



GOVERNMENT
USERS CONFERENCE

JUNE 8-9
2010
CHANTILLY, VA

Thai, the Tiger of Text Analysis: An Introduction to Thai Text Processing Issues

Rattima Nitisaroj
Linguist
Basis Technology Corp.



Human Language Technology From Arabia to Afghanistan

English vs. Thai

The next morning, as soon as the sun was up, they started on their way, and soon saw a beautiful green glow in the sky just before them.

เช้าต่อมา ทันทีที่ที่ตะวันขึ้นพวกเขา ก็ออกเดินทางกันและใน
ไม่ช้าก็เห็นแสงสุกเขียวดวงมทาบท้องฟ้าอยู่เบื้องหน้า

Why Word Segmentation?

- Words are required in every area of text and speech processing
 - Information Retrieval
 - Machine Translation
 - Text-to-Speech Synthesis
 - Automatic Speech Recognition

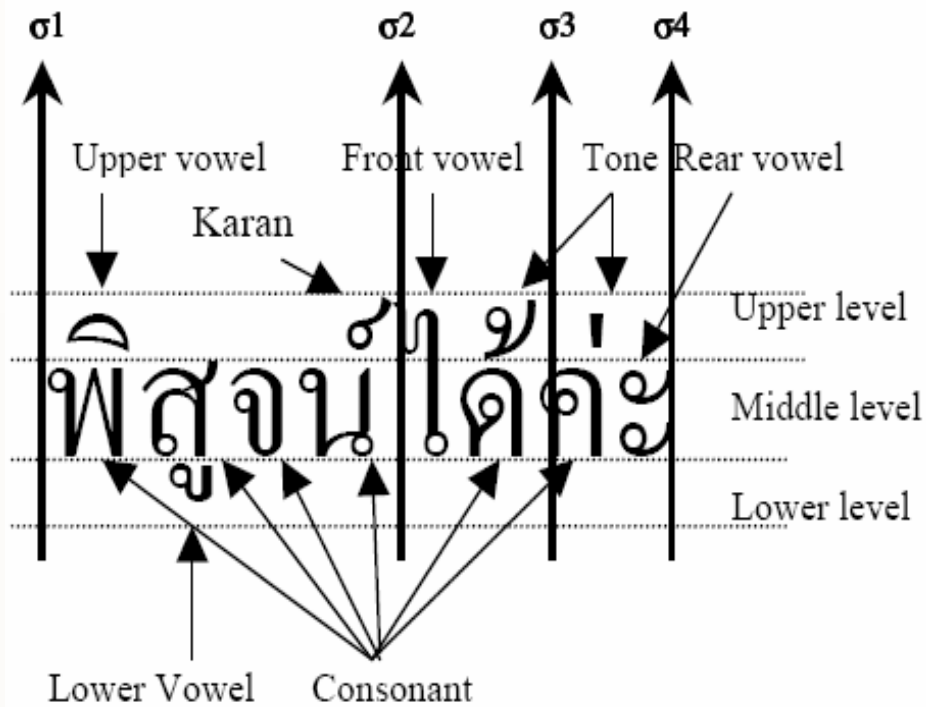
Why Syllable Segmentation?

- A syllable is easier to define and more consistent in analysis than a word
- Important for determining how a character is pronounced, benefiting text-to-speech and automatic Romanization

Thai Characters

- Thai is an alphabetical language
- 44 characters for 21 consonant sounds
- 19 characters for 24 vowel sounds (18 single vowels and 6 diphthongs)
- 4 characters for tone markers (5 tones)
- Special symbols and numbers

Thai Characters



(Theeramunkong et al. 2000, Figure 1)

Letter-to-Sound Relationship

- A character represents different sounds depending on the syllable position

ราว /ra:w/

การ /ka:n/

ทาน /tha:n/

บาท /bà:t/

Letter-to-Sound Relationship

- However, the syllable position does not guarantee the same pronunciation

บัณ**ฑิต** /ban-**dìt**/

มณ**ฑา** /mon-**thà:**/

เท /**thè:**/

เทา /**thau**/

Letter-to-Sound Relationship

- A sequence of consonants can represent
 - a single sound
ทราย /sa:y/
 - a consonant cluster sound
จันทร /can-thra:/
 - or map to one leading syllable and an initial consonant sound
กันทร /kan-tha-ra:-ko:n/

Letter-to-Sound Relationship

- A vowel sound may not be represented by any vowel character

ปน /pon/

กรรณ /kan/

Letter-to-Sound Relationship

- When the special character Karan is placed above a character, it sometimes indicates that one or more preceding character is not pronounced.

