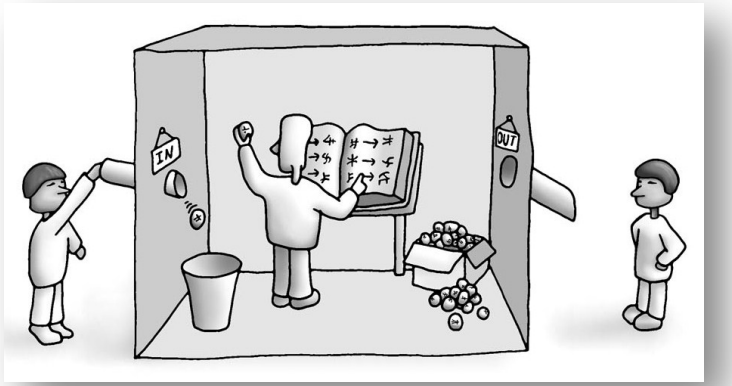
Thai Text Processing and Its Applications

Virach Sornlertlamvanich
Sirindhorn International Institute of Technology (SIIT),
Thammasat University, Thailand
virach@siit.tu.ac.th

NLP Challenges

- **Natural Language Processing** is no more for the **language study** only.

Turing Test as a tool to understand whether "**machines can think**".



NLP@Sapienza

AlphaGo

AlphaGo's algorithm uses a combination of neural networks, machine learning and Monte Carlo tree search techniques, combined with extensive training, both from human and computer play.

Searle's Chinese Room argument 1980 in "Minds, Brains, and Programs"
Accurate NLP: machine translation, summarization, machine reading, question answering, information retrieval, social text understanding, opinion mining, etc.

NLP Challenges

- **Internet, Big Data, Machine Learning, Deep Learning** have brought along the possibilities.

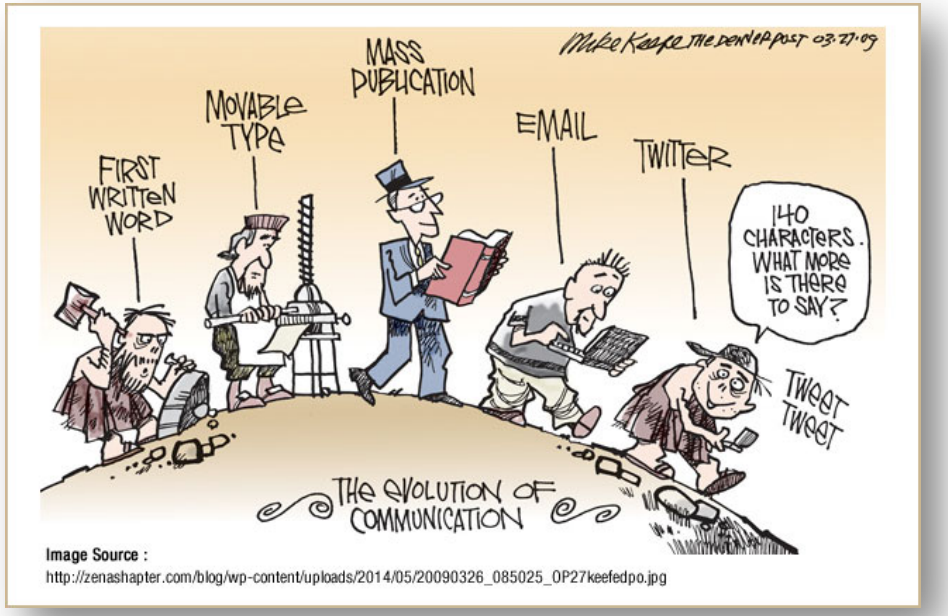


Image Source :
http://zenashapter.com/blog/wp-content/uploads/2014/05/20090326_085025_0P27keefedpo.jpg

