Cross-Language IR at University of Tsukuba

Automatic Transliteration for Japanese, English, and Korean

Atsushi Fujii, Tetsuya Ishikawa
University of Tsukuba

Motivation

- We developed an automatic transliteration method for Japanese and English CLIR
- the method has been used in commercial CL patent service
- In NTCIR-4 CLIR, we applied our method to Korean and realized JEK transliteration in a single framework

Classification of CLIR methods

- query translation method
- document translation method
- interlingual method (thesauri and LSI)
- hybrid method (combining QT and DT)

Query Translation

- translate compound query terms

1. consult a dictionary to derive all the possible word/phrase translation candidates
2. transliterate out-of-dictionary loanwords on a phonogram-by-phonogram basis
3. resolve translation ambiguity through a probabilistic method

Query Translation (cont.)

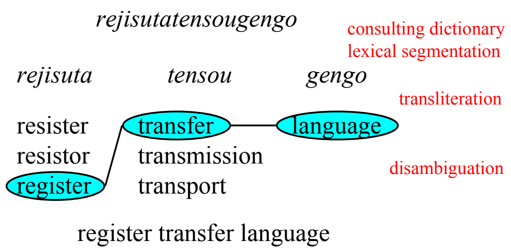
- compound query S and a translation candidate T

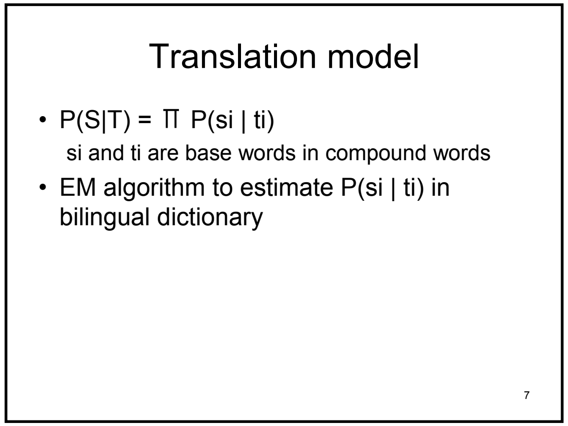
  S = s1, s2, …, sN
  T = t1, t2, …, tN

- compute $P(T|S) = \underbrace{P(S|T)}_{\text{translation model}} \cdot \underbrace{P(T)}_{\text{language model}}$

- select the candidate with max P(T|S)

Example of J-E Query Translation

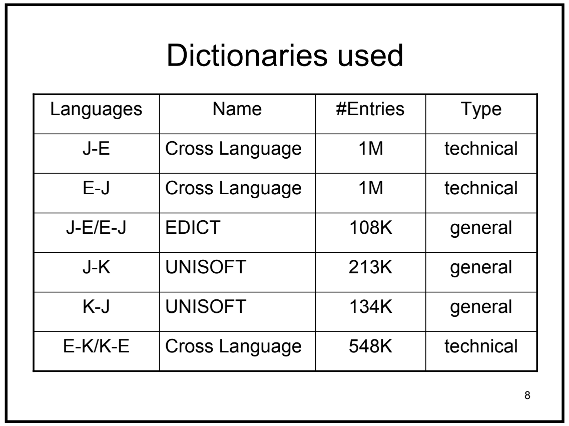




Translation model

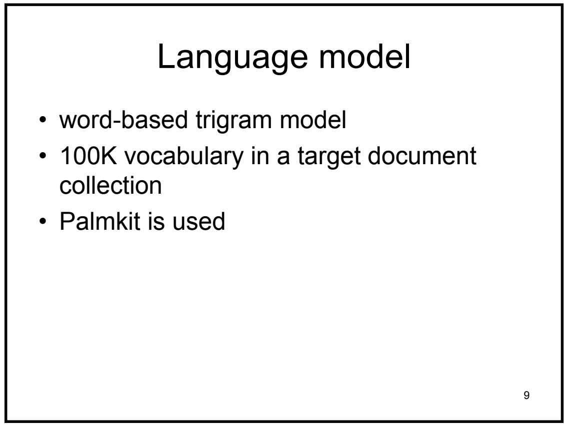
- $P(S|T) = \prod P(si \mid ti)$

  si and ti are base words in compound words
- EM algorithm to estimate P(si | ti) in bilingual dictionary