นั~~์~~

พั~~์~~กต~~์~~

พระ~~์~~ลักษ~~์~~มณ~~์~~

Letter-to-Sound Relationship

- Homographs: the pronunciation varies according to the semantic context

เพลลา /phlau/ ‘axle’

 /phe:-la:/ ‘time’

สระ /sà/ ‘pool’

 /sà-rà/ ‘vowel’

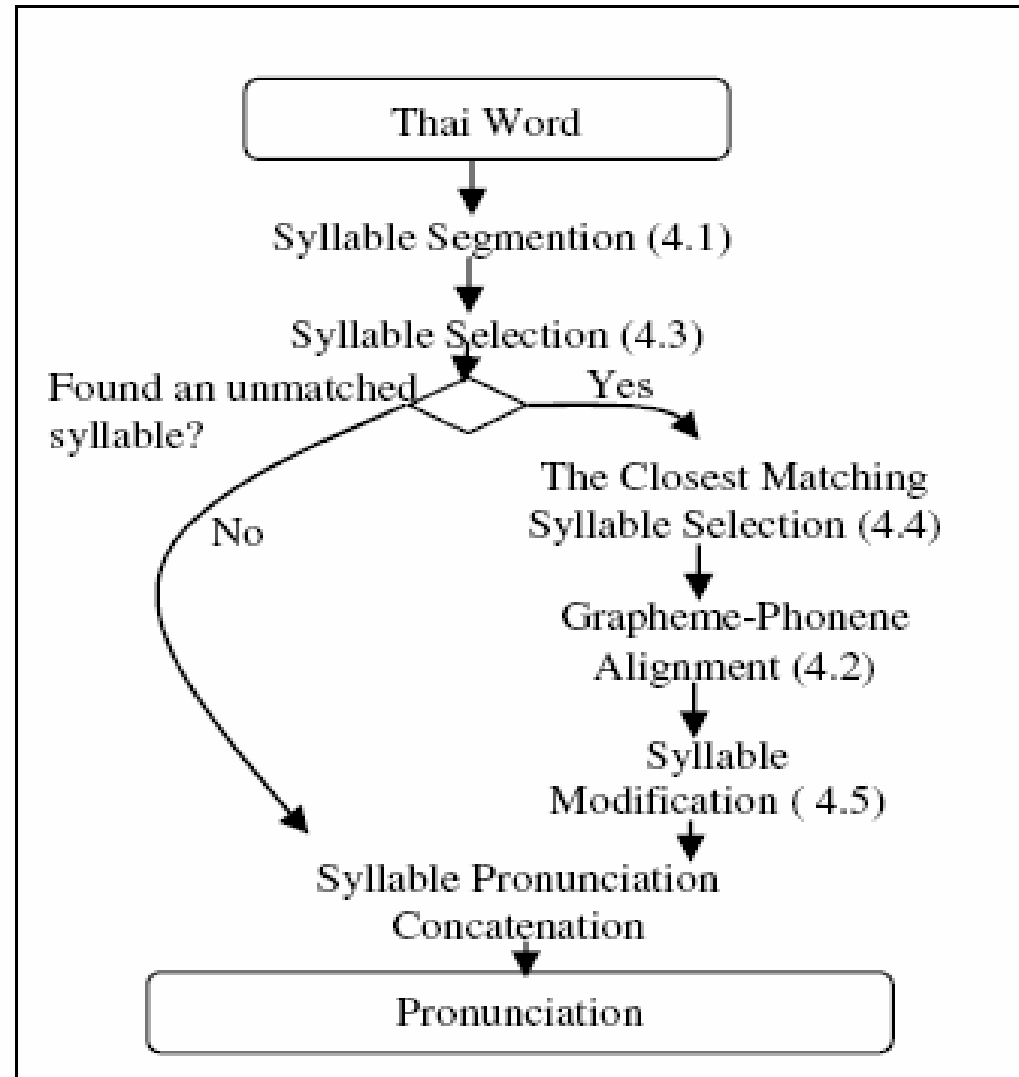
Letter-to-Sound Conversion

- Aroonmanakun & Rivepiboon (2004)
 - Segment a character string into syllables using a trigram model of syllables
 - Generate all possible pronunciations for each syllable
 - Choose the pronunciation of each syllable based on the results of word segmentation and the statistical information of pronunciation

