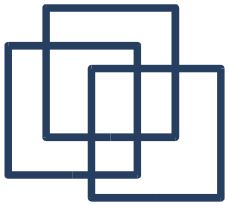
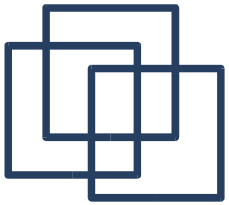


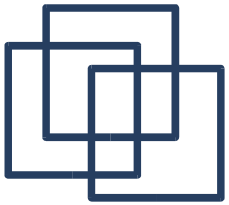
Implementations that Unify the Language Processing

Virach Sornlertlamvanich
Thai Computational Linguistics Laboratory (TCL), NICT
virach@tcllab.org



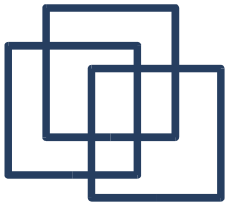
Outline

- Motivation
- Non-segmenting language
- Language identification
- Multi-lingual search engine
- Term-based ontology Alignment
- TCL's computation lexicon



Motivation

- Reliance on word segmentation
- Consistency in recognizing a word
- Updating the contemporary word list
- To establish an unified language processing

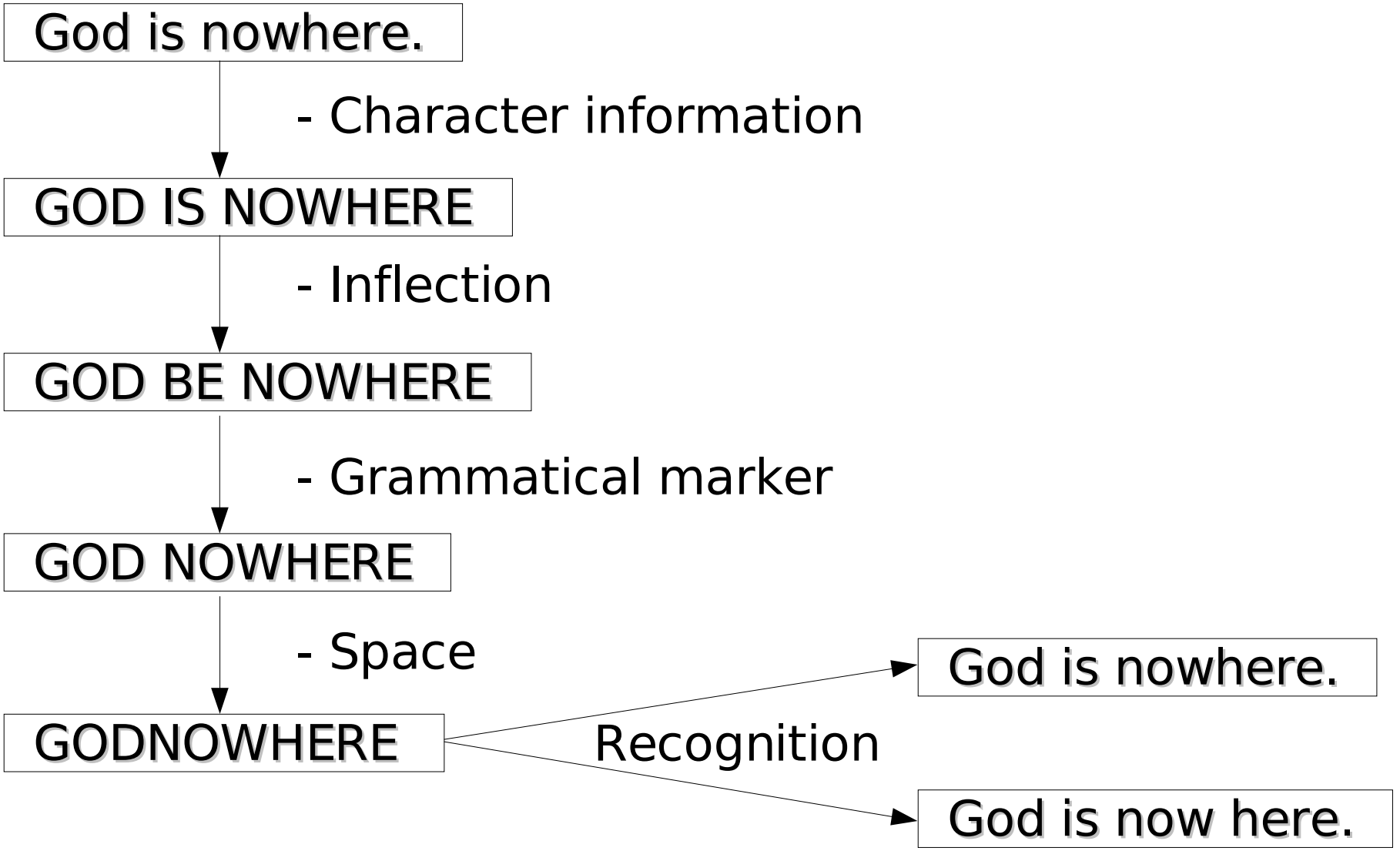


Thai Language as a Non-Segmenting Language

- No explicit word boudary marker e.g. capital letter, space character, punctuation mark, etc.
- No inflection
- No grammatical marker

How to determine word and sentence boundary?

Non-Segmenting Language (an example)



Difficulty in Word Segmentation

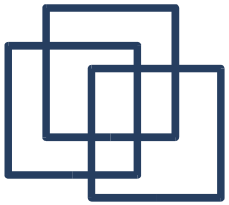
- **Ambiguity of being a word**

แบบนอก	->	แบบ | นอก	<> แบ | บน | อก
มีที่นา	->	มี | ที่นา	<> มี | ที่ | นา
ขอบอกขอบใจ	->	ขอบอก | ขอบใจ	<> ขอ | บอก | ขอบใจ
ขนมอบกรอบ	->	ขนม | อบ | กรอบ	<> ขน | มอบ | กรอบ
ร้านข้าวซอยลำดวน	->	ร้าน | ข้าวซอย | ลำดวน	<> ร้าน | ข้าว | ซอย | ลำดวน

- **Unknown word**

นาตาลี	->	นา | ตา | ลี
อยุธยาอะลิอันซ์ซีพี	->	อยุธยา | อะลิอันซ์ | ซี | พี
กาลิเลโอ	->	กา | ลิเล | โอ

- **Dictionary information** i.e. POS, thesaurus



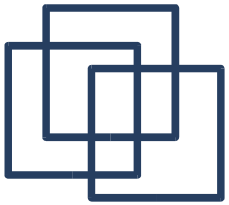
Word-Based Approach

- Word segmentation (accuracy for Thai)
  - Longest matching: 92%
  - Maximal matching: 93%
  - POS tri-gram: 96%
- Sentence segmentation (accuracy for Thai)
  - POS tri-gram: 84.57%
  - Feature-based approach (Winnow): 89.13%

