Word Segmentation and Transliteration in Chinese and Japanese

Masato Hagiwara
Rakuten Institute of Technology, New York

CUNY NLP Seminar 4/5/2013
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)
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)

何日怎么章鱼说

How do you say octopus in Japanese?

Japanese | octopus | how | say
---|---|---|---
(day) | (Japanese) | (octopus) | (fish) | (how) | (say)
Examples Where Maximum Match Fails

Police gun-kill (perf.) that (mes.) escapee

Police with gun kill (perf.) that escapee

Heuristic rules
“Word Binding” Scores
Heuristic Approaches – Minimum Bunsetsu Number

今 日 本 当 に 良 い 天 気 で す ね

Today, really, good weather is (part.)

# of Bunsetsu = 4

Now, Japan, really, good weather is (part.)

# of Bunsetsu = 5

Optimizes for a whole sentence

楽 R 天 Rakuten [Yoshimura et al. 1983]
What is Bunsetsu?

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

창대의 밑에는,
 걱정일 것 같은 얼굴을 하고 있는 아버지와 엄마.
‘마아훈’이 눈을 뜨는 것을
 지금이나 지금일까하고 기다리고 있습니다.

수술은 끝났지만,
 의사에는 어려운 얼굴을 하고 있었습니다.

앞으로도 죽.
 ‘마아훈’은 병과 함께
 살아가지 않으면 안 됩니다.

みずを　いれて　くしゃくしゃ，かきまわし，
どろどろに，とかすと，べちゃべちゃ，どろの
つちのペンキが，できました。
Minimum Bunsetsu Number

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

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

where

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

A Special Case of Minimum Cost Methods

[Yoshimura et al. 1983]
Word-based Models

\[
\min \sum_{i=1}^{N} \left[ \text{cost}_1(w_i) + \text{cost}_2(w_{i-1}, w_i) \right]
\]

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

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

Kudo et al. 2004

Tokyo → Kyoto
Character-based Models 1 – Character Tagging

Police used a gun to kill the escaped criminal

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

LMR Tagging

- 警 [B]
- 察 [B]
- 用 [B]
- 枪 [B]
- 杀 [B]

- 用 [S]

- 计算机 [B] [I] [E]

[Xue and Shen 2003] [Peng et al. 2004]
Character-based Models 2 – “Boundary” Tagging

boundary tags

segmented words

boundary tags

SVM Logistic Regression

$x_l, x_r, x_{l-1}x_l, x_lx_r, \ldots$
$c(x_l), c(x_r), \ldots$
$l_s, r_s, i_s, \ldots$(dict. feat.)

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]

[Neubig et al. 2011]
One-at-a-Time PoS Tagging Models

boundary tags

segmented words

POS Tags

Enables Domain Adaptation through Partial Annotation

[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”

F-measure

0 10 20 30 40 50 60 70 80 90 100

Iterations

94.40%

94.90%

95.40%

95.90%

96.40%

2step-Lr/part
2step-Lr/sent
2step-crf/sent
joint/sent
Character-based Joint Models

警察用枪杀了那个逃犯

警 [B-Noun]
察 [B-Noun]
用 [B-Noun]
枪 [B-Noun]
杀 [B-Noun]

警 [S-Noun]
察 [S-Noun]
用 [S-Noun]
枪 [S-Noun]
杀 [S-Noun]

警 [I-Noun]
察 [I-Noun]
用 [I-Noun]
枪 [I-Noun]
杀 [I-Noun]

警 [E-Noun]
察 [E-Noun]
用 [E-Noun]
枪 [E-Noun]
杀 [E-Noun]

用 [S-Prep]

计算机 [B-Noun] [I-Noun] [E-Noun]

...
Stack Decoding Models

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

[Zhang, Clark 2008] [Okanohara, Tsujii 2008]
Chinese/Japanese WS Evolution

**Chinese**

- Heuristics
  - Maximum Forward Match
- Character-based
  - One-at-a-Time Models
- Character-based
  - All-at-once Models
- Stack Decoding Models
- Boundary Tagging Models