Letter-to-Sound Conversion

- Example-based

(Charoenpornasawat &
Schultz 2006, figure 2)



Word Segmentation



What is a word?

- Orthography-based
 - ice cream vs. ice-cream
- Concept-based

proceedings

หนังสือรวมบทความทางวิชาการในการประชุมสัมมนา

What is a word?

- Orthography-based
 - ice cream vs. ice-cream
- Concept-based

proceedings

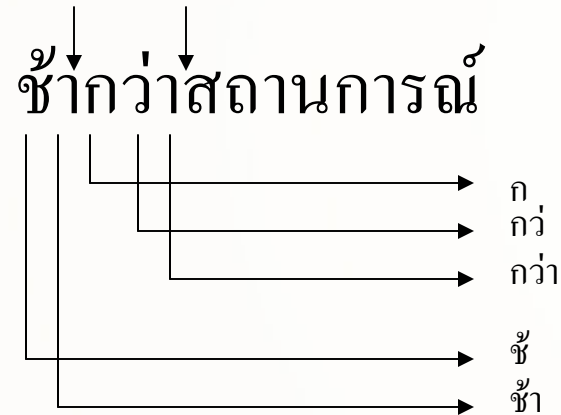
หนังสือรวมบทความทางวิชาการในการประชุมสัมมนา
book collect article about academic in meeting seminar

- Degree of agreement among five speakers in segmenting 3,070 syllables into words: 0.75 (Aroonmanakun 2002)

Thai Word Segmentation

- Dictionary-based
 - Longest Matching (Poowarawan 1986)
 - Maximal Matching (Sornlertlamvanich 1993)
- Corpus-based
 - Context words and POS tags (Meknavin & Charoenpornasawat 1997, Charoenpornasawat 1999)
 - Syllable-based trigram model and maximum collocation strength (Aroonmanakun 2002)
 - Character types and character-based n -gram model (Haruechaiyasak et al. 2008)

Dictionary-based



ไปห้ามเหสี

ไปห้ามเหสี

go carry deviate color

ไปห้ามเหสี

go see queen

Segmentation Ambiguity

- Context independent

ขอมอบ

ขอม**อบ**

ข**อม**อบ

Khmer **bake**

beg give

ชนก**ลุ่ม**

ชนก**ลุ่ม**

ชนก**ลุ่ม**

father **low**

people group

Segmentation Ambiguity

- Context dependent

ดวงตากกลมโต

eye round big

นั่งตากลมทะเล

expose wind sea

Corpus-based

- Meknavin & Charoenpornswat (1997), Charoenpornswat (1999)
 - Presence of a particular word within +/- 10 words ex. if ตากลม has แปลว่ ‘glittering’ in context, the string is segmented as ตากลม ‘round eyes’
 - Collocations – test for a pattern of up to 2 contiguous words and/or POS tags around target ex. if มากกว่า *number measurement*, the result is มากกว่า

Corpus-based

- Aroonmanakun (2002)
 - Word segmentation as a process of segmenting syllables (using a trigram model) and merging syllables (based on collocation strength)
 - Approximately 200 syllable patterns are defined

Corpus-based

- Haruechaiyasak et al. (2008)
 - Training corpus contains characters tagged with information on character type and whether they are word-initial or intra-word characters
 - n -gram model of characters preceding and following word boundary. n varies from 3 to 11. Results improve with larger n .