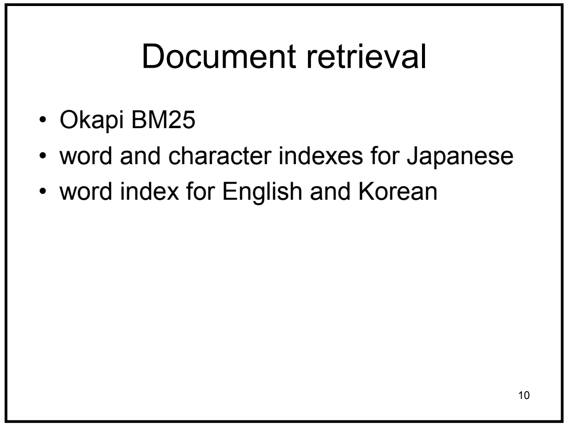
Dictionaries used

Languages	Name	#Entries	Type
J-E	Cross Language	1M	technical
E-J	Cross Language	1M	technical
J-E/E-J	EDICT	108K	general
J-K	UNISOFT	213K	general
K-J	UNISOFT	134K	general
E-K/K-E	Cross Language	548K	technical



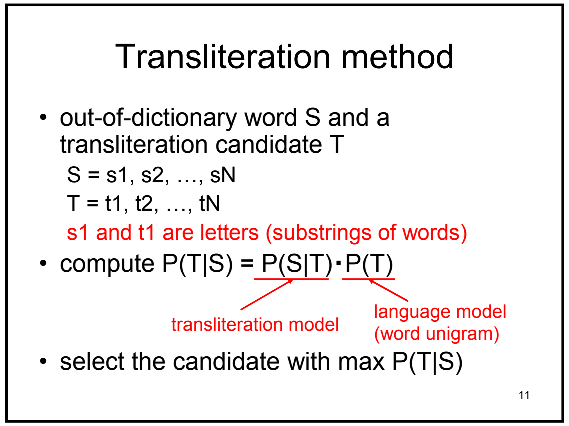
Language model

- word-based trigram model
- 100K vocabulary in a target document collection
- Palmkit is used



Document retrieval

- Okapi BM25
- word and character indexes for Japanese
- word index for English and Korean



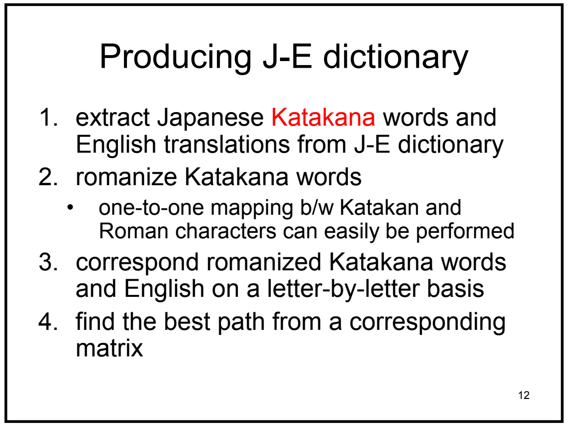
Transliteration method

- out-of-dictionary word S and a transliteration candidate T

  S = s1, s2, …, sN

  T = t1, t2, …, tN

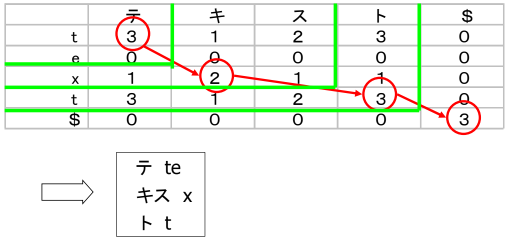
  s1 and t1 are letters (substrings of words)
- compute $P(T|S) = \underbrace{P(S|T)}_{\text{transliteration model}} \cdot \underbrace{P(T)}_{\text{language model (word unigram)}}$
- select the candidate with max P(T|S)