**Japanese**

- Heuristics
  - Minimum Bunsetsu Number
- Semi-Markov
  - Word-based Models

**Models**

- Generative Models
- Discriminative Models
<table>
<thead>
<tr>
<th>Chinese/Japanese WS Evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heuristics</strong></td>
</tr>
<tr>
<td>Word-based</td>
</tr>
<tr>
<td>Pipeline (One-at-a-time)</td>
</tr>
<tr>
<td>Generative</td>
</tr>
<tr>
<td>Viterbi</td>
</tr>
<tr>
<td><strong>Statistics</strong></td>
</tr>
<tr>
<td>Character-based</td>
</tr>
<tr>
<td>Joint (All-at-Once)</td>
</tr>
<tr>
<td>Discriminative</td>
</tr>
<tr>
<td>Discriminative</td>
</tr>
<tr>
<td>→ Discriminative</td>
</tr>
<tr>
<td>→ Pros and Cons</td>
</tr>
</tbody>
</table>
Transliteration
(“Semantic” Transliteration Models)
Transliteration

Phonetic translation between languages with different writing systems

New York / 纽约 niuyue / ニューヨーク nyuuyyooku

Obama / 奥巴马 aobama / オバマ obama
Phoneme-based Methods

English word

$P(w)$
golf ball

English phoneme

$P(e|w)$
G A A L F B A O L

Japanese phoneme

$P(j|e)$
g o r u h u b o o r u

Japanese word

$P(k|j)$
ゴルフボール

Trains a large WFST (from Japanese to English words)

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

[Knight and Graehl 1998]
Direct Orthographical Mapping

Joint Source Channel Model

Transliteration Prob. \(= \) Prod. of TU n-gram probs.

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

\[P(\text{flextime} \rightarrow \text{furekkusutai}me) = P(f \rightarrow fu | \text{BOW}) \times P(\text{le} \rightarrow \text{re} | f \rightarrow fu) \times P(x \rightarrow \text{kkusu} | \text{le} \rightarrow \text{re}) \times \ldots
\]

TU Probability Estimation

TU Probability Table

\[P(\text{fl} \rightarrow \text{flek} | \cdot) = XXX\]
\[P(\text{ext} \rightarrow \text{ku} | \cdot) = YYY\]
\[P(\text{p} \rightarrow \text{pi} | \cdot) = ZZZ\]
\[\ldots\]

Current Alignment

Freq. \(\rightarrow\) Prob.

EM Algorithm

Viterbi Algorithm

Training Corpus

[Li et al. 2004]
Multiple Language Origins

piaget / piaje ピアジェ
French origin
French model

target / taagetto ターゲット
English origin
English model

亚历山大 Yalishanda / Alexander
Indo-European origin
Chinese Transliteration Model

山本 Yamamoto / Yamamoto
Japanese origin
Japanese Reading Model

Maryan / Malian 玛丽安
Female name
Female model

Marino / Malinuo 马里诺
Male name
Male model

[Li et al. 2007]
Latent Class Transliteration

**Class transliteration [Li et al. 2007]**

\[ P_{LI}(t|s) = \sum_c P(t, c|s) = \sum_c P(c|s)P(t|c, s) \]

s: source
t: target

Explicit language detection

**Latent Class Transliteration [Hagiwara & Sekine 2011]**

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

z: latent class
K: # of latent classes
(determined using dev. sets)

Latent class distribution

[Hagiwara and Sekine 2011]
Iterative Learning via EM Algorithm

Transliteration Model

\[ P(\alpha \rightarrow \beta | z = k) \]

<table>
<thead>
<tr>
<th></th>
<th>Lx</th>
<th>Ly</th>
<th>Lz</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(pi→pi)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(ar→aa)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(get→je)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(get→getto)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Training Pairs

piaget $\rightarrow$ piaje
target $\rightarrow$ taaget
...

\[ \gamma_{nk} \]

Transliteration probability
Based on viterbi search

P(pi/a/get $\rightarrow$ pi/a/j)
tar/get $\rightarrow$ taa/getto
...

[Hagiwara and Sekine 2011]
Iterative Learning via EM Algorithm

Transliteration Model

\[ P(\alpha \rightarrow \beta | z = k) \]

<table>
<thead>
<tr>
<th></th>
<th>Lx</th>
<th>Ly</th>
<th>Lz</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(pi→pi)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(ar→aa)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(get→je)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(get→getto)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Training Pairs

piaget → piaje
target → taaget
...