Result Comparison

Approach	Algorithm	Precision	Recall	F1
Dictionary-based	LM-Lexitron	88.21	86.91	87.55
	LM-Domain	95.20	88.55	91.75
	MM-Lexitron	88.34	87.39	87.86
	MM-Domain	95.27	88.92	91.98
Machine-learning-based	Naïve Bayes	69.70	60.60	64.90
	Decision Tree	80.10	75.10	77.50
	Support Vector Machine	92.87	88.71	90.74
	Conditional Random Field	95.79	94.98	95.38

LM = Lexical Matching, MM = Maximal Matching

Search Problems



Word Boundary

Query: ข้า 'galangal'

Results:

ข้า เป็นชื่อชาวเขาเผ่าหนึ่ง 'Akha'

ข้า สารหน้ารู้ 'news'

Incorrect or Inconsistent Transliteration

Internet

อินเทอร์เน็ต

อินเทอร์เน็ต

อินเทอร์เน็ต

อินเทอร์เน็ต

อินเทอร์เน็ต

Variations in Romanization

- Romanization based on orthographic form, pronunciation, or both

วุฒิสักดิ์

Wutthisak

Vuttisak

Vudhisakdi

Wuthisak

Wutisuk

Wudhisak

Vudhisukdi

Wutisak

- Can one of the transliterated or romanized variants be used in a single search which covers all variants at once?

Compounds

จุดโคจรใกล้สุดจากขั้วโลก 'perigee'

point orbit near most from pole earth

สลากกินแบ่งรัฐบาล 'lottery'

lots eat distribute government

รถโดยสารประจำทาง 'bus'

car take regular way

Compounds

จุดโคจรใกล้สุดจากขั้วโลก 'perigee'

point orbit near most from pole earth

สลากกินแบ่งรัฐบาล 'lottery'

lots eat distribute government

รถโดยสารประจำทาง 'bus'

car take regular way

Compounds

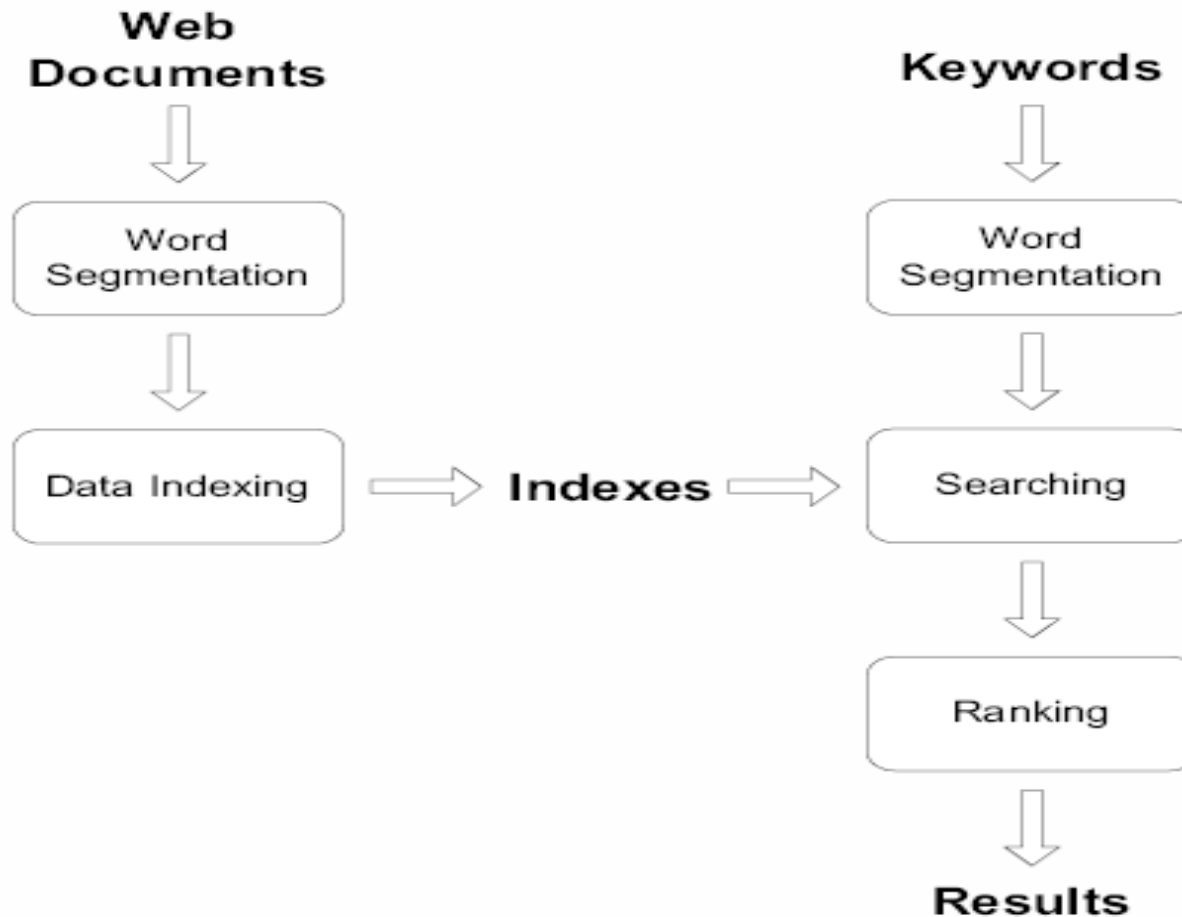
จุดโคจรใกล้สุดจากขั้วโลก ‘perigee’
point orbit near most from pole earth

