between known and unknown words
- Flexible sets of features (e.g., long distance constraints)

楽天 Rakuten [Zhang, Clark 2008] [Okanohara, Tsujii 2008]

Chinese/Japanese WS Evolution



Chinese/Japanese WS Evolution

Heuristics	*Statistics*
	→ ***Statistics***
Word-based	*Character-based*
	→ ***Ja: word, Zh: character***
Pipeline	*Joint*
(One-at-a-time)	*(All-at-Once)*
	→ ***Joint***
Generative	*Discriminative*
	→ ***Discriminative***
Viterbi	*Stack Decoding*
	→ ***Pros and Cons***

楽天 Rakuten

Transliteration
(“Semantic” Transliteration Models)

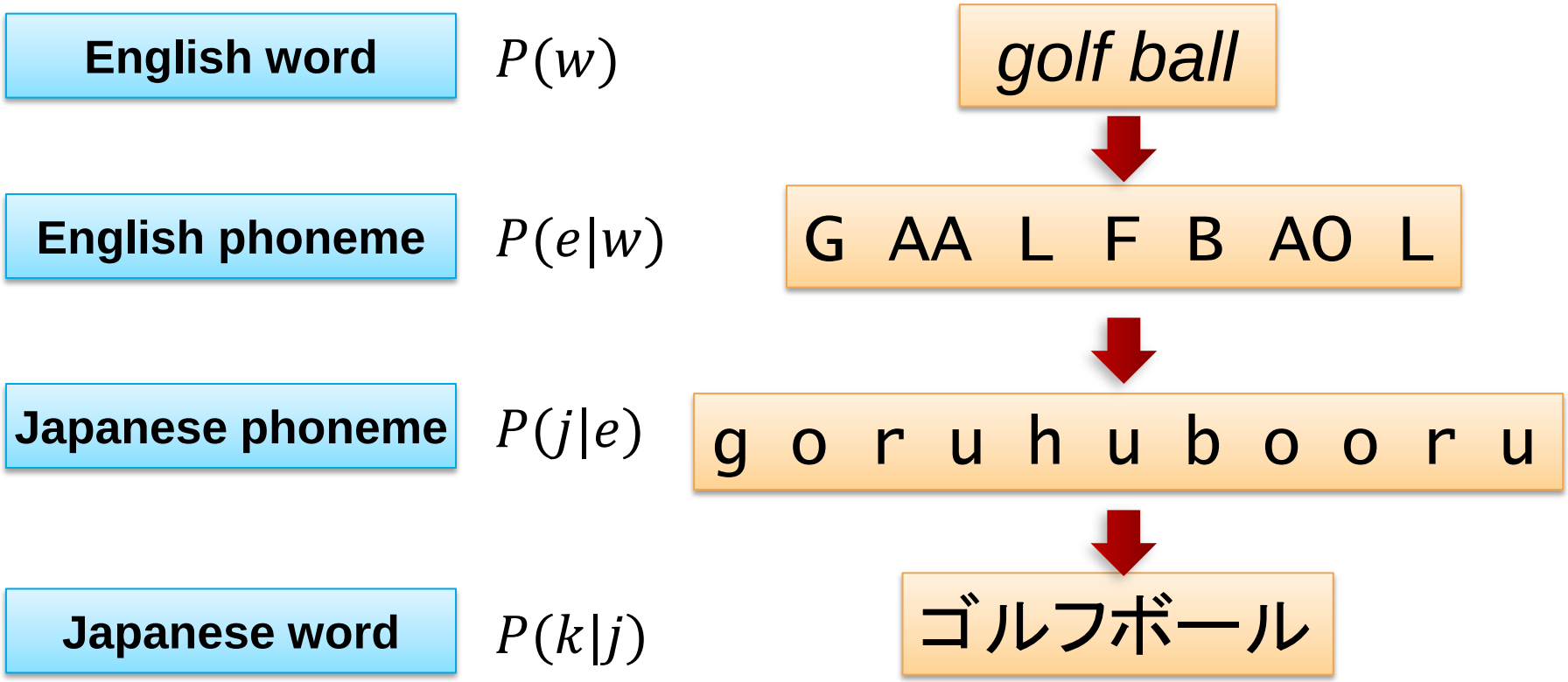
Transliteration

Phonetic translation between languages with different writing systems

New York / 纽约 *niuyue* / ニューヨーク *nyuuyooku*

Obama / 奥巴马 *aobama* / オバマ *obama*

Phoneme-based Methods



Trains a large WFST (from Japanese to English words)

$$P(w)P(e|w)P(j|e)P(k|j)$$

楽天 Rakuten [Knight and Graehl 1998]

Direct Orthographical Mapping

Joint Source Channel Model

Transliteration Prob. = Prod. of TU n-gram probs.