Producing J-E dictionary

1. extract Japanese Katakana words and English translations from J-E dictionary
2. romanize Katakana words
   - one-to-one mapping b/w Katakan and Roman characters can easily be performed
3. correspond romanized Katakana words and English on a letter-by-letter basis
4. find the best path from a corresponding matrix

Example matrix



	テ	キ	ス	ト	$
t	3	1	2	3	0
e	0	0	0	0	0
x	1	2	1	1	0
t	3	1	2	3	0
$	0	0	0	0	3

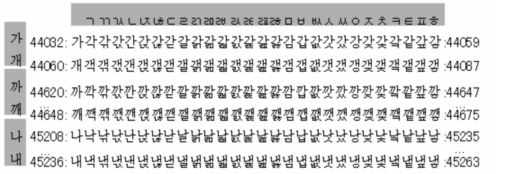
テ te
キス x
ト t

Producing J-K dictionary

- In EUC-KR, characters are coded independent of pronunciation
- one-to-one mapping b/w Hangul and Roman characters cannot easily be performed
  - # of Hangul characters is approx. 11,000
  - # of common characters is approx. 2,000
- we used Unicode, in which character is coded according to pronunciation

Romanizing Korean words

- first consonant changes every 21 lines
- vowel changes every line and repeats every 21 lines
- last consonant changes every column



ㄱ ㄲ ㄳ ㄴ ㄵ ㄶ ㄷ ㄹ ㄺ ㄻ ㄼ ㄽ ㄾ ㄿ ㅀ ㅁ ㅂ ㅄ ㅅ ㅆ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ

가 44032: 가각갂갃간갅갆갇갈갉갊갋갌갍갎갏감갑값갓갔강갖갗갘같갚갛 :44059
개 44060: 개객갞갟갠갡갢갣갤갥갦갧갨갩갪갫갬갭갮갯갰갱갲갳갴갵갶갷 :44087
…
까 44620: 까깍깎깏깐깑깒깓깔깕깖깗깘깙깚깛깜깝깞깟깠깡깢깣깤깥깦깧 :44647
깨 44648: 깨깩깪깫깬깭깮깯깰깱깲깳깴깵깶깷깸깹깺깻깼깽깾깿꺀꺁꺂꺃 :44675
…
나 45208: 나낙낚낛난낝낞낟날낡낢낣낤낥낦낧남납낪낫났낭낮낯낰낱낲낳 :45235
내 45236: 내낵낶낷낸낹낺낻낼낽낾낿냀냁냂냃냄냅냆냇냈냉냊냋냌냍냎냏 :45263
…

specific Hangul characters can be identified by pronunciation

Example of transliteration

Topic ID	Japanese	English	Korean
005	ダイオキシン	dioxin	다이옥신
006	マイケル・ジョーダン	Michael Jordan	마이클 조던
008	バイアグラ	viagra	비아그라
031	ユーゴスラビア	Yugoslavia	유고슬라비아

Experiments (J/E)

<TITLE>, mean average precision (rigid)

Languages	#Entries	w/o transliteration		w/ transliteration
J-E	1M	0.2174	<	0.2182
E-J	1M	0.1250	=	0.1250
J-E (EDICT)	108K	0.1147	<	0.1383
E-J (EDICT)	108K	0.0612	<	0.0857

transliteration was effective for small dictionaries

Experiments (Korean)

<TITLE>, mean average precision (rigid)

Languages	w/o transliteration		w/ transliteration
J-K	0.2177	<	0.2457
K-J	0.1486	<	0.1746
E-K	0.2026	<	0.2153
K-E	0.1017	<	0.1231

transliteration was also effective for Korean