สลากกินแบ่งรัฐบาล ‘lottery’
lots eat distribute government

รถโดยสารประจำทาง ‘bus’
car take regular way

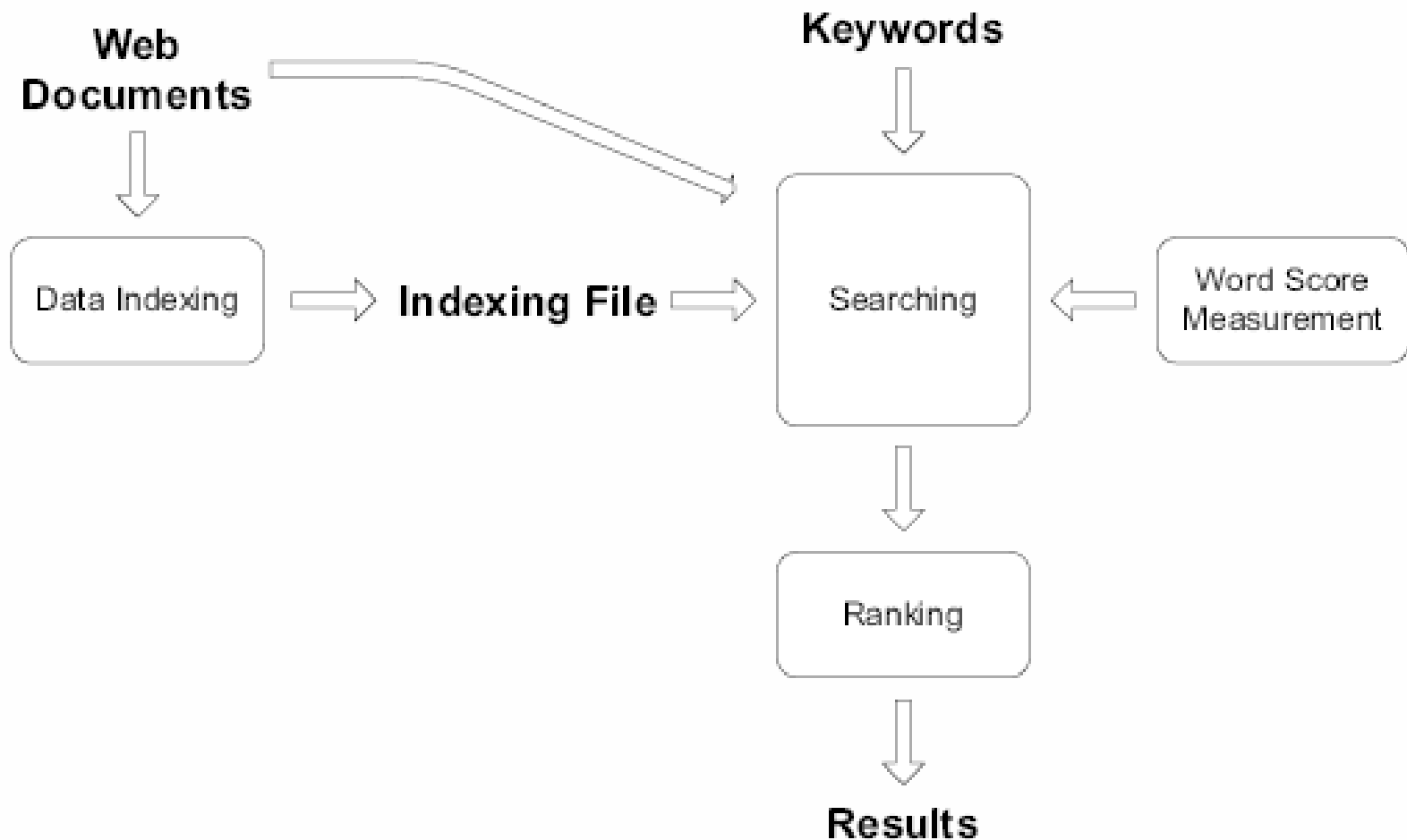
Thai Search Methods

- Dictionary-based



Thai Search Methods

- Suffix array-based



Suffix Array Construction

Text	a	b	r	a	c	a	d	a	b	r	a