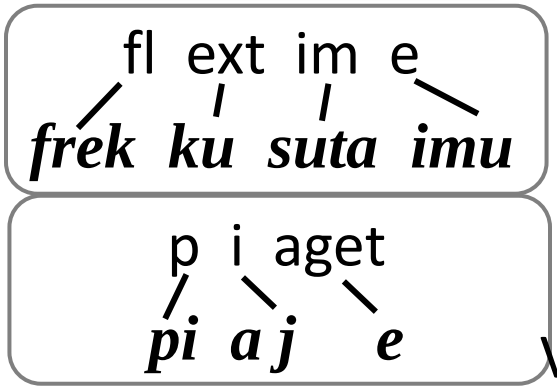
$$P_{JSC}(\langle s,t\rangle) = \prod_{i=1}^{f} P(u_i|u_{i-n+1},\ldots,u_{i-1})$$

P(flextime→***furekkusutaime***)

= P(f→***fu***|BOW) × P(le→***re***|f→***fu***) × P(x→***kkusu***|le→***re***) × …

TU Probability Estimation

Current Alignment



Freq.→Prob.

EM Algorithm

Viterbi Algorithm

TU Probability Table

P(fl→***flek*** | ▪) = XXX

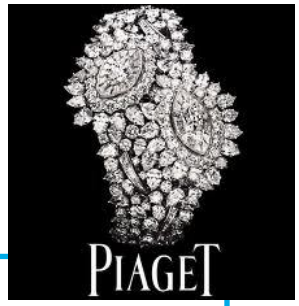
P(ext→***ku*** | ▪) = YYY

P(p→***pi*** | ▪) = ZZZ

…

楽天 Rakuten [Li et al. 2004]

Multiple Language Origins



piaget / *piaje* ピアジェ
target / *taagetto* ターゲット

→ French origin / English origin → French model / English model

亚历山大 Yalishanda / Alexander
山本 Yamamoto / Yamamoto

→ Indo-European origin / Japanese origin → Chinese Transliteration Model / Japanese Reading Model

Marian / Malian 玛丽安
Marino / Malinuo 马里诺

→ Female name / Male name → Female model / Male model

楽天 Rakuten [Li et al. 2007]

Latent Class Transliteration

Class transliteration [Li et al. 2007]

$$P_{LI}(\boldsymbol{t}|\boldsymbol{s}) = \sum_{c} P(\boldsymbol{t}, c|\boldsymbol{s}) = \sum_{c} P(c|\boldsymbol{s}) P(\boldsymbol{t}|c, \boldsymbol{s})$$

Explicit language detection

s: source

t: target

Latent Class Transliteration [Hagiwara&Sekine 2011]

$$P_{\mathrm{LST}}(\langle \boldsymbol{s}, \boldsymbol{t} \rangle) = \sum_{z=1}^{K} P(z) \prod_{i=1}^{f} P(u_i|z)$$