Meaningful Bits

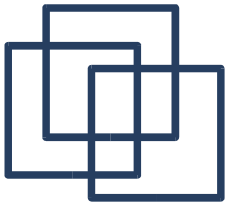
ADLTSUGKNOWLEDGEBWGZKTILA

ปร้จตเสีศฐาดีความรู้ษะภูกฮเศฉ



Language Identification

- Identify the language of a given text based on String Kernels
- Advantages:
  - Identify the language from the text directly, regardless its coding system
  - Not require linguistic presuppositions about the data
  - Derive properties of n-gram language model
  - Apply to any kernel classifiers



String Kernels

- A kernel := the inner product function between two vectors,

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

Explicit Mapping ($1 \le r \le 2$)

u = yzxxz, v = xyzxxxy

- By organizing all possible substrings in the lexicographic order, for substrings in Σ^1, we get:

Σ^1	$\phi_1(u)$	$\phi_1(v)$	$\phi_1(u)\cdot\phi_1(v)$
x	$2\lambda^1$	$4\lambda^1$	$8\lambda^2$
y	λ^1	$2\lambda^1$	$2\lambda^2$
z	$2\lambda^1$	λ^1	$2\lambda^2$

$$K_2(u,v) = 8\lambda^2 + 2\lambda^2 + 2\lambda^2 = 12\lambda^2$$

Explicit Mapping ($1 \le r \le 2$)

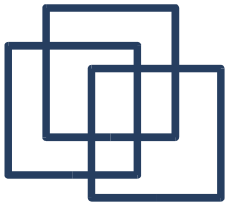
$u = yzxxz,\ v = xyzxxxy$

- For substrings in Σ^2, we get:

Σ^2	$\phi_2(u)$	$\phi_2(v)$	$\phi_2(u)\cdot\phi_2(v)$
xx	λ^2	$2\lambda^2$	$2\lambda^4$
xy	0	$2\lambda^2$	0
xz	λ^2	0	0
yx	0	0	0
yy	0	0	0
yz	λ^2	λ^2	λ^4
zx	λ^2	λ^2	λ^4
zy	0	0	0
zz	0	0	0

$$K_2(u,v) = 2\lambda^4 + \lambda^4 + \lambda^4 = 4\lambda^4$$

$$K_r(u,v) = K_1(u,v) + K_2(u,v) = 12\lambda^2 + 4\lambda^4$$



Brute-Force Matching

1 2 3 4 5 6 7

v : x y z x x x y

u : y

y = λ^2

y z = $\lambda^2 \cdot \lambda^2$

y

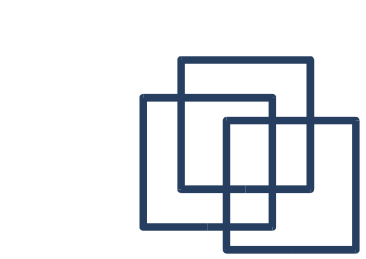
y

y

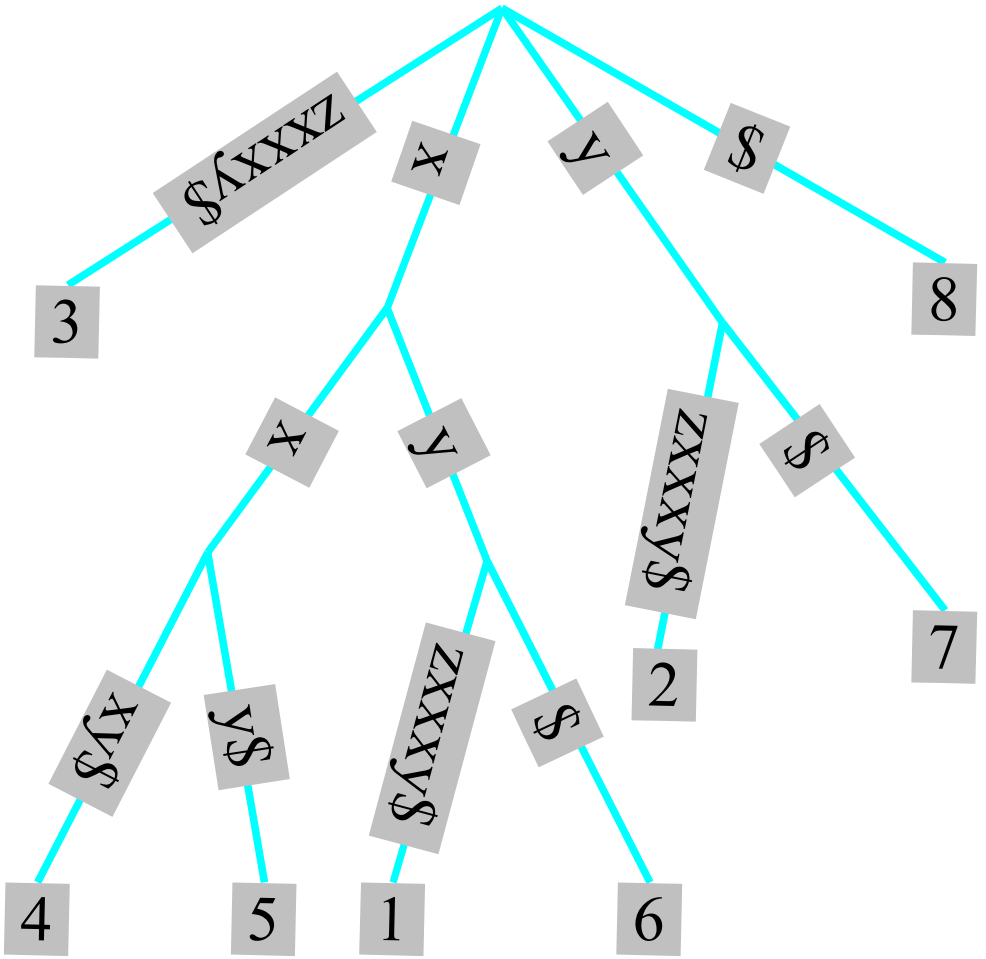
y

y = λ^2

The computational complexity is O(r|u||v|)

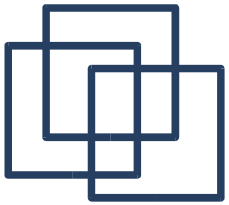


Faster Matching with Suffix Trees



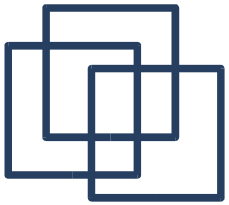
Suffix tree for the string v = xyzxxxy$

The computational complexity is O(c|u|+|v|)



Language Identification --Training and Test Corpus--

- Centroid-based and SVM classification methods based on string kernel
- 5 fold cross validation
- 3 groups of 20 languages
  - Asian: Thai, Chinese, Japanese, Korean
  - Roman alphabet: English, French, Italian, Portuguese, Spanish, Swedish, German, Hungarian
  - Slavic family: Czech, Polish, Croatian, Slovak, Slovenian, Bulgarian, Russian, Greek