Index	0	1	2	3	4	5	6	7	8	9	10



Suffix	Index
a b r a c a d a b r a	0
b r a c a d a b r a	1
r a c a d a b r a	2
a c a d a b r a	3
c a d a b r a	4
a d a b r a	5
d a b r a	6
a b r a	7
b r a	8
r a	9
a	10

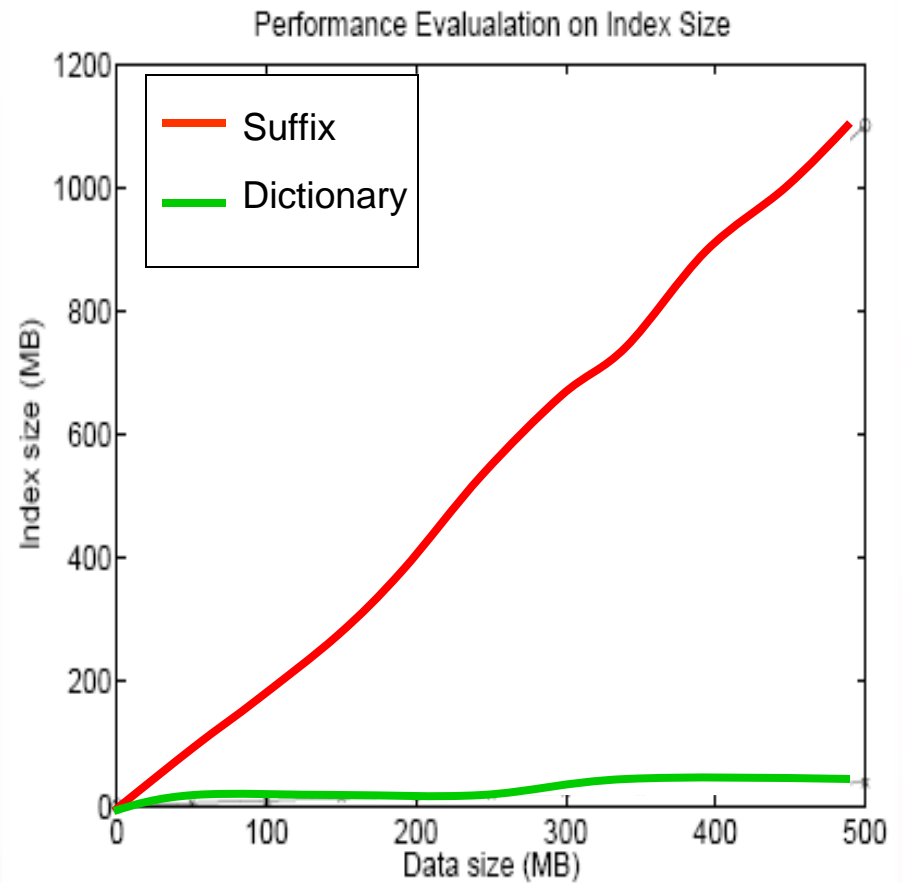
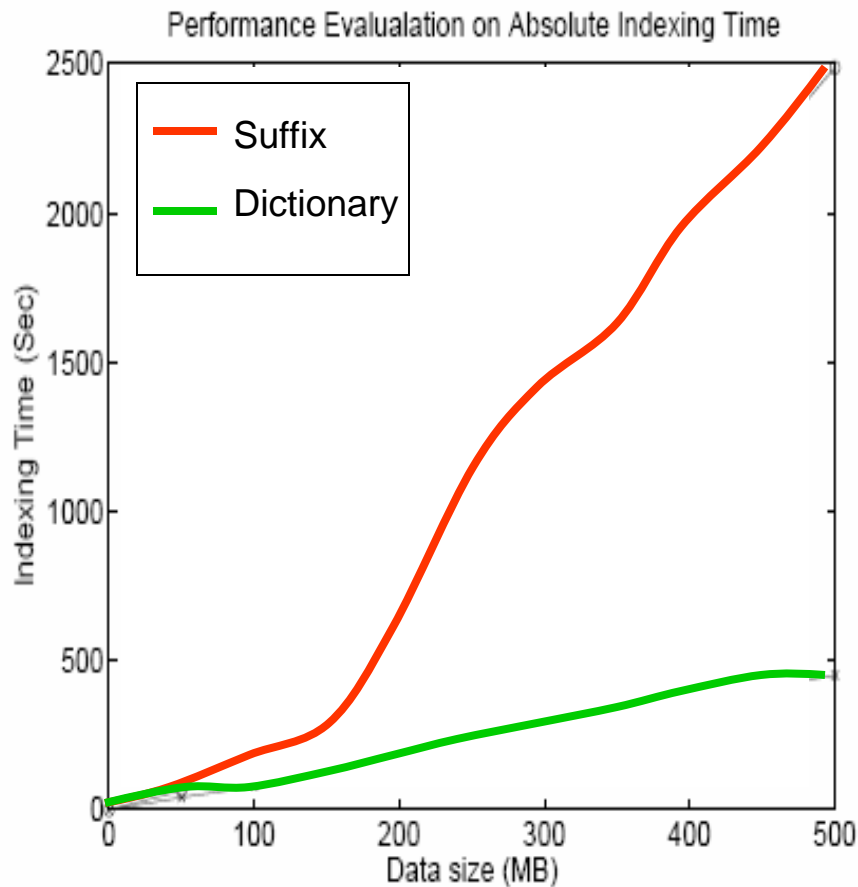