Latent class distribution

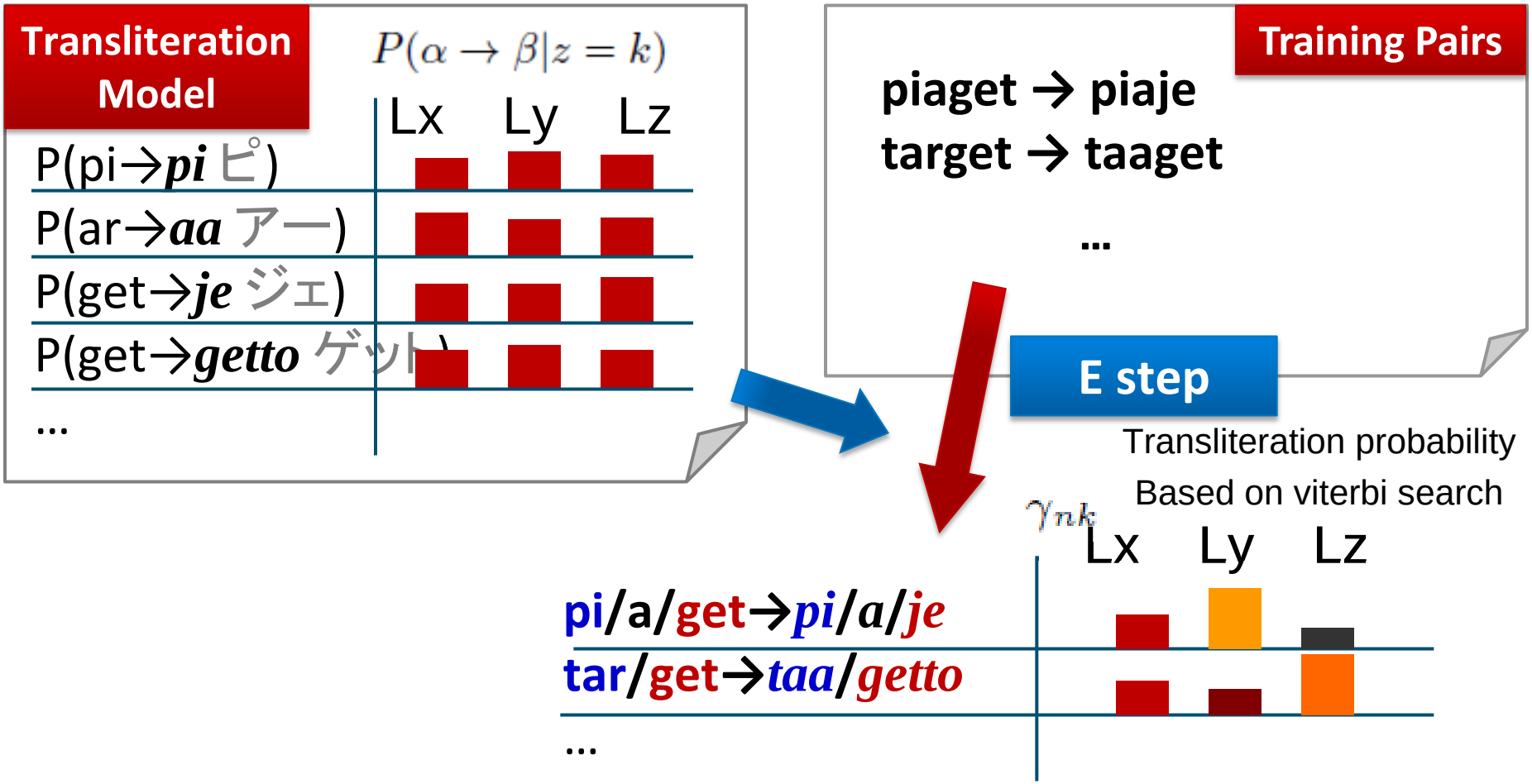
z: latent class

K: # of latent classes

(determined using dev. sets)

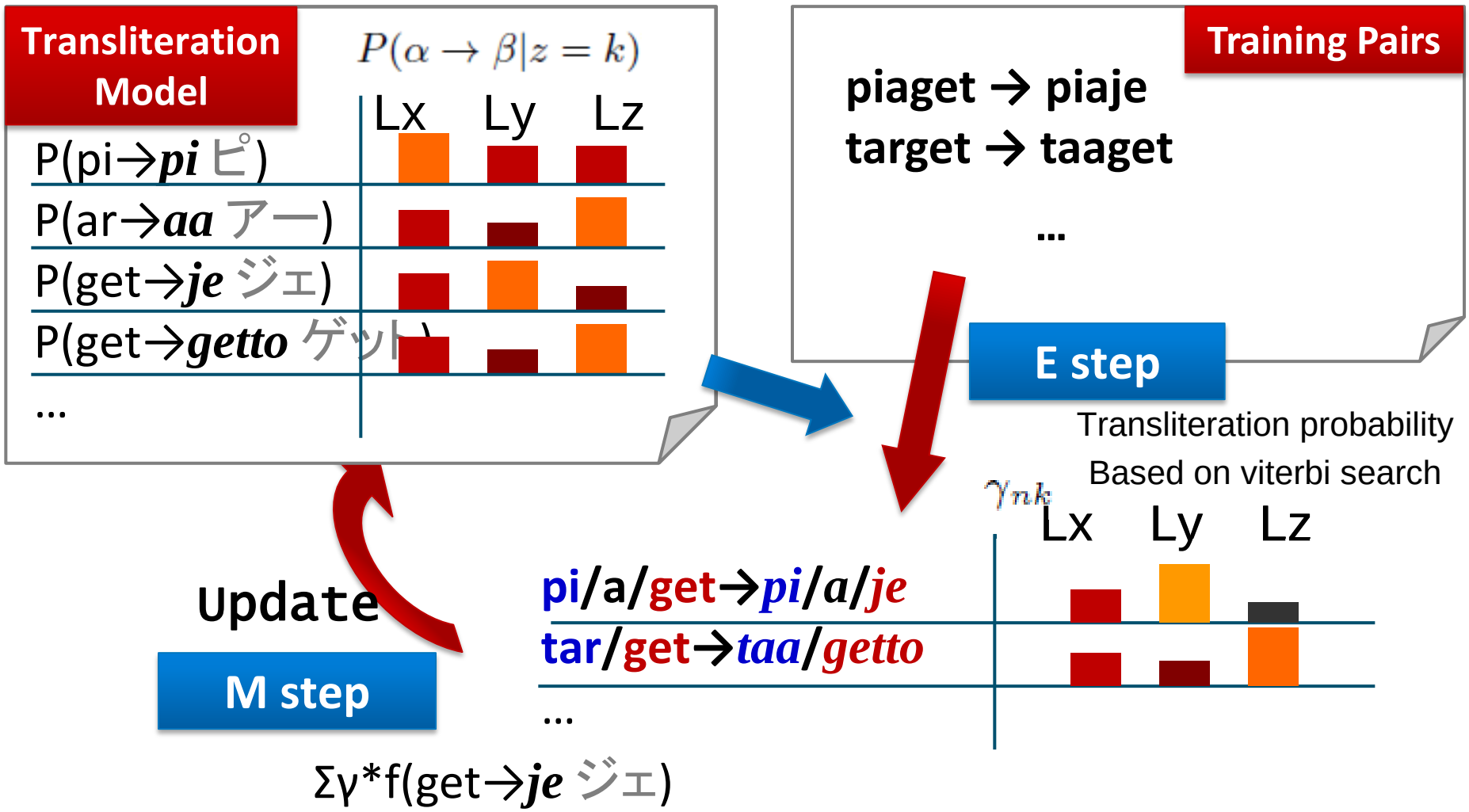
楽天 Rakuten [Hagiwara and Sekine 2011]