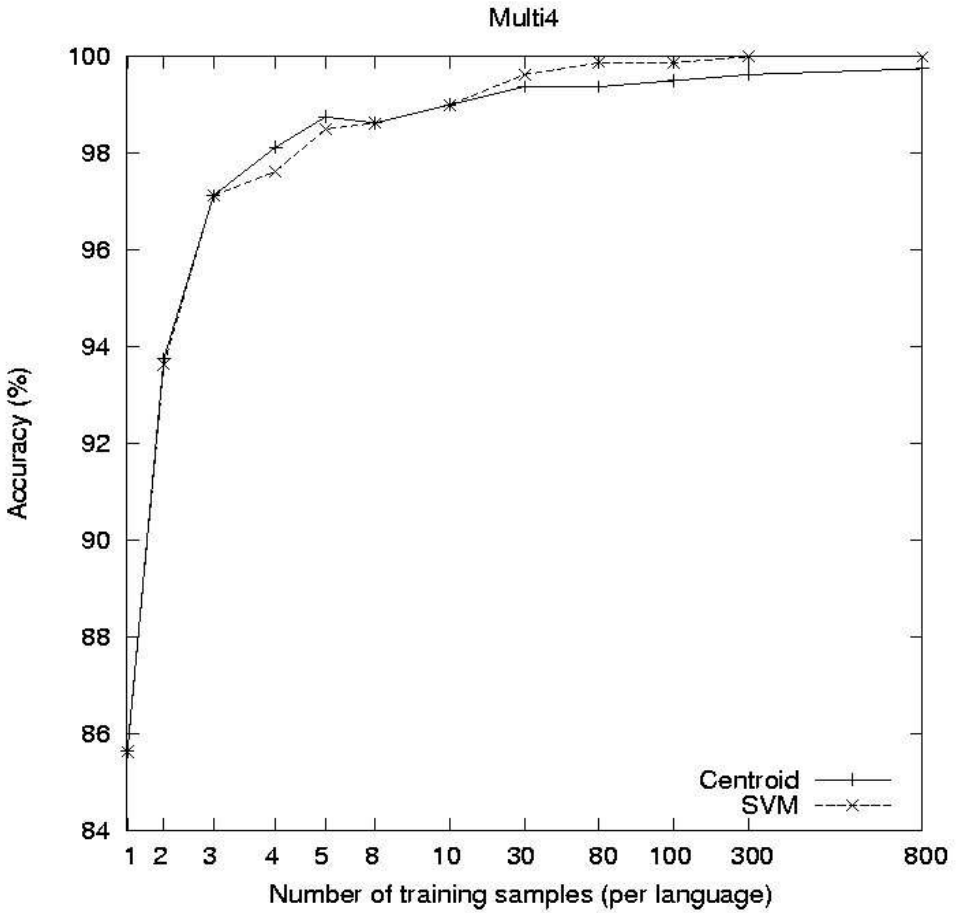


Language Identification --Training and Test Corpus--

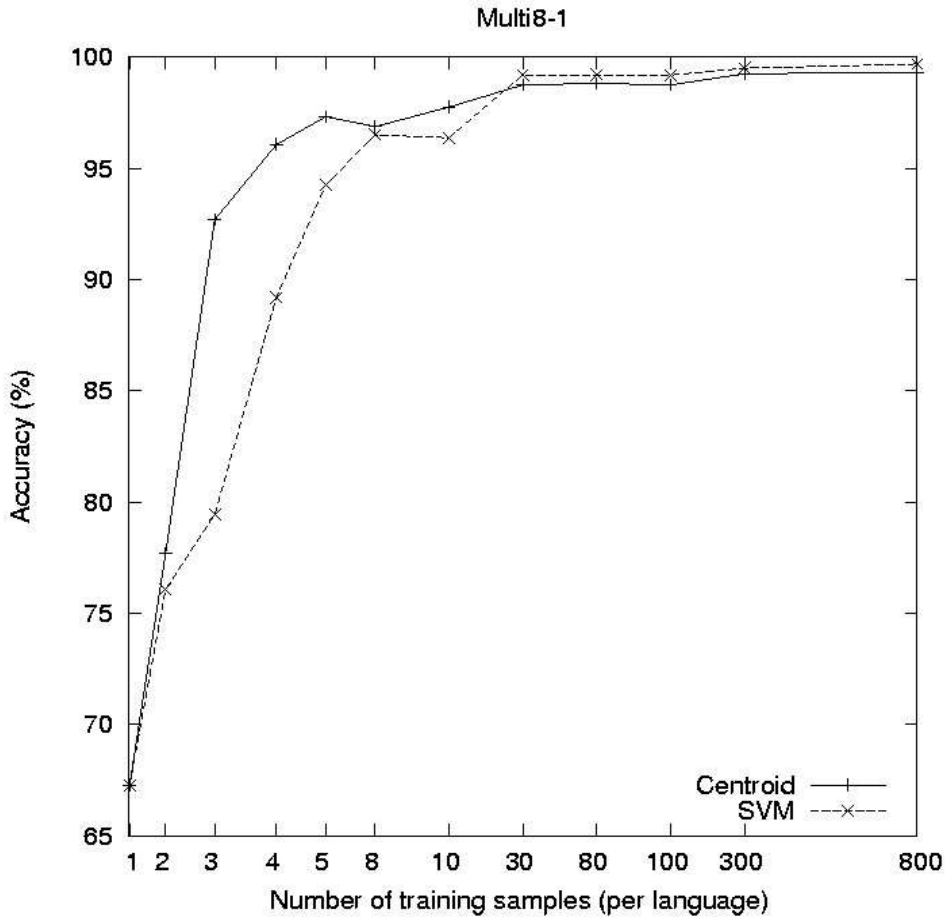
	Language	Encoding	Size(KB)		Language	Encoding	Size(KB)
1	English	ISO-8859-1	474	11	Croatian	Windows-1250	207
2	French	ISO-8859-1	421	12	Slovak	Windows-1250	214
3	Italian	ISO-8859-1	202	13	Slovenian	Windows-1250	212
4	Portuguese	ISO-8859-1	257	14	Bulgarian	Windows-1251	200
5	Spanish	ISO-8859-1	213	15	Russian	Windows-1251	213
6	Swedish	ISO-8859-1	213	16	Greek	ISO-8859-7	279
7	German	ISO-8859-1	206	17	Thai	TIS-620	210
8	Hungarian	Windows-1250	206	18	Chinese	Big5	201
9	Czech	Windows-1250	295	19	Japanese	EUC-JP	416
10	Polish	Windows-1250	218	20	Korean	EUC-KR	204