Sorted Suffix	Index
a	10
a b r a	7
a b r a c a d a b r a	0
a c a d a b r a	3
a d a b r a	5
b r a	8
b r a c a d a b r a	1
c a d a b r a	4
d a b r a	6
r a	9
r a c a d a b r a	2



10	7	0	3	5	8	1	4	6	9	2
----	---	---	---	---	---	---	---	---	---	---

Thai Search Methods



(Haruechaiyasak et al. 2006, figures 3 & 4)

Sentence Segmentation



Sentence Segmentation

สามิต 14 ว่า "นั่นแหละ กระสุนหัวเปราะต่างจากกระสุน
ธรรมดาตรงที่มันไม่ได้ทำด้วยตะกั่วหุ้มด้วยปลอก
ทองแดง แต่ทำมาจากวัสดุผสมหลายอย่าง บางอย่างอัด
ด้วยแรงมหาศาล บางอย่างอัดด้วยกาว หัวเป็นลูกปราชญ์หุ้ม
ด้วยเรซิน คิดค้นมาตั้งแต่ปี 2517 มีกระสุนหลาย
ขนาด ตั้งแต่ .25 ถึง .45 กระสุนแบบนี้จะว่าอันตรายก็
อันตราย แต่จะว่าปลอดภัยก็ปลอดภัย"

(Win Leowarin's *Kattakam Jakrasi*)