Iterative Learning via EM Algorithm



楽天 Rakuten [Hagiwara and Sekine 2011]

Iterative Learning via EM Algorithm

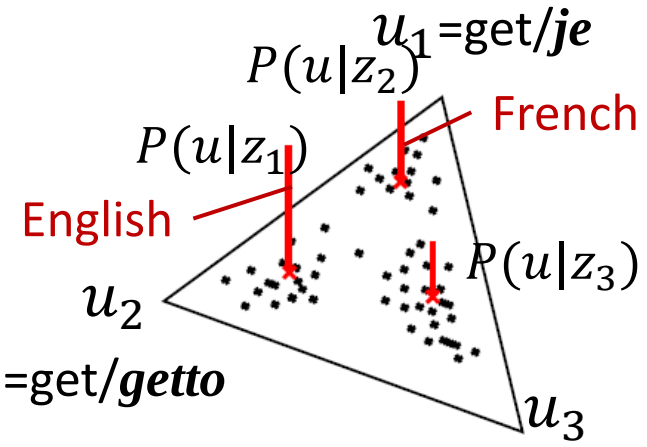


楽天 Rakuten [Hagiwara and Sekine 2011]

Latent Semantic Transliteration Model using Dirichlet Mixture

Latent Class Transliteration [Hagiwara&Sekine 11]

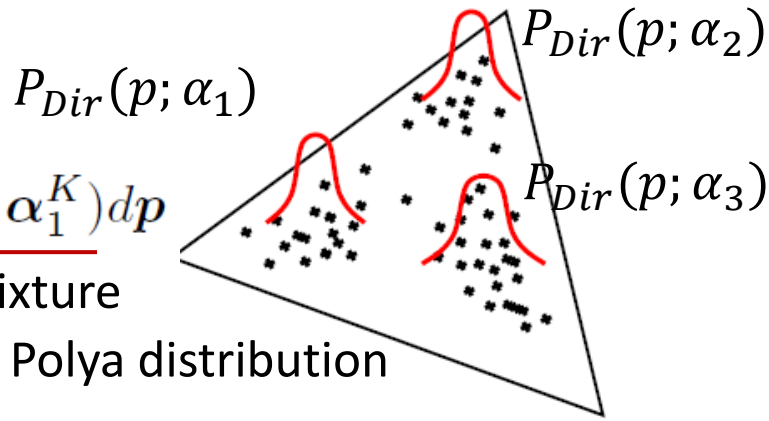
$$P_{\mathrm{LST}}(\langle \boldsymbol{s}, \boldsymbol{t} \rangle) = \sum_{z=1}^{K} P(z) \prod_{i=1}^{f} P(u_i|z)$$