Evaluation

Asian Languages

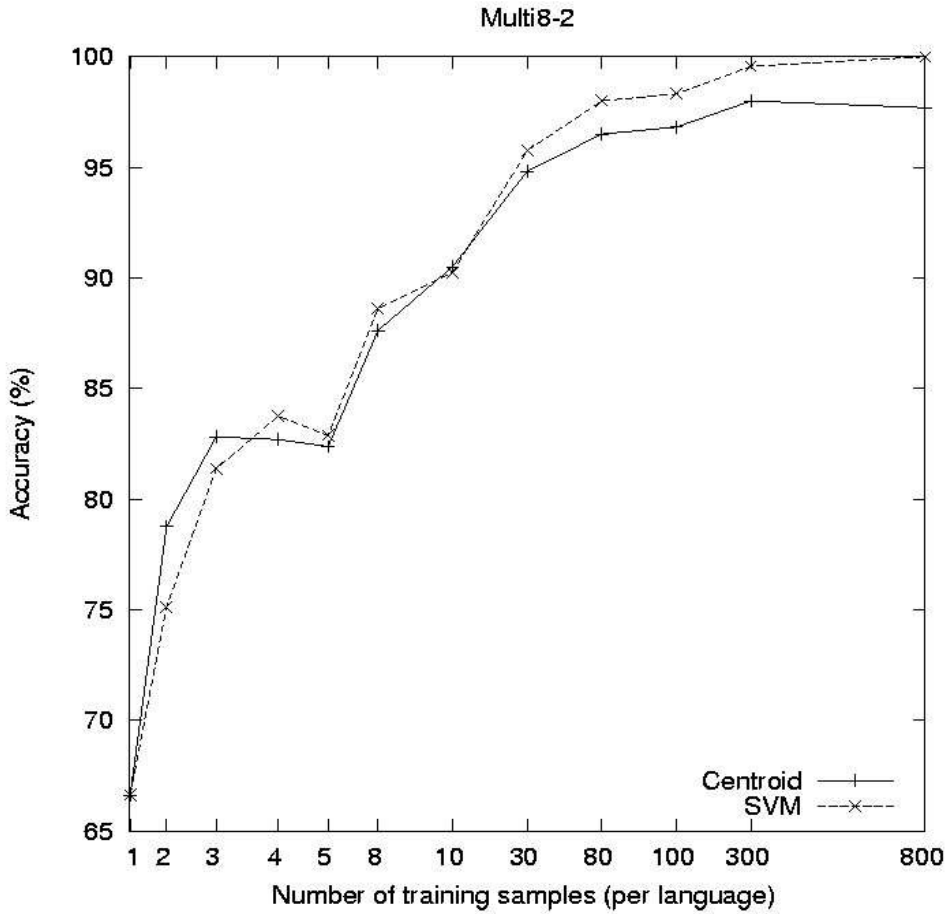


Roman Alphabet

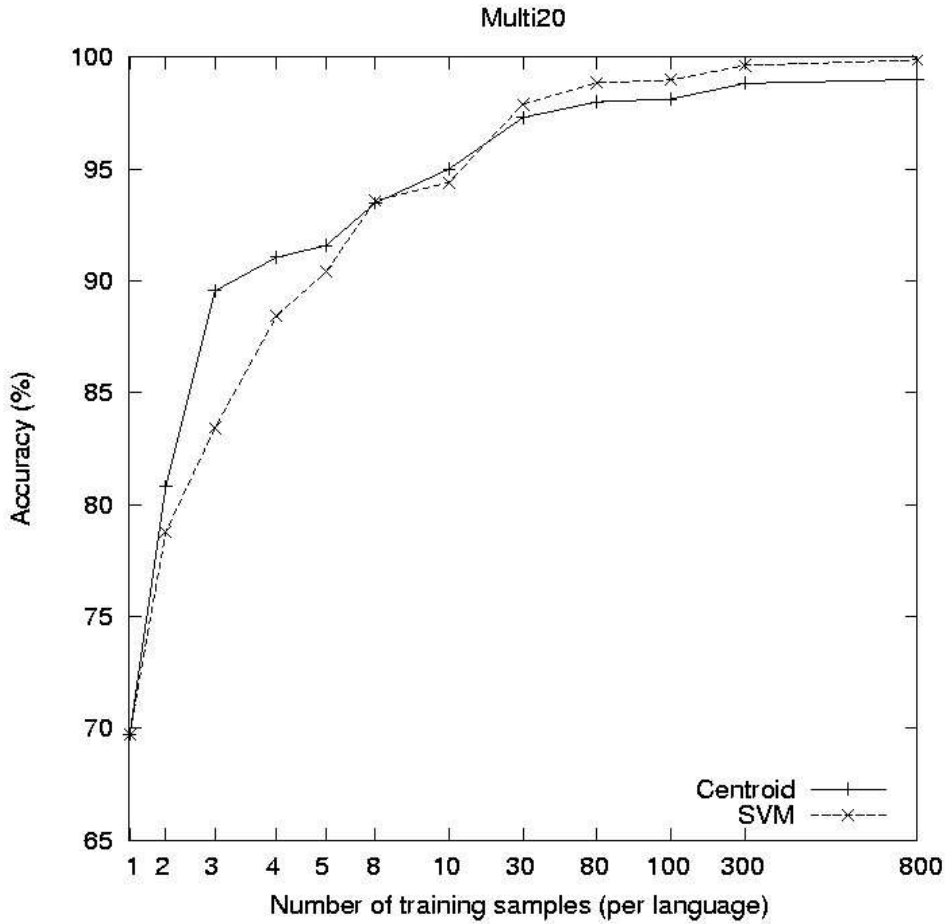


Evaluation

Slavic Family

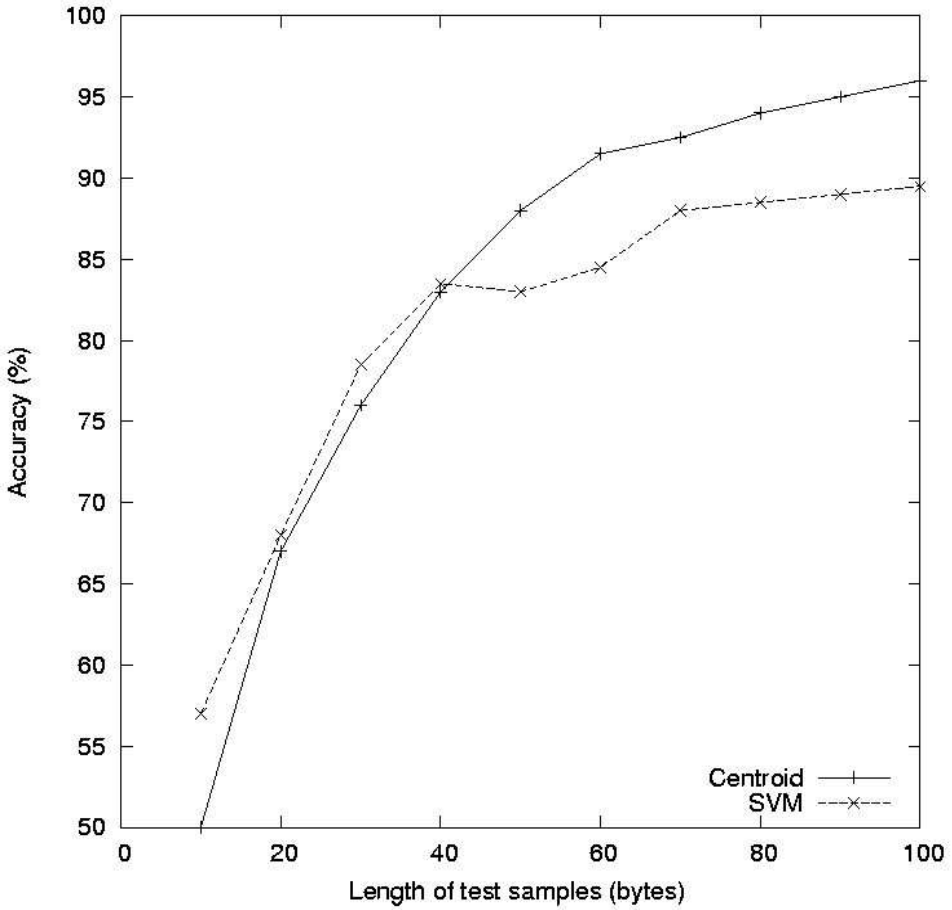


20 Languages

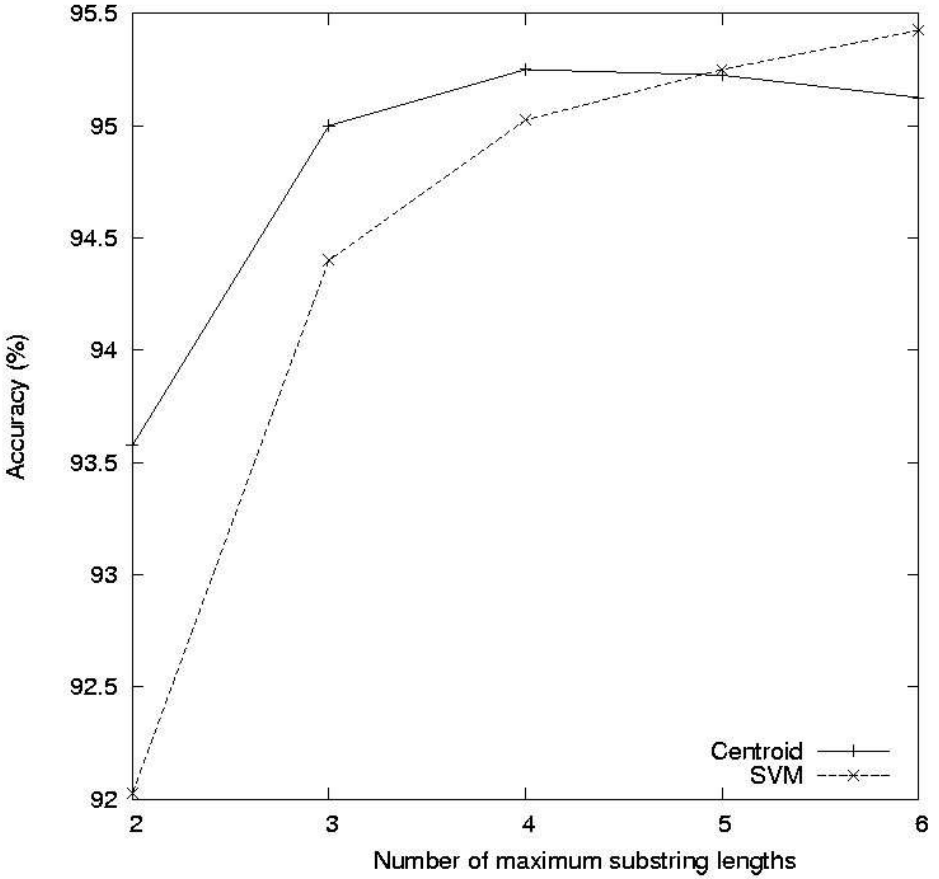


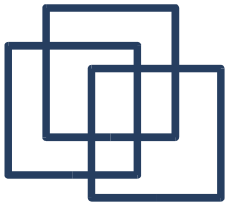
Evaluation

Length of Test Sample



Maximum Length of Substrings





Language Identification

- SVM shows its discrimimation power over Centroid based method under longer substrings, larger training set and test samples environment
- Both methods are good enough for discriminating the close language family
- Substrings can well represent the language in string kernels approach

TCL
Thai Computational Linguistics Laboratory

Langi: language identifier based on substring kernel

Result

Greek

Input string

Σύμφωνα με την ESPRESSO "σε περίεργο τροχαίο δυστύχημα είχε βρει τραγικό θάνατο ο αντεισαγγελέας του Αρείου Πάγου Ανδρέας Ιλούδας αφού είχε πρώτα παραδώσει στα χέρια του επίτιμου εισαγγελέα του ανωτάτου δικαστηρίου Γιώργου Πλαγιανάκου, πόρισμα με απίστευτα στοιχεία για αποφυλακίσεις κρατουμένων για υποθέσεις "νεφωτικών".

Select the identification method

centroid-based method | Identify | Reset

centroid-based method
support vector machines

Text samples including 20 languages

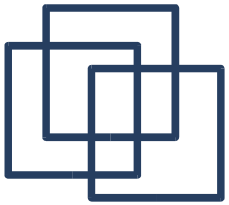
Bulgarian
Сигнал за домашната разправия е получен вчера около.

Chinese
新華社今天上午供本報特稿

Croatian
Najviši hrvatski dužnosnici uvjereni su da će do toga datuma to i učiniti.

Czech
Senát proto říká, že měl mít na posouzení novely víc času.