Sentence Segmentation

สามิต 14 ว่า "นั่นแหละ กระสุนหัวเราะต่างจากกระสุน
ธรรมดาตรงที่มันไม่ได้ทำด้วยตะกั่วหุ้มด้วยปลอก
ทองแดง↓ แต่ทำมาจากวัสดุผสมหลายอย่าง บางอย่างอัด
ด้วยแรงมหาศาล บางอย่างอัดด้วยกาว หัวเป็นลูกปราชญ์หุ้ม
ด้วยเรซิน คัดค้นมาตั้งแต่ปี↓ 2517 มีกระสุนหลาย
ขนาด ตั้งแต่ .25 ถึง .45 กระสุนแบบนี้จะว่าอันตรายก็
อันตราย แต่จะว่าปลอดภัยก็ปลอดภัย"

(Win Leowarin's *Kattakam Jakrasi*)

Use of Space in Thai Writing

- Between sentences
- Between phrases or clauses
- Before and after numerals
- Between list elements
- Between first and last names
- Before and/or after some special symbols and punctuation marks
- Only 30% of all space use in Thai is for sentence breaking (Charoenpornasawat & Sornlertlamvanich 2001)

Some Cues to Distinguish Space

- A space after a final particle is a sentence boundary. Examples of final particles: ไหม (question particle), ครับ (polite, male),ค่ะ (polite, female)
- A space after a discourse marker is not a sentence boundary. Examples of discourse markers: ดังนั้น 'therefore', เพราะฉะนั้น 'as a result', ต่อไปนี้ 'next'

Automatic Sentence Segmentation

- Classification of space into sentence breaking and non-sentence breaking space as trigram POS tagging problem (Mittrapiyanurak & Sornlertlamvanich 2000)
- Feature-based classification (Charoenpornasawat & Sornlertlamvanich 2001)
 - Number of words to the left and to the right
 - 2 words and POS tags before and after target space

Automatic Sentence Segmentation

- Results

	Trigram (%)	Feature-based (%)
Correct break	75.97	77.27
False break	8.13	1.74

References

- Aroonmanakun, W. 2002. Collocation and Thai word segmentation. In *Proceedings of the 5th Symposium on Natural Language Processing*, Hua Hin, Thailand, pp. 68-75.
- Aroonmanakun, W., and W. Rivepiboon. 2004. A unified model of Thai romanization and word segmentation. In *Proceedings of PACLIC 18*, Tokyo, Japan, pp. 205-124.
- Charoenpornasawat, P. 1999. *Feature-based Thai Word Segmentation*, Master Thesis, Chulalongkorn University.
- Charoenpornasawat, P., and T. Schultz. 2006. Example-based grapheme-to-phoneme conversion for Thai. In *Proceedings of Interspeech 2006*, Pittsburgh, PA, pp. 1268-1271.
- Charoenpornasawat, P., and V. Sornlertlamvanich. 2001. Automatic sentence break disambiguation for Thai. In *Proceedings of ICCPOL2001*, Korea, pp. 231-235.
- Haruechaiyasak, C., C. Damrongrat, C. Sangkeettrakarn, S. Kongyong, and N. Angkawattanawit. 2006. Sansarn Look!: A platform for developing Thai-language information retrieval systems. In *Proceedings of ITC-CSCC 2006*.
- Haruechaiyasak, C., S. Kongyong, and M. Dailey. 2008. A comparative study on Thai word segmentation approaches. In *Proceedings of ECTI-CON2008*, pp. 125-128.

References

- Meknavin, S., P. Charoenpornswat, and B. Kijirikul. 1997. Feature-based Thai word segmentation. In *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, Phuket, Thailand.
- Mittrapiyanurak, P., and V. Sornlertlamvanich. 2000. The automatic Thai sentence extraction. In *Proceedings of the 4th Symposium on Natural Language Processing*, Chiang Mai, Thailand, pp. 23-28.
- Poowarawan, Y. 1986. Dictionary-based Thai syllable separation. In *Proceedings of the 9th Electronics Engineering Conference*.
- Sornlertlamvanich, V. 1993. Word segmentation for Thai in machine translation system, *Machine Translation*, NECTEC, Bangkok, Thailand, pp. 50-56.
- Theeramunkong, T., V. Sornlertlamvanich, T. Tanhermhong, and W. Chinnan. 2000. Character cluster based Thai information retrieval. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, Hong Kong, pp. 75-80.
- Tongchim, S., P. Srichaivattana, V. Sornlertlamvanich, and H. Isahara. 2006. Blind evaluation for Thai search engines. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, May 24-26.