Latent Semantic Transliteration using Dirichlet Mixture (Proposed)

$$P_{DM}(\langle \boldsymbol{s}, \boldsymbol{t} \rangle) = \int_{K} \underbrace{P_{Mul}(\langle \boldsymbol{s}, \boldsymbol{t} \rangle; \boldsymbol{p})}_{\text{Multinomial}} \underbrace{P_{DM}(\boldsymbol{p}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^K)}_{\text{Dirichlet Mixture}} d\boldsymbol{p}$$

$$\propto \sum_{k=1} \lambda_k P_{Polya}(\langle \boldsymbol{s}, \boldsymbol{t} \rangle; \boldsymbol{\alpha}_1^K) \quad \text{Polya distribution}$$



楽天 Rakuten [Hagiwara and Sekine 2012]

Discriminative Transliteration Model

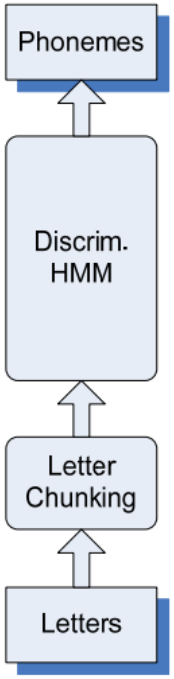
construct [kænstrʌkt]

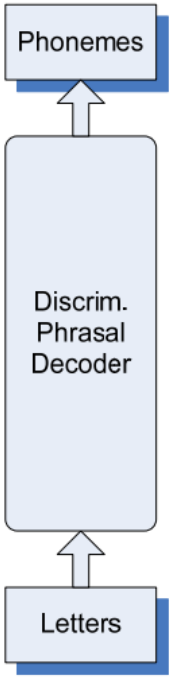
features:

(s, s), (ns, s), (st, s), (ons, s), (nst, s) ...

(n, s)

(s, ns), (ns, ns), (st, ns), (ons, ns), (nst, ns), ...



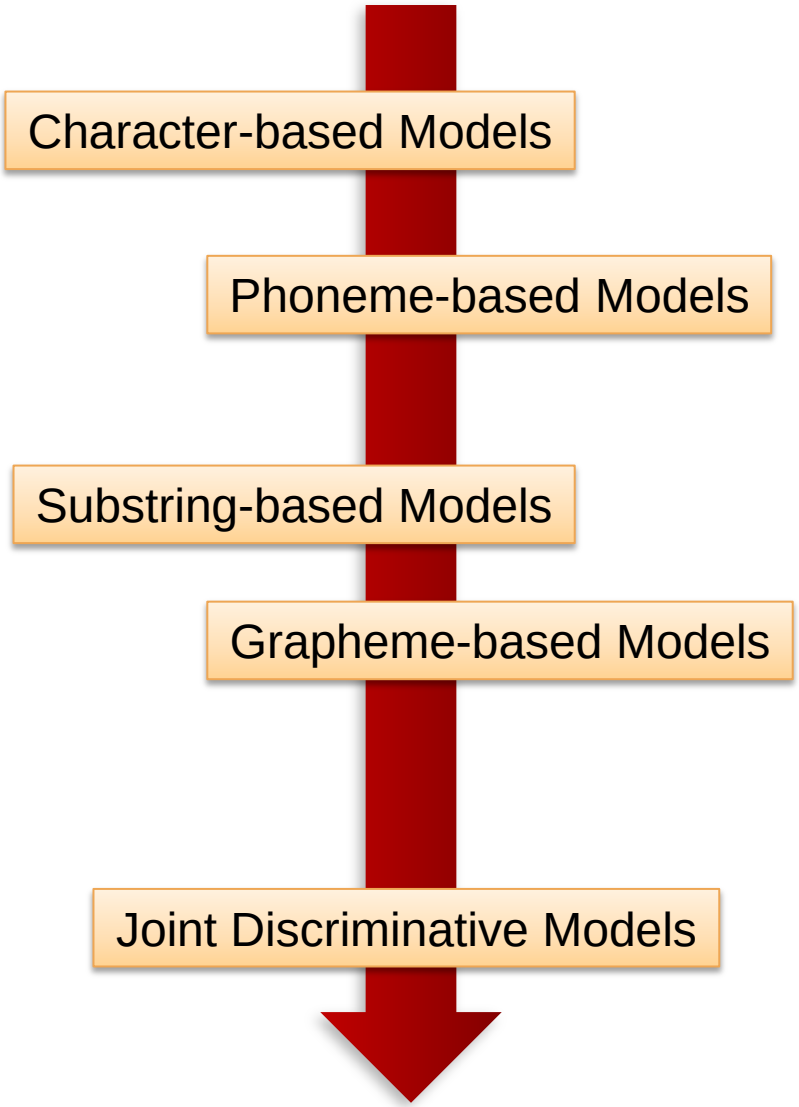


Predicting: $\hat{y} = \arg\max_{y'} [\alpha \cdot \Phi(x, y')]$

search: monotone search for phrasal decoder

楽天 Rakuten [Jiampojamarn et al. 2008] [Cherry and Suzuki 2009]