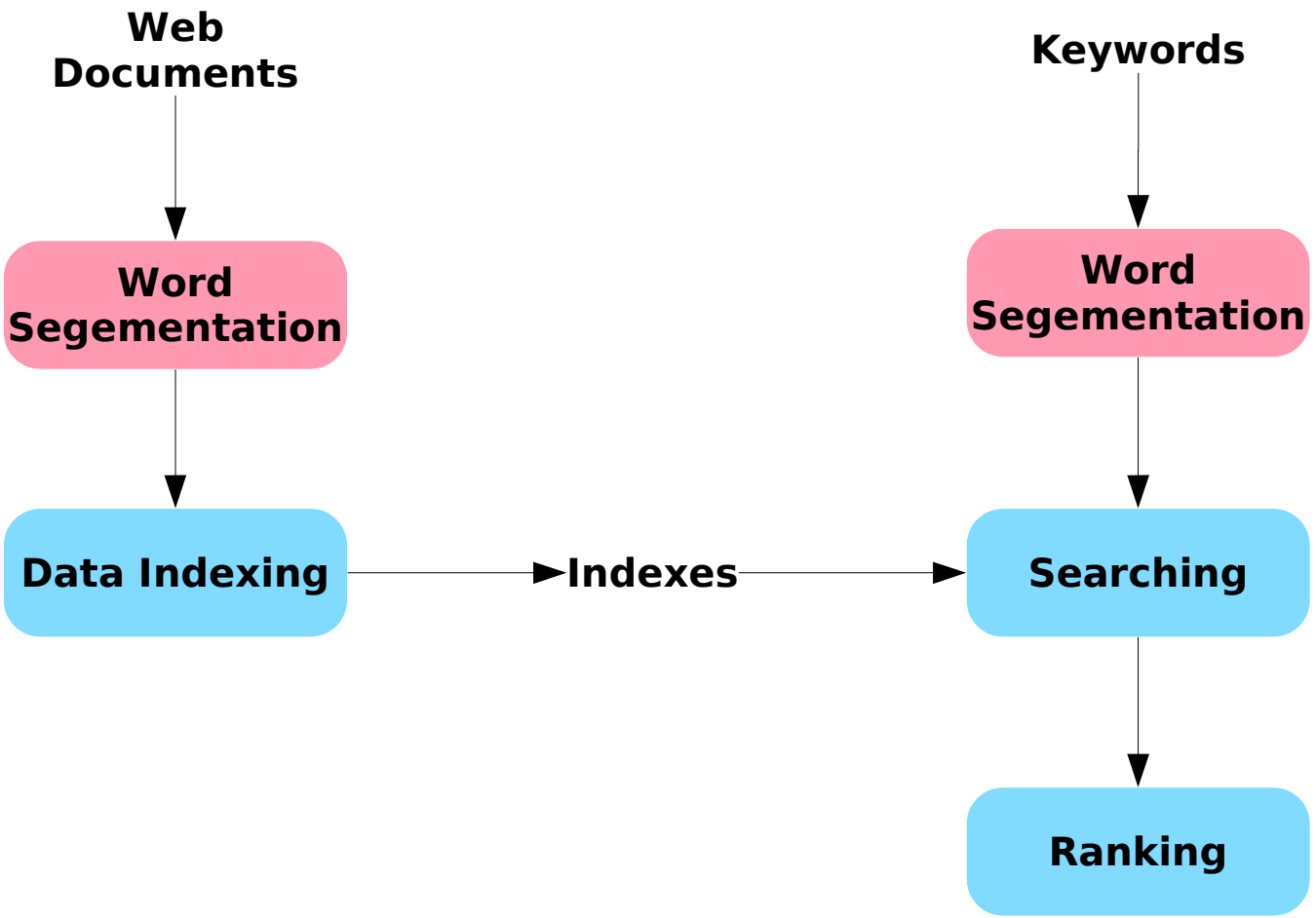
English



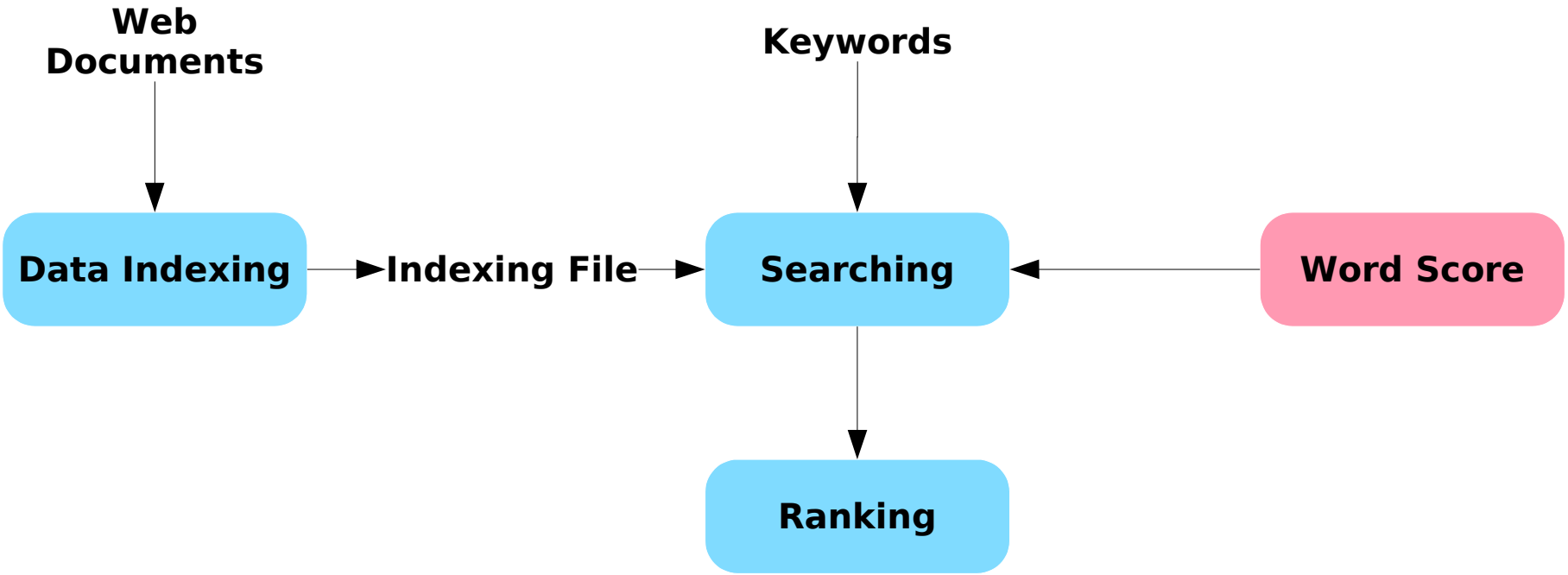
Dictionary-less Search Engine

- To overcome the limitation of vocabulary for making index
- To deal with the out-of-vocabulary problem
- To avoid the incomplete word segmentation result
- To avoid multiple search in case of phrase search
- To make it extensible for multi-lingual search

