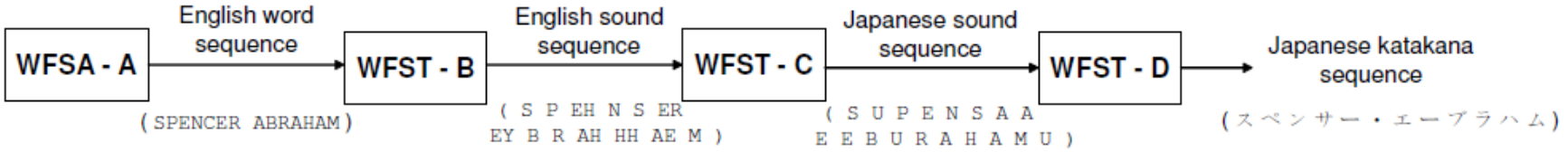
Non-Stochastic Learning of Cross-Language Transliteration Rules From Small Dataset

Варвара Логачева, Эдуард Клышинский
Институт прикладной математики
им. М.В.Келдыша
г. Москва

Существующие методы транскрипции (Knight and Grael)

K. Knight and J. Graehl. 1998. Machine transliteration. Computational Linguistics, 24(4):599–612.



WFSA-A – разделение входа на слова
WFST-B – генерация фонемного состава английского языка
WFST-C – преобразование в фонемный японского языка
WFST-D – преобразование в азбуку катакана
На каждом этапе используется вероятностная информация

Статистический метод – обучение конечного автомата
Задача обратной транслитерации (японский → английский)

Генерация правил

adria;адрия
adrian;адриан
adriana;адриана
adriane;адриан
adrianna;адрианна
adrianne;адрианн
adrien;адриан
adriene;адриен
adrienn;адрианн
adrienna;адрианна
aldema;альдема
aldena;альдена
aldenaide;альденед
aldenaise;альденез

A → A
AE → A
{<}AI → Э
{*}AI{*} → Е
{<}AY → Э
{*}AY → Е
AIL{>} → АЙ
AILL{*} → АЙ
AILLE{>} → АЙ

Первичные правила

isVowel (l) = $\begin{cases} \textbf{true}, \text{ если } l \text{ – гласная} \\ \textbf{false}, \text{ если } l \text{ – согласная} \end{cases}$

Слово w = $l_1 l_2 \ldots l_n$

Граница группы – между l_i и l_{i+1} такими, что

isVowel (l_i) ≠ isVowel (l_{i+1})

R | u | gg | ie | r | o → R | u | dzh | e | r | o

M | a | cch | i → M | a | kk | i

Учет контекста

l → л

anjela – анжела

cella – селла

l → ль

almelda – альмельда

avital – авиталь

l {a, o, u, i, e, l} → л

l {>, t, d, p, m} → ль

Деление на слоги

A|d r i a|n a

A|l d e|n a

A|l m e|l i|n e

A|д р и а|н а

A|л ь д е|н а

A|л ь м е|л и н

Порождение сложных правил

Каждую пару слогов можно представить как

$<p_{i1}, \ldots, p_{ik}, \mathbf{p_x}, p_{ik+1}, \ldots, p_{in}> \rightarrow$
$<c_{i1}, \ldots, c_{ik}, \mathbf{c_x}, c_{ik+1}, \ldots, c_{im}>,$
где $\mathbf{p_x} \rightarrow \mathbf{c_x}$ – подстрока, не удовлетворяющая ни одному из существующих правил.

Можно выделить три случая несоответствия $\mathbf{p_x}$ правилам.

Порождение сложных правил

- **px** = ∅, **cx** ≠ ∅
- **px** ≠ ∅, **cx** = ∅
- **px** ≠ ∅, **cx** ≠ ∅

lde → льде

de → ьде

d → ьд

_ → ь

l → л

e → е

d → д

l →

ль

Результаты: качество обучения

Исходный язык	**CT**	**UCT**	**ATV**
Японский	7005 (**99%**)	4778 (**68%**)	1,38
Китайский	4468 (**95%**)	4173 (**89%**)	1,06
Немецкий	3484 (**82%**)	3247 (**77%**)	1,07
Арабский	2102 (**99%**)	1793 (**85%**)	1,19
Шведский	1576 (**88%**)	905 (**50%**)	1,61
Польский	1424 (**99%**)	1174 (**81%**)	1,2
Испанский	1025 (**98%**)	777 (**74%**)	1,33
Французский	678 (**89%**)	227 (**29%**)	2,66
Румынский	565 (**97%**)	295 (**52%**)	1,78
Словенский	502 (**99%**)	354 (**70%**)	1,3
Тагальский	257 (**87%**)	225 (**76%**)	1,14
Монгольский	231 (**100%**)	227 (**98%**)	1,02
Хинди	160 (**99%**)	152 (**94%**)	1,05

Результаты: качество транскрипции

Язык	**Тестовая выборка**	**CT**	**UCT**	**ANL**	**AE**
Японский	701	664 (**94%**)	496 (**70%**)	0,043	1,125
Китайский	468	436 (**93%**)	406 (**86%**)	0,026	1,23
Немецкий	421	339 (**80%**)	325 (**77%**)	0,032	1,176
Арабский	211	169 (**80%**)	160 (**75%**)	0,054	1,5
Шведский	178	149 (**83%**)	82 (**46%**)	0,131	1,33
Польский	144	138 (**95%**)	123 (**85%**)	0,027	1,5
Испанский	105	99 (**94%**)	76 (**72%**)	0,038	1,125

Конец