E step

Transliteration probability

Based on viterbi search

\[ \gamma_{nk} \]

<table>
<thead>
<tr>
<th></th>
<th>Lx</th>
<th>Ly</th>
<th>Lz</th>
</tr>
</thead>
<tbody>
<tr>
<td>pi/a/get</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tar/get</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Update

\[ \Sigma \gamma * f(get \rightarrow je) \]

M step

[Hagiwara and Sekine 2011]
Latent Semantic Transliteration Model using Dirichlet Mixture

**Latent Class Transliteration** [Hagiwara & Sekine 11]

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

**Latent Semantic Transliteration using Dirichlet Mixture (Proposed)**

\[
P_{DM}(\langle s, t \rangle) = \int P_{Mul}(\langle s, t \rangle; p) P_{DM}(p; \lambda, \alpha^K_1) dp
\]

\[
\propto \sum_{k=1}^{K} \lambda_k P_{Polya}(\langle s, t \rangle; \alpha^K_1)
\]

\[
P_{Dir}(p; \alpha_1)
\]

\[
P_{Dir}(p; \alpha_2)
\]

\[
P_{Dir}(p; \alpha_3)
\]

\[
P(u | z_1)
\]

\[
P(u | z_2)
\]

\[
P(u | z_3)
\]

\[
\begin{align*}
u_1 &= \text{get/jé} \\
u_2 &= \text{get/getto} \\
u_3 &= \text{French}
\end{align*}
\]

Multinomial Dirichlet Mixture [Hagiwara and Sekine 2012]
Discriminative Transliteration Model

features:

\[(s, s), (ns, s), (st, s), (ons, s), (nst, s) \ldots\]
\[(n, s)\]
\[(s, ns), (ns, ns), (st, ns), (ons, ns), (nst, ns) \ldots\]

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

search: monotone search for phrasal decoder

[Discriminative Transliteration Model] [Jiampojamarn et al. 2008] [Cherry and Suzuki 2009]
Transliteration Evolution

- Character-based Models
- Phoneme-based Models
- Substring-based Models
- Grapheme-based Models
- Joint Discriminative Models

Generative Models

Discriminative Models
Transliteration Model Evolution

<table>
<thead>
<tr>
<th>Character</th>
<th>Substring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoneme</td>
<td>Grapheme</td>
</tr>
<tr>
<td>Uniform</td>
<td>Semantic</td>
</tr>
<tr>
<td>Generative</td>
<td>Discriminative</td>
</tr>
</tbody>
</table>

Characters $\rightarrow$ Substring $\rightarrow$ Grapheme $\rightarrow$ Semantic $\rightarrow$ Discriminative
Integrated Models
Compound Noun and Transliteration

ブラキッシュレッド (burakissyurededo)

ブラキッシュレッド

ブラキッ
bracki

シュレッド
shred

ブラックオバマ (beilakeaobama)

ベラック
barack

オバマ
obama

English Language Model

Transliteration Model

ブラキッシュ
blackish

レッド
red
Source/Target Language Statistics

aktionsplan

aktionsplan(852) 852 0

aktion(960) action

plan(710) plan

aktions(5) ???

plan(710) plan

akt(224) ???

ion(1) ???

plan(710) plan

German Corpus

Bilingual Resources

[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)

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
dictionary
ジャンクフード junk food

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

[Kaji, Kitsuregawa 2011]
Language Projection via “Online” Transliteration

Transliteration Model

\[ \phi^\text{LMP}_1(w_i) = \log p(w_i) \]

\[ \phi^\text{LMP}_2(w_{i-1}, w_i) = \log p(w_{i-1}, w_i) \]

[Hagiwara, Sekine 2013]
Agenda

- Word Segmentation
- Transliteration
- Integrated Models
References – Chinese Word Segmentation


[Xue and Shen 2003] Nianwen Xue and Libin Shen. 

[Xue 2003] Nianwen Xue, 


[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.

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


[Okanohara and Tsujii 2008] 岡野原 大輔, 辻井 潤一
References – Transliteration


References – Integrated Models