Dictionary-based Search Engine --Architecture--



Dictionary-less Search Engine --Architecture--



Dictionary-less Search Engine --Word Score--

$$MI_L(abxy) = \frac{p(abxy)}{p(ab)\cdot p(xy)}$$

$$MI_R(xycd) = \frac{p(xycd)}{p(xy)\cdot p(cd)}$$

$$wscore_L(xy|ab) = 1 - norm(MI_L(abxy))$$

$$wscore_R(xy|cd) = 1 - norm(MI_R(xycd))$$

Dictionary-based VS Dictionary-less --Evaluation--

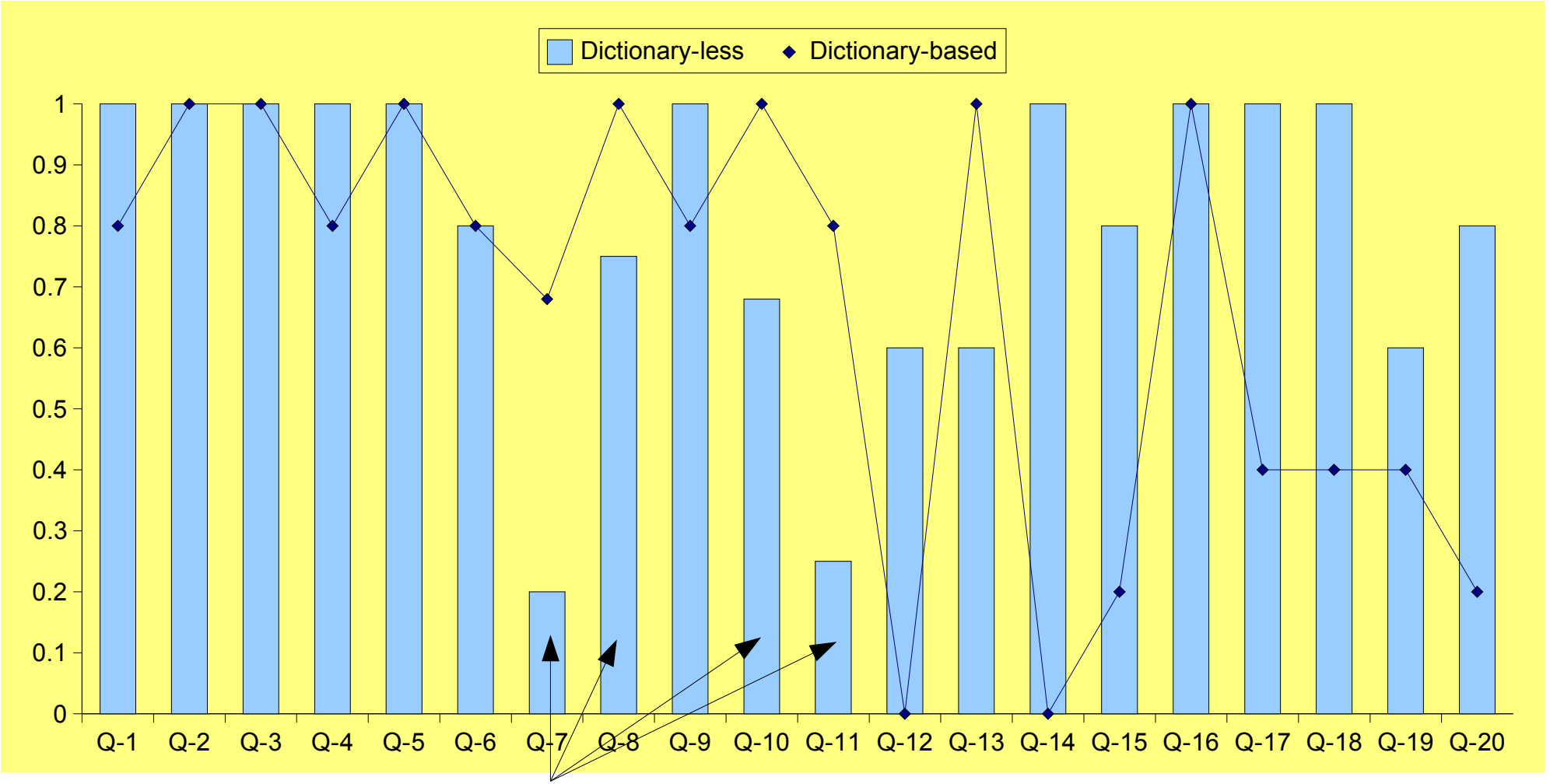
- Evaluate top 10 results of 20 queries by 5 evaluators
- Relevant if 3 out of 5 evaluators agree on each result
- Satisfaction on the result for each query is the evarage on the relevant

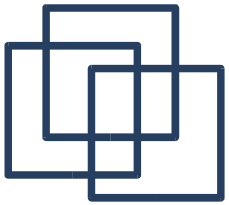
List of Queries

	Unsegmented queries	Segmented queries
1	บริจาค, สืนามิ	บริจาค, สืนามิ
2	เส้นตาย, ซัดดม	เส้นตาย, ซัดดม
3	มนัส, โชว์ตัว, จีน	มนัส, โชว์|ตัว, จีน
4	ผลกระทบ, ราคาน้ำมันแพง	ผลก|ระ|ทบ, ราคา|น้ำมัน|แพง
5	ไข้หวัดนก	ไข้หวัด|นก
6	ทุจริต, การเลือกตั้ง	ทุจริต, การ|เลือกตั้ง
7	จับกุม, ผู้ก่อการร้าย, ภาคใต้	จับกุม, ผู้ก่อการร้าย, ภาค|ใต้
8	นโยบาย, แก้ไข, ปัญหายาเสพติด	นโยบาย, แก้ไข, ปัญหา|ยา|เสพติด
9	ทดลอง, ลดค่าทางด่วน	ทดลอง, ลด|ค่า|ทางด่วน
10	ซื้อคืน, สัมปทาน, รถไฟฟ้า	ซื้อ|คืน, สัมปทาน, รถไฟฟ้า

	Unsegmented queries	Segmented queries
11	ลงทุน, ในพม่า	ลงทุน, ใน|พม่า
12	ส่งเสริม, การท่องเที่ยว, ไทย	ส่งเสริม, การ|ท่องเที่ยว, ไทย
13	เลือกตั้ง, ประธานาธิบดี, ปาเลสไตน์	เลือกตั้ง, ประธานาธิบดี, ปา|เลสไตน์
14	เลือกตั้ง, ผู้ว่า, กทม.	เลือกตั้ง, ผู้|ว่า, กทม|.
15	สินค้าไทย, การส่งออก	สินค้า|ไทย, การ|ส่ง|ออก
16	แปรรูปรัฐวิสาหกิจ	แปรรูป|รัฐวิสาหกิจ
17	อุ้ม, นายสมชาย	อุ้ม, นาย|สม|ชาย
18	ฉลองปีใหม่	ฉลอง|ปี|ใหม่
19	พรทิพย์, ลาออก	พร|ทิพย์, ลา|ออก
20	สวนสนุก	สวน|สนุก