Transliteration Evolution



Transliteration Model Evolution

Character	*Substring* → ***Substring***
Phoneme	*Grapheme* → ***Grapheme***
Uniform	*Semantic* → ***Semantic***
Generative	*Discriminative* → ***Discriminative***

Integrated Models

Compound Noun and Transliteration

ブラキッシュレッド
(*burakissyureddo*)

ブラキッ　シュレッド

*bracki　shred

ブラキッシュ　レッド

blackish　red

Transliteration Model

贝拉克奥巴马
(*beilakeaobama*)

贝拉克　奥巴马

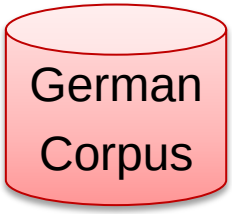
barack　obama

English Language Model

楽天 Rakuten

Source/Target Language Statistics

aktionsplan

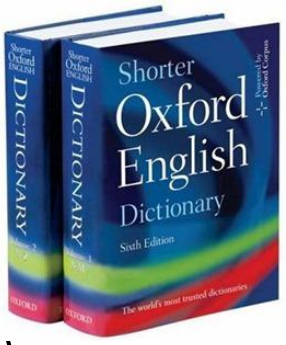


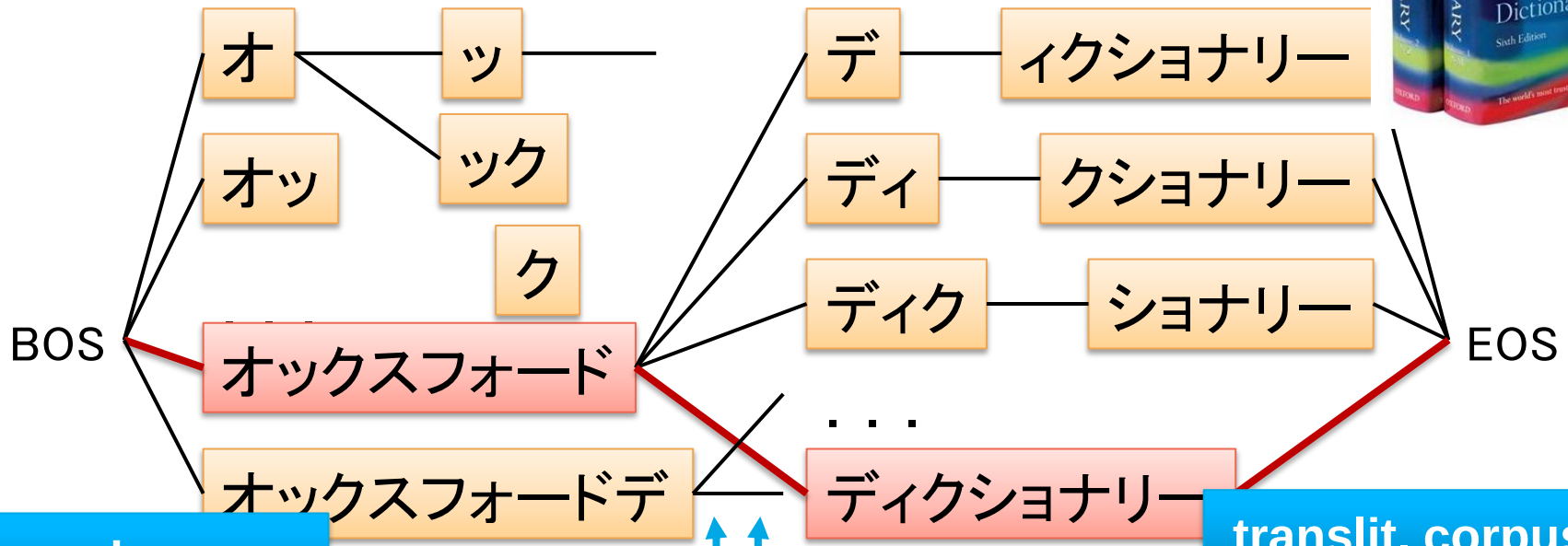
aktionsplan(852) ???		852	0
aktion(960) action	plan(710) plan	825.6	2
aktions(5) ???	plan(710) plan	59.6	1
akt(224) ???	ion(1) ???	plan(710) plan	54.2	1