Facebook:-

Adds 0.5 petabyte (10^{15}) of data every 24 hours

Twitter:-

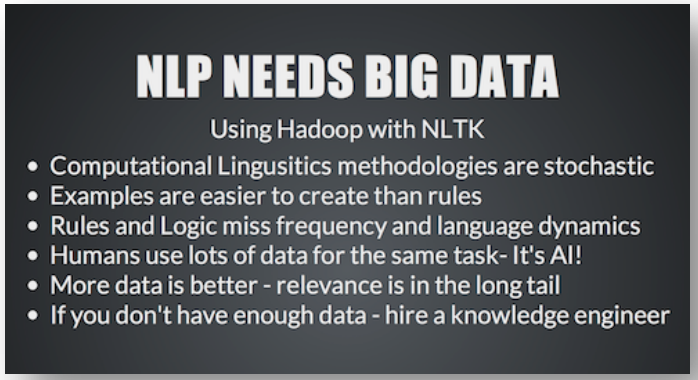
Adds 340 million tweets per day

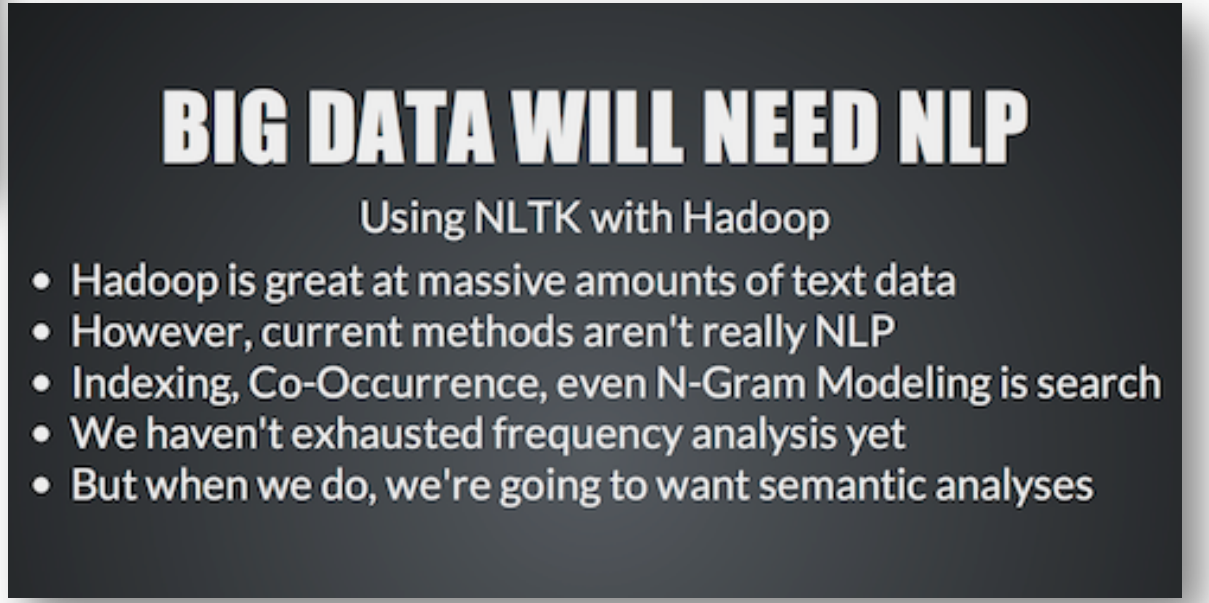
Youtube:-

Adds 100 hours of new videos every minute

Germin8, Social Intelligence

NLP Challenges





Bird Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

Data Community DC (DC2)

Three NLP Fundamental Problems

- Word Segmentation/Tokenization
- Named Entity Recognition/Keyword Extraction
- Semantic Relation Extraction

Three NLP Constructive Applications

- Linked Data → Knowledge Map
- Keyword Tracking → Social Movement
- Hyper Local News → Urban-Rural Info Gap