Dictionary-based VS Dictionary-less --Evaluation--





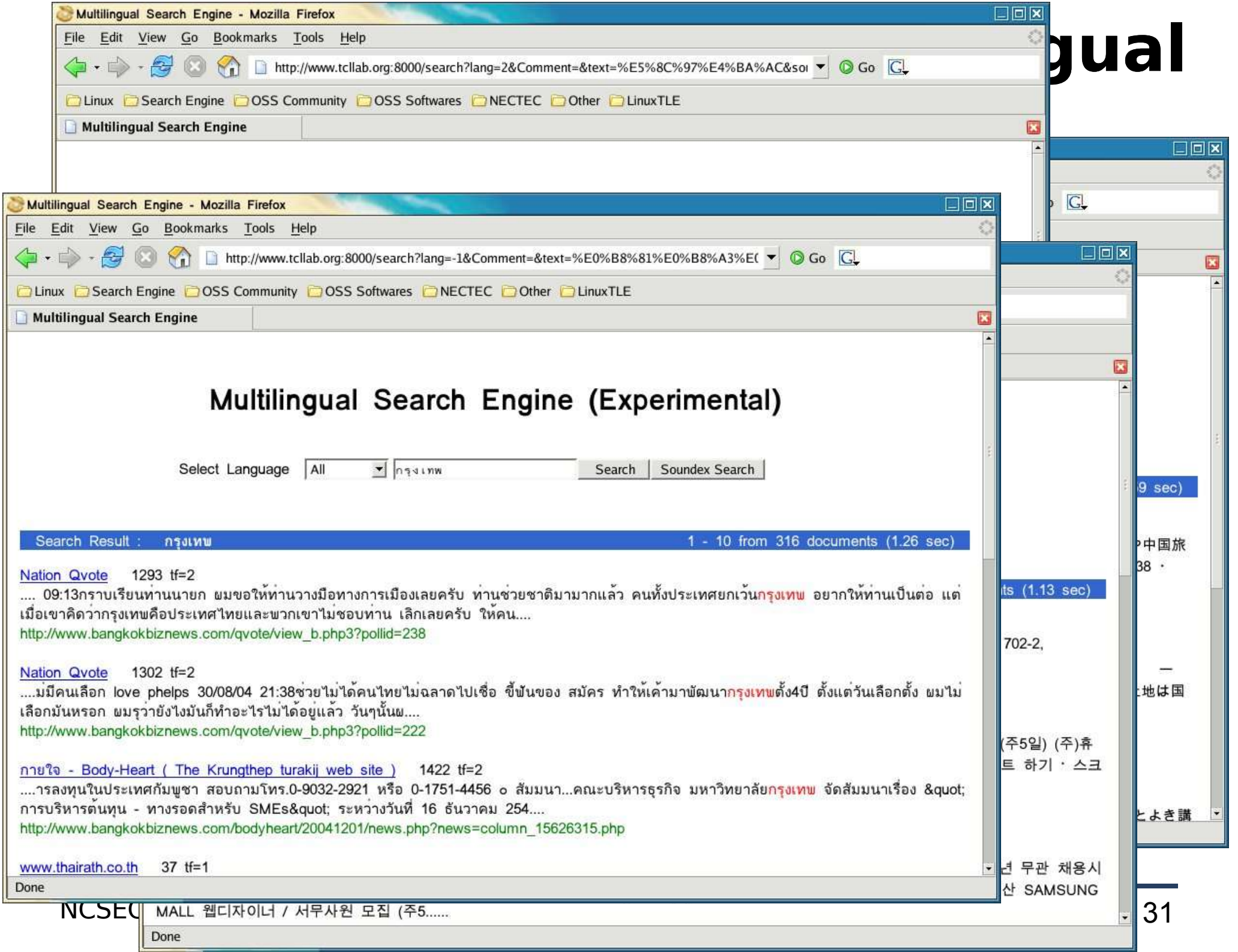
Evaluation

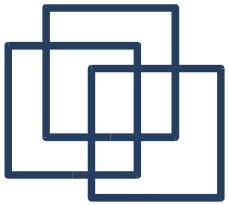
- Dictionary-based gets inferior results in case of excessive segmentation to be general words i.e. การส่งออก (การ|ส่ง|ออก), but not in the case of being non-general words i.e. ผลกระทบ (ผลก|ระ|ทบ)

Applying to Multi-Lingual Document Search

- 5853 articles from newspaper (65MB)

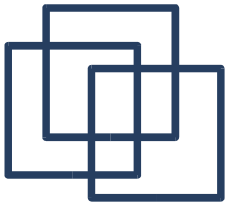
Language	Size (MB)
Thai	15.6
Chinese	28.1
Japanese	18.8
Korean	34.6
Total	**97.1**


Multilingual Search Engine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.tcllab.org:8000/search?lang=2&Comment=&text=%E5%8C%97%E4%BA%AC&so
Go
Linux Search Engine OSS Community OSS Softwares NECTEC Other LinuxTLE
Multilingual Search Engine
Multilingual Search Engine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.tcllab.org:8000/search?lang=-1&Comment=&text=%E0%B8%81%E0%B8%A3%E
Go
Linux Search Engine OSS Community OSS Softwares NECTEC Other LinuxTLE
Multilingual Search Engine
Multilingual Search Engine (Experimental)
Select Language All กรุงเทพ Search Soundex Search
Search Result : กรุงเทพ 1 - 10 from 316 documents (1.26 sec)
Nation Qvote 1293 tf=2
.... 09:13กราบเรียนท่านนายก ผมขอให้ท่านวางมือทางการเมืองเลยครับ ท่านช่วยชาติมามากแล้ว คนทั้งประเทศยกเว้นกรุงเทพ อยากให้ท่านเป็นต่อ แต่เมื่อเขาคิดว่ากรุงเทพคือประเทศไทยและพวกเขาไม่ชอบท่าน เลิกเลยครับ ให้คน....
http://www.bangkokbiznews.com/qvote/view_b.php3?pollid=238
Nation Qvote 1302 tf=2
....มีคนเลือก love phelps 30/08/04 21:38ช่วยไม่ได้คนไทยไม่ฉลาดไปเชื่อ ขี้ฟันของ สมัคร ทำให้เค้ามาพัฒนากรุงเทพตั้ง4ปี ตั้งแต่วันเลือกตั้ง ผมไม่เลือกมันหรอก ผมรู้ว่ายังไงมันก็ทำอะไรไม่ได้อยู่แล้ว วันๆนั้นผ....
http://www.bangkokbiznews.com/qvote/view_b.php3?pollid=222
กายใจ - Body-Heart (The Krungthep turakij web site) 1422 tf=2
....ารลงทุนในประเทศกัมพูชา สอบถามโทร.0-9032-2921 หรือ 0-1751-4456 o สัมมนา...คณะบริหารธุรกิจ มหาวิทยาลัยกรุงเทพ จัดสัมมนาเรื่อง " การบริหารต้นทุน - ทางรอดสำหรับ SMEs" ระหว่างวันที่ 16 ธันวาคม 254....
http://www.bangkokbiznews.com/bodyheart/20041201/news.php?news=column_15626315.php
www.thairath.co.th 37 tf=1
Done
MALL 웹디자이너 / 서무사원 모집 (주5.....
Done



Concluding Remark

- Language independent consideration is required for multi-lingual text precessing
- Individual language has a unique bit sequence
- A term is a frequent use of a string



Credits

- Prapass Srichaivattana
  - Dictionary-less search engine
- Canasai Kruengkrai
  - Language identification
  - Term-based ontology alignment
- Shisanu Tongchim
  - Dictionary-less search engine
- Thatsanee Charoenporn
  - Term-based ontology alignment