楽天 Rakuten [Koehn and Knight 2003]

Use of Monolingual Paraphrase and Transliteration

オックスフォードディクショナリー
(*okkusufoododikushonarii*)





paraphrases

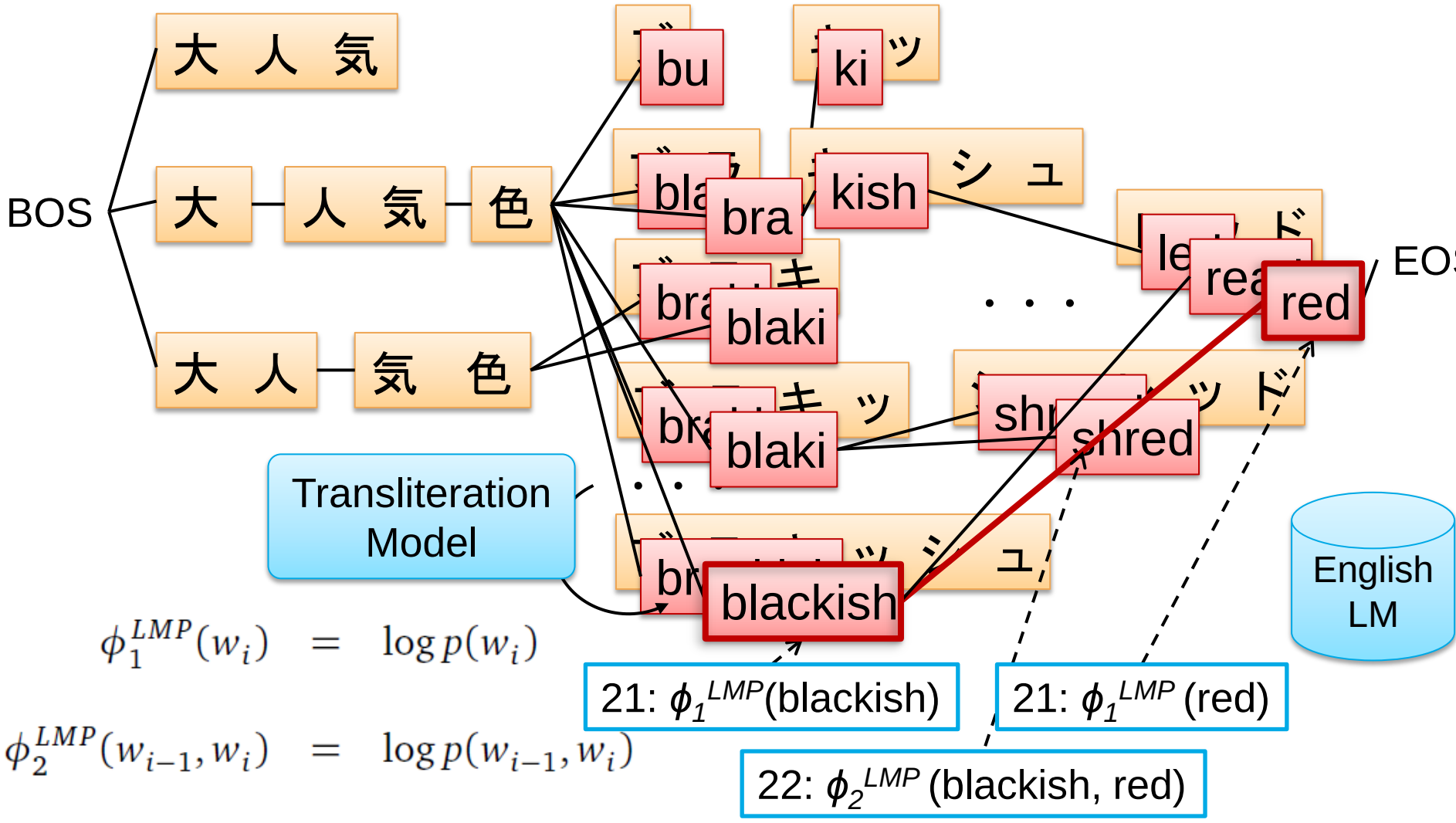
アンチョビパスタ anchovy pasta
アンチョビ・パスタ anchovy pasta
アンチョビのパスタ anchovy pasta

translit. corpus

オックスフォード oxford
ディクショナリー dictionary
ジャンク フード junk food

楽天 Rakuten [Kaji, Kitsuregawa 2011]