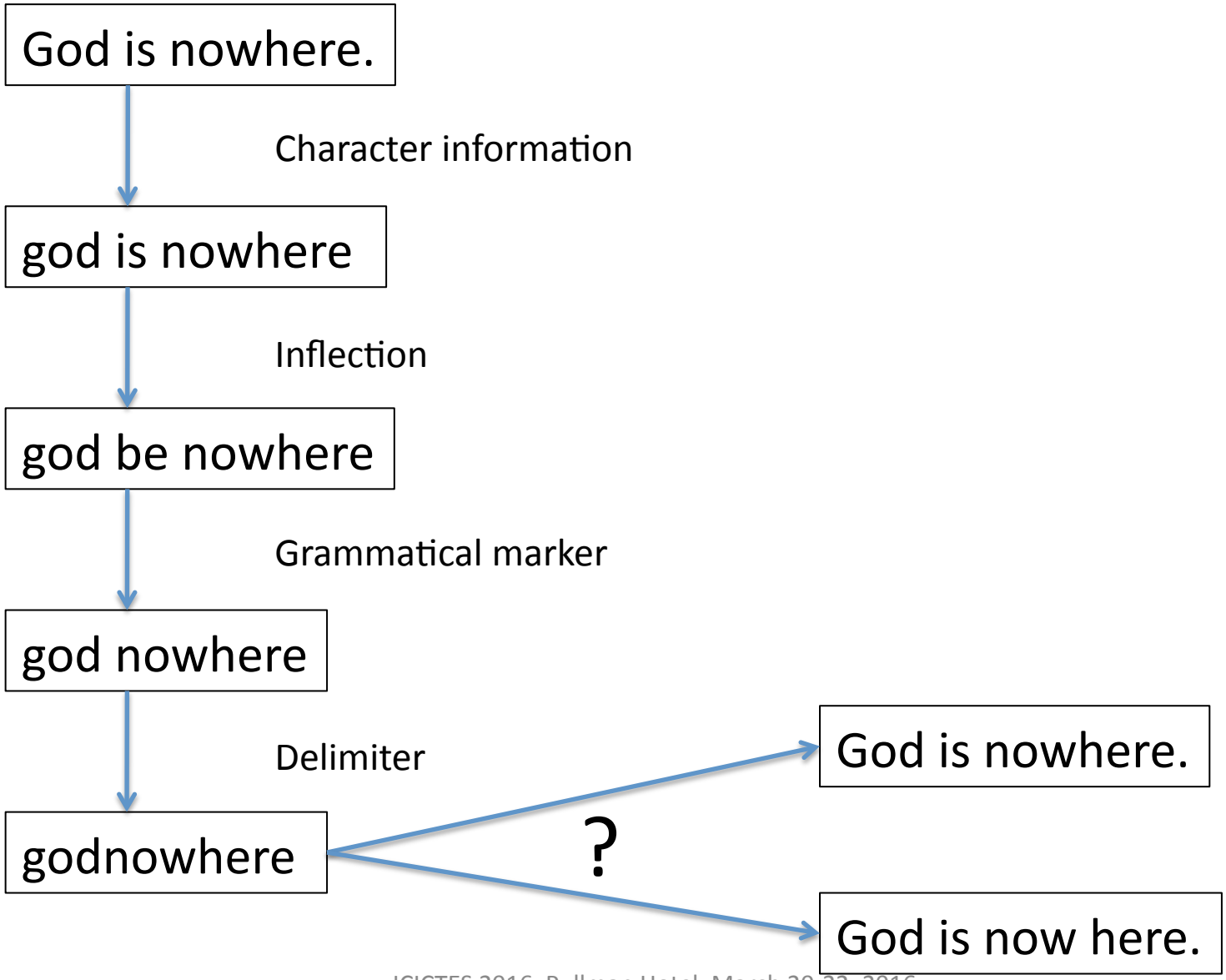
WORD SEGMENTATION

Thai Language as a Non-Segmenting Language

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

How to determine word and sentence boundary?!?!

Simulating a Non-Segmenting Language



Word-Based Approach

Word segmentation (accuracy for Thai)

- Longest matching: 92%
- Maximal matching: 93%
- POS tri-gram: 96%
- Machine learning: 97%

Sentence segmentation (accuracy for Thai)

- POS tri-gram: 84.57%
- Feature-based approach (Winnow): 89.13%

Character-Based Approach
Term Extraction

- Automatic Corpus-Based Thai Word Extraction with the **C4.5** Learning Algorithm
- C4.5-trained decision tree for determining potential word boundary from **MI, Entropy** and some **linguistic information**
- Capable of discovering new words in document without assistance from static dictionary

Attributes(1) : Left and Right Mutual Information

$$MI_L(x\mathbf{y}z) = \frac{p(x\mathbf{y}z)}{p(x)p(\mathbf{y}z)}$$

x | $\mathbf{y}z$

$$MI_R(x\mathbf{y}z) = \frac{p(x\mathbf{y}z)}{p(x\mathbf{y})p(z)}$$

$x\mathbf{y}$ | z

where

x is the leftmost character of string $x\mathbf{y}z$

$\mathbf{y}$ is the middle substring of $x\mathbf{y}z$

z is the rightmost character of string $x\mathbf{y}z$

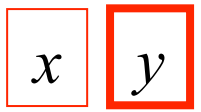
$p(\ )$ is the probability function.

High mutual information implies that $x\mathbf{y}z$ co-occurs more than expected by chance. If *xyz* is a word, its MI_L and MI_R must be high.

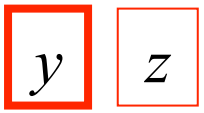
…e**function**… and ...f**unction**...

Attributes(2) : Left and Right Entropy

$$H_L(\mathbf{y}) = -\sum_{\text{all } x \in \mathbf{A}} p(x\mathbf{y} \mid \mathbf{y}) \cdot log_2 p(x\mathbf{y} \mid \mathbf{y})$$



$$H_R(\mathbf{y}) = -\sum_{\text{all } z \in \mathbf{A}} p(\mathbf{y}z \mid \mathbf{y}) \cdot log_2 p(\mathbf{y}z \mid \mathbf{y})$$



where

x is the leftmost character of string *xyz*

y is the middle substring of *xyz*

z is the rightmost character of string *xyz*

p( ) is the probability function.

Entropy shows the variety of characters before and after a word.
If ***y*** is a word, its left and right entropy must be high.
...?**function**... and ...?**unction**...

Attributes(3) : Frequency, Length Function Word

- **Frequency**
  Words tend to be used more often than non-word string sequences.
- **Length**
  Short strings are likely to happen by chance. The long and short strings should be treated differently.
- **Function Word**
  Function words are used mostly in phrases. They are useful to disambiguate words and phrases.

$$Func(s) = \begin{cases} 1 & \text{if } s \text{ contains a function word} \\ 0 & \text{otherwise} \end{cases}$$

Attributes(4): First Two and Last Two Characters

- **Frequency of the first-two characters** of the considered string which appears in the first-two characters of words in the dictionary

  high frequency -> the beginning of the considered string conforms to the Thai spelling system.

  Ex.

  function: how likely *fu* can be the beginning of word.

  This idea can be also applied to **the last-two characters**.

Experimental Results (1)

The Precision of Word Extraction

	No. of strings extracted by the decision tree	No. of words extracted	No. of non-word strings extracted
Training Set	1882 (100%)	1643 (87.3%)	239 (12.7%)
Test Set	1815 (100%)	1526 (84.1%)	289 (15.9%)

The Recall of Word Extraction

	No. of words that has more than 2 occurrences in corpus	No. of words extracted by the decision tree	No. of words in corpus that are found RID
Training Set	2933 (100%)	1643 (56.0%)	1833 (62.5%)
Test Set	2720 (100%)	1526 (56.1%)	1580 (58.1%)

Remark: These precision and recall are measured against 30,000 strings that occur more than 2 times in the corpus and conform to some simple Thai spelling rules.

Experimental Results (2)

Word Extraction VS. a Dictionary

	No. of words extracted by the decision tree	No. of words extracted by the decision tree which is in RID	No. of words extracted by the decision tree which is not in RID
Training Set	1643 (100.0%)	1082 (65.9%)	561 (34.1%)
Test Set	1526 (100.0%)	1046 (68.5%)	480 (31.5%)

Remark: RID is referred to the Thai-Thai dictionary published by The Royal Institute in 1982.

KEYWORD AND SEMANTIC RELATION EXTRACTION

Semantic Link Generation

- Semantic Representation of the description
  - **Keyword Extraction**
    - Extract keywords in text documents and link them to appropriate articles
  - **Semantic Relation Extraction**
    - Extract commons syntactic patterns between two keywords and generalize them to a triple (e_i , r_{ij} , e_j)
- Linked Data
  - Set of triple (e_i , r_{ij} , e_j)

Keyword Extraction

- Some keywords are readily available in the set **tags**, but many of them are still missing.
- Our task is to extract those missing keywords from the **description** and **title**.

Focused NE in Cultural Domain Database
- Cultural attraction (Place)
- Cultural person (Person)
- Cultural artifact (Object)

Training Data Preparation

- Generate a keyword list from **tags** and **titles** that are not shorter than 5 characters and not longer than 30 characters
- POS tagged **descriptions**
- Label the POS tagged descriptions with the keyword list

Labeling

- Apply BIO tagging
  - B: beginning position of a keyword
  - I: intermediate (or end) position of a keyword
  - O: other words
- If several matches are possible, select the longest one

Training Data

- Description

ผ้าซิ่นลายมัดหมี่บ้านปทุมแก้ว เป็นงานฝีมือพื้นบ้าน .....

Training Data

- Description

ผ้าซิ่นลายมัดหมี่บ้านปทุมแก้ว เป็นงานฝีมือพื้นบ้าน …..

- Segmented/Tagged/Labeled Description

Word	POS tag	Label	
ผ้าซิ่น	N	B-K	
ลายมัดหมี่บ้านปทุมแก้ว	N	I-K	
<space>	P	O	
เป็น	V	O	
งานฝีมือพื้นบ้าน	N	O	
……		…..	…..

- Keyword List

(extracted from tag and title)

…..

ผ้า

…..

ผ้าซิ่น

ผ้าซิ่นลายมัดหมี่บ้านปทุมแก้ว

…..

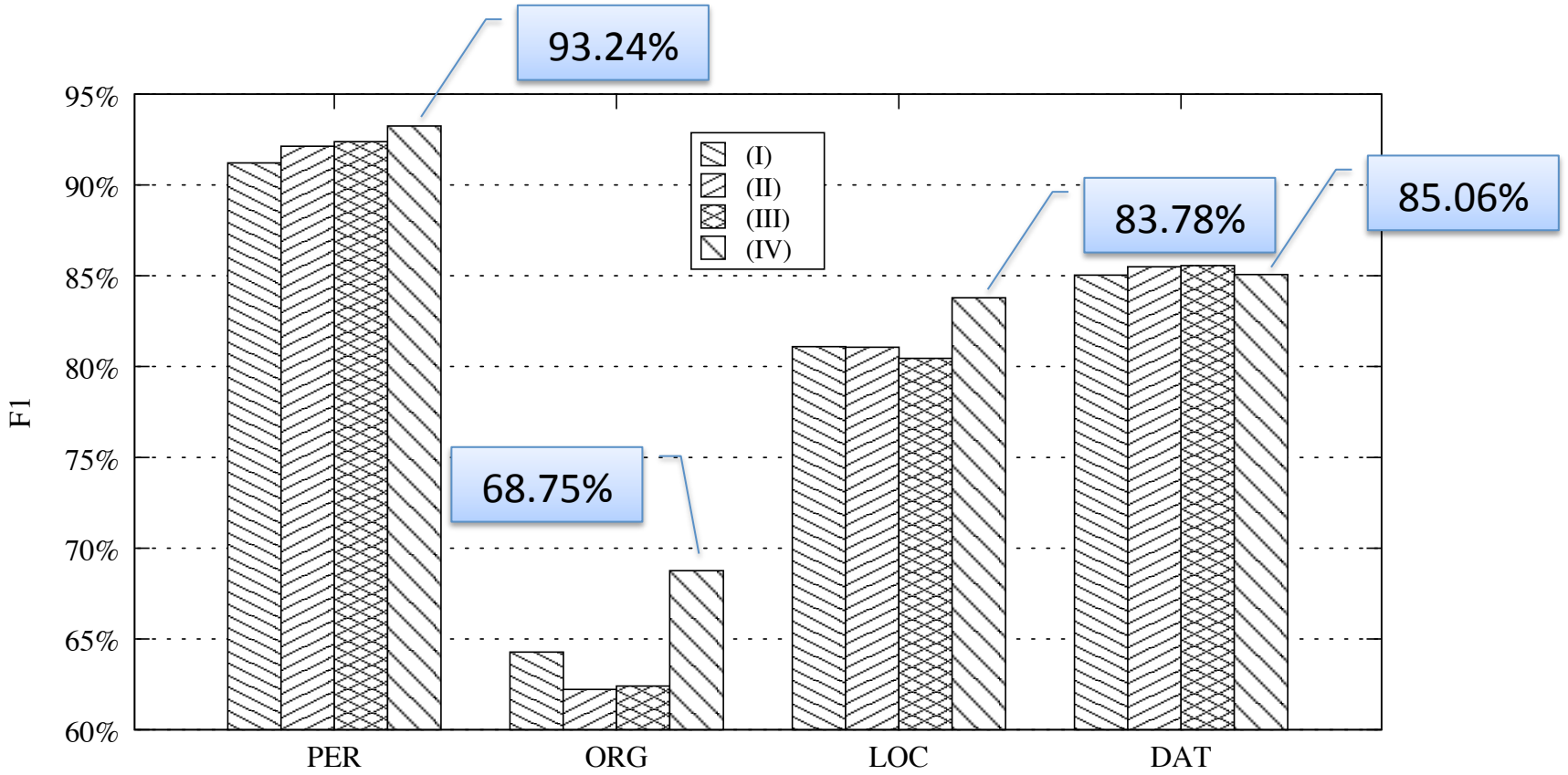
…..

Preliminary Experiment Result

- Based on Margin Infused Relaxed Algorithm (MIRA), Crammer et al., 2005

(I): word 1, 2 grams + label bigrams $\langle w_j \rangle, j \in [-2, 2] \times y_0$ $\langle w_j, w_{j+1} \rangle, j \in [-2, 1] \times y_0$ $\langle y_{-1}, y_0 \rangle$	**(III): (II) + POS 3 grams** $\langle p_j, p_{j+1}, p_{j+2} \rangle, j \in [-2, 0] \times y_0$
(II): (I) + POS 1,2 grams $\langle p_j \rangle, j \in [-2, 2] \times y_0$ $\langle p_j, p_{j+1} \rangle, j \in [-2, 1] \times y_0$	**(IV): (III) + k-char prefixes/suffixes** $\langle P_k(w_0) \rangle, k \in [2, 3] \times y_0$ $\langle S_k(w_0) \rangle, k \in [2, 3] \times y_0$ $\langle P_k(w_0), S_k(w_0) \rangle, k \in [2, 3] \times y_0$

F-Measure



Average evaluation result on NE annotated online news
Recall=0.8256, Precision=0.9061, F1=0.8640

Semantic Relation Acquisition

- Extract commons syntactic patterns between two nouns
- Our task is to acquire triples (e_i , r_{ij} , e_j), where
  - e_i and e_j are entities (keywords)
  - r_{ij} is a relationship between them

Semantic Relation Template

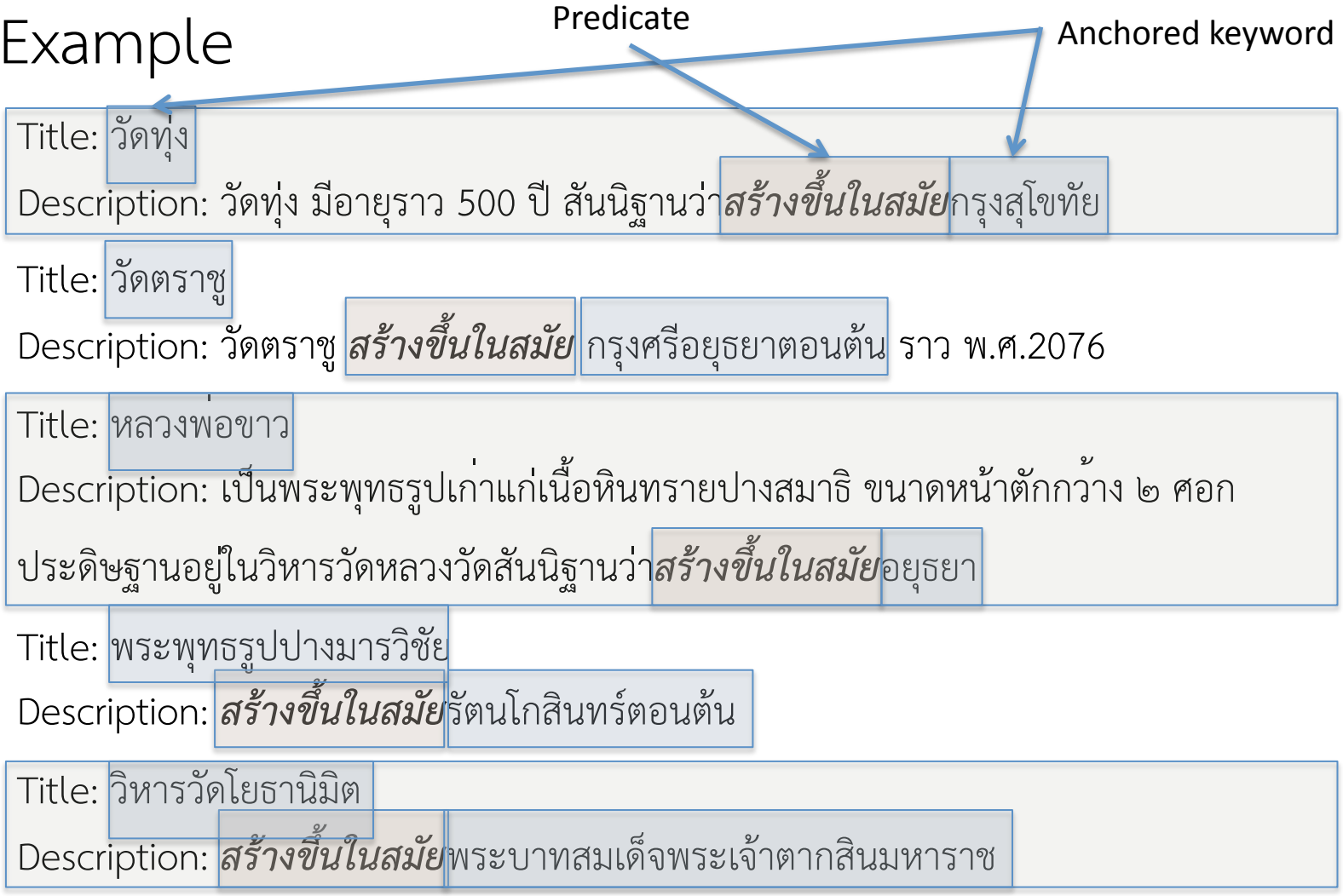
Domain	Relation	Surface	Argument
Cultural attraction	ISLOCATEDAT	ตั้งอยู่ที่	LOC
	ISBUILTIN	สร้าง(ขึ้น)*ใน สร้าง(ขึ้น)*เมื่อ ตั้ง(ขึ้น)*เมื่อ	DATE
	ISBUILTBY	สร้าง(ขึ้น)*โดย ตั้ง(ขึ้น)*โดย	PER, ORG
	HASOLDNAME	เดิมชื่อ ชื่อเดิม	LOC, ORG
Cultural person	MARRIEDWITH	สมรสกับ	PER
	HASFATHERNAME	บิดาชื่อ	PER
	HASMOTHERNAME	มารดาชื่อ	PER
	HASOLDNAME	เดิมชื่อ ชื่อเดิม	PER
	HASBIRTHDATE	เกิด(เมื่อ)*	DATE
	BECOMEMONKIN	อุปสมบทเมื่อ	DATE
Cultural artifact	ISMADEBY	ผลิต(ขึ้น)*โดย ทำ(ขึ้น)*โดย ผลงานโดย	PER, ORG
	ISSOLDAT	จำหน่ายที่	LOC, ORG

Relation Instances Found in Word Distance

Relation	Argument	Distance					
		0	1	2	3	4	5
Cultural attraction							
ISLOCATEDAT	LOC	356	574	591	624	678	757
ISBUILDIN	DATE	3825	11487	11538	11573	11633	11667
ISBUILDBY	PER, ORG	131	202	218	234	249	257
HASOLDNAME	LOC, ORG	0	9	21	26	27	29
Cultural person							
MARRIEDWITH	PER	132	177	177	177	177	177
HASFATHERNAME	PER	120	372	372	373	373	373
HASMOTHERNAME	PER	97	383	383	383	383	383
HASOLDNAME	PER	51	259	273	277	277	283
HASBIRTHDATE	DATE	4122	4745	4801	4947	4966	5075
BECOMEMONKIN	DATE	346	435	435	436	436	436
Cultural artifact							
ISMADEBY	PER, ORG	62	107	109	125	129	130
ISSOLDAT	LOC, ORG	31	31	56	59	62	64

Extract Common Syntactic Pattern of a Predicate between Two Keywords

- Example



Extract Common Syntactic Pattern of a Predicate between Two Keywords

- Example

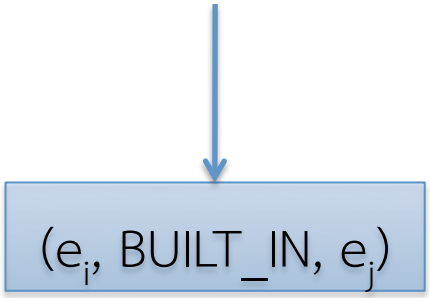
(วัดทุ่ง, ***สร้างขึ้นในสมัย***, กรุงสุโขทัย)

(วัดตราชู, ***สร้างขึ้นในสมัย***, กรุงศรีอยุธยาตอนต้น)

(หลวงพ่อขาว, ***สร้างขึ้นในสมัย***, อยุธยา)

(พระพุทธรูปปางมารวิชัย, ***สร้างขึ้นในสมัย***, รัตนโกสินทร์ตอนต้น)

(วิหารวัดโยธานิมิต, ***สร้างขึ้นในสมัย***, พระบาทสมเด็จพระเจ้าตากสินมหาราช)



Extract Common Syntactic Pattern of a Predicate between Two Keywords

- Example

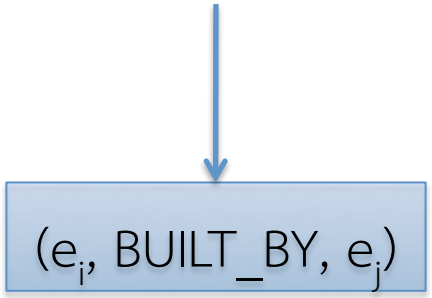
(กระโจมไฟบ้านโรงถ่าน, ***สร้างโดย***, อพท.)

(ศาลเจ้าตากสิน วัดบ้านค่าย, ***สร้างโดย***, พระครูพิพัฒน์ชยาภรณ์)

(วัดทุ่งโฮ้งใต้, ***สร้างโดย***, กลุ่มชาวลาวพวน)

(ศาลพระพรหม, ***สร้างโดย***, เทศบาลนครเชียงราย)

(วงเวียนนิมิตร, ***สร้างโดย***, เทศบาลนครภูเก็ต)

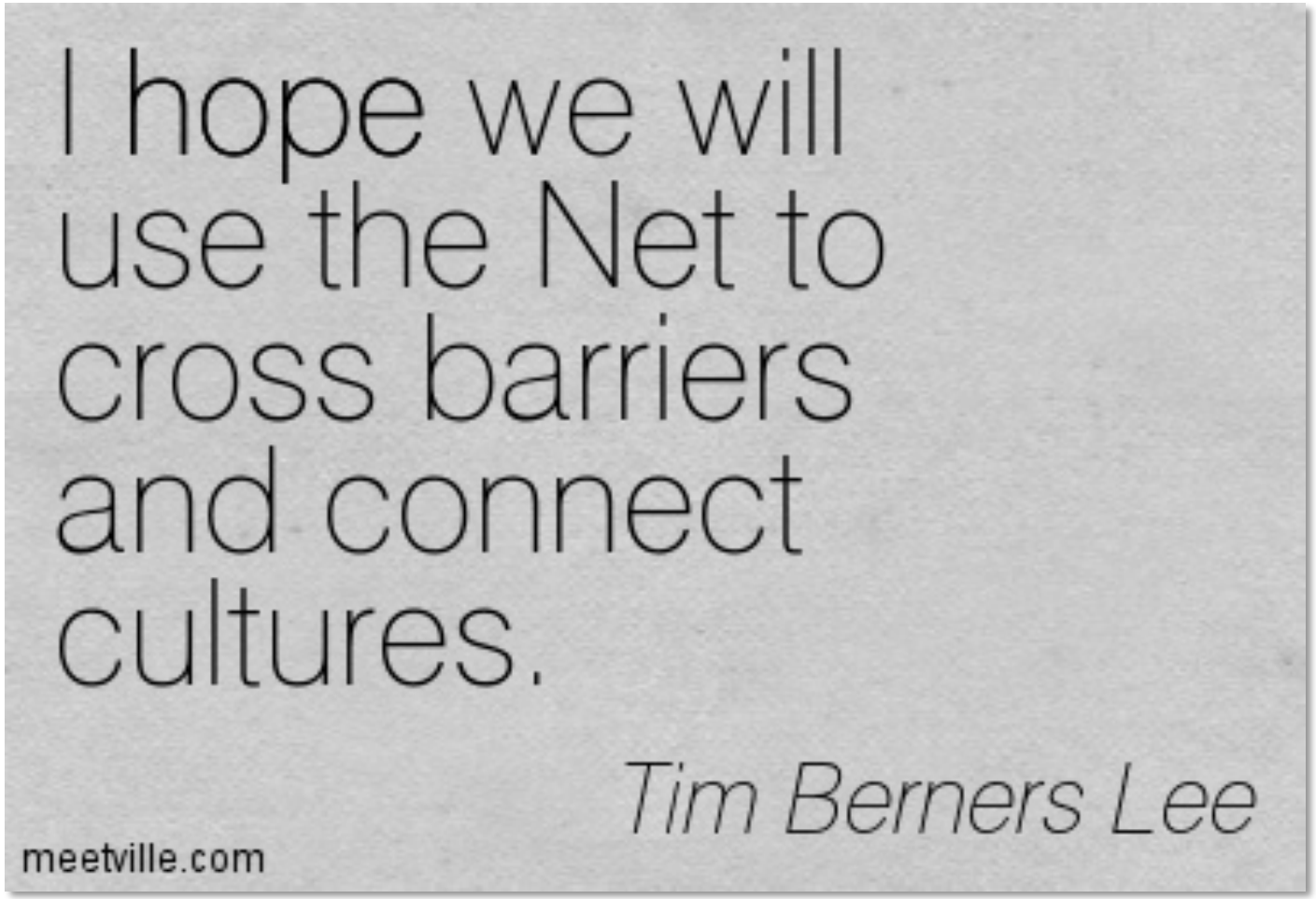


Accuracy of Relation Extraction

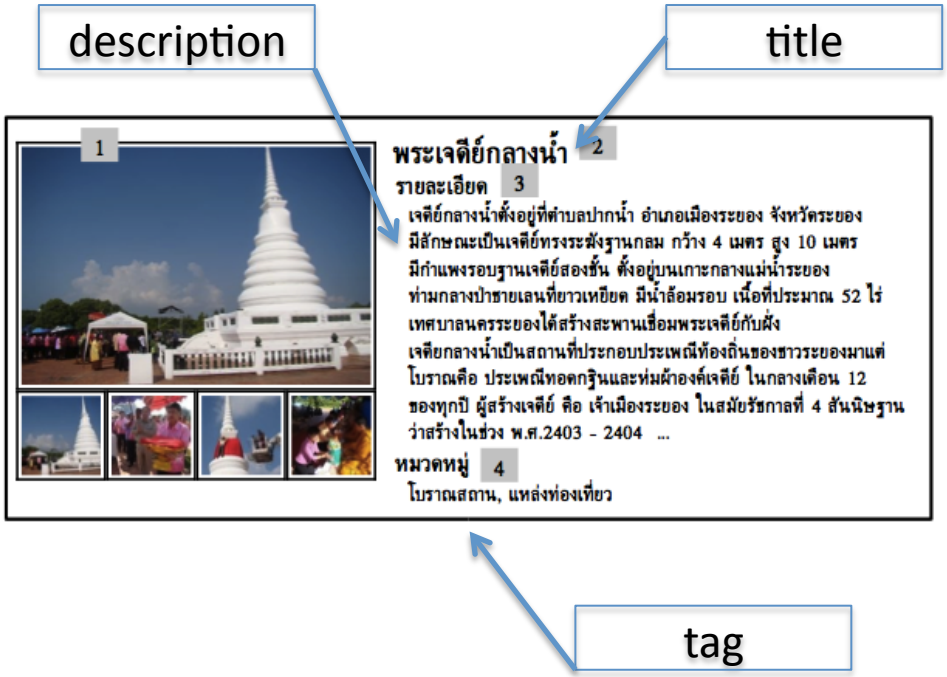
Relation	Argument	#Sample	#Correct	#Incorrect	Accuracy
Cultural attraction					
ISLOCATEDAT	LOC	50	49	1	98%
ISBUILDIN	DATE	50	48	2	96%
ISBUILDBY	PER, ORG	50	48	2	96%
HASOLDNAME	LOC, ORG	27	23	4	85%
Cultural person					
MARRIEDWITH	PER	50	49	1	98%
HASFATHERNAME	PER	50	48	2	96%
HASMOTHERNAME	PER	50	49	1	98%
HASOLDNAME	PER	50	47	3	94%
HASBIRTHDATE	DATE	50	48	2	96%
BECOMEMONKIN	DATE	50	50	0	100%
Cultural artifact					
ISMADEBY	PER, ORG	50	44	6	88%
ISSOLDAT	LOC, ORG	50	49	1	98%

NLP Applications

KNOWLEDGE MAP GENERATION


I hope we will
use the Net to
cross barriers
and connect
cultures.
Tim Berners Lee
meetville.com

Types of Semantic Relation



Domain	Relation	Surface	Argument
Cultural attraction	ISLOCATEDAT	ตั้งอยู่ที่	LOC
	ISBUILTIN	สร้าง(ขึ้น)*ใน สร้าง(ขึ้น)*เมื่อ ตั้ง(ขึ้น)*เมื่อ	DATE
	ISBUILTBY	สร้าง(ขึ้น)*โดย ตั้ง(ขึ้น)*โดย	PER, ORG
	HASOLDNAME	เดิมชื่อ ชื่อเดิม	LOC, ORG
Cultural person	MARRIEDWITH	สมรสกับ	PER
	HASFATHERNAME	บิดาชื่อ	PER
	HASMOTHERNAME	มารดาชื่อ	PER
	HASOLDNAME	เดิมชื่อ ชื่อเดิม	PER