Language Projection via “Online” Transliteration



$$\phi_1^{LMP}(w_i) = \log p(w_i)$$

$$\phi_2^{LMP}(w_{i-1}, w_i) = \log p(w_{i-1}, w_i)$$

楽天 Rakuten [Hagiwara, Sekine 2013]

Agenda

Word Segmentation

Transliteration

Integrated Models

References – Chinese Word Segmentation

[Wong and Chan 1996] Pak-kwong Wong and Chorkin Chen.
Chinese Word Segmentation based on Maximum Matching and Word Binding Force, COLING 1996.

[Xue and Shen 2003] Nianwen Xue and Libin Shen.
Chinese Word Segmentation as LMR Tagging, SIGHAN 2003.

[Xue 2003] Nianwen Xue,
Chinese Word Segmentation as Character Tagging, Computational Linguistics and Chinese Language Processing, 2003.

[Peng et al. 2004] Fuchun peng, Fangfang Feng, Andrew McCallum.
Chinese Segmentation and New Word Detection using Conditional Random Fields, COLING 2004.

[Kruengkrai et al. 2009] Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, Hitoshi Isahara.
An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging, ACL/IJCNLP 2009.

[Ng and Low 2004] Hwee Tou Ng and Jin Kiat Low.
Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? EMNLP 2004.

[Zhang and Clark 2008] Yue Zhang and Stephen Clark.
Joint Word Segmentation and POS Tagging using a Single Perceptron, ACL 2008.

References – Japanese Morphological Analysis

[Yoshimura et al. 1983] 吉村 賢治, 日高 達, 吉田 将
文節数最小法を用いたべた書き日本語文の形態素解析, 情報処理学会論文誌, 1983.

[Kudo et al. 2004] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto
Applying Conditional Random Fields to Japanese Morphological Analysis, EMNLP 2004.

[Nakagawa and Uchimoto 2007] Tetsuji Nakagawa and Kiyotaka Uchimoto.
A Hybrid Approach to Word Segmentation and POS Tagging, ACL 2007.

[Neubig et al. 2011] Graham Neubig, Yosuke Nakata, Shinsuke Mori.
Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, ACL 2011.

[Okanohara and Tsujii 2008] 岡野原 大輔, 辻井 潤一
Shift-Reduce操作に基づく未知語を考慮した形態素解析, JNLP 2008.

References – Transliteration

[Knight and Graehl 1998] Kevin Knight and Jonathan Graehl.
Machine Transliteration, Computational Linguistics, 1998.

[Li et al. 2004] Haizhou Li, Min Zhang, Jian Su.
A Joint Source-Channel Model for Machine Transliteration, ACL 2004.

[Li et al. 2007] Haizhou Li* Khe Chai Sim, Jin-Shea Kuo, Minghui Dong.
Semantic Transliteration of Personal Names, ACL 2007.

[Hagiwara and Sekine 2011] Masato Hagiwara and Satoshi Sekine.
Latent Class Transliteration based on Source Language Origins. ALC-HLT, 2011.

[Hagiwara and Sekine 2012] Masato Hagiwara and Satoshi Sekine.
Latent Semantic Transliteration using Dirichlet Mixture. NEWS 2012.

[Jiampojamarn et al. 2007] Sittichai Jiampojamarn, Grzegorz Kondrak and Tarek Sherif.
Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion,
NAACL 2007.

[Jiampojamarn et al. 2008] Sittichai Jiampojamarn, Colin Cherry, Grzegorz Kondrak.
Joint Processing and Discriminative Training for Letter-to-Phoneme Conversion, NAACL 2008.

[Cherry and Suzuki 2008] Colin Cherry and Hisami Suzuki.
Discriminative Substring Decoding for Transliteration, EMNLP 2008.

References – Integrated Models

[Koehn and Knight 2003] Philipp Koehn and Kevin Knight.
Empirical Methods for Compound Splitting, EACL 2003.

[Kaji and Kitsuregawa 2011] Nobuhiro Kaji and Masaru Kitsuregawa.
Splitting Noun Compounds via Monolingual and Bilingual Paraphrasing: A Study on Japanese Katakana Words, EMNLP 2011

[Hagiwara and Sekine 2013] Masato Hagiwara and Satoshi Sekine.
Accurate Word Segmentation using Transliteration and Language Model Projection,
ACL 2013 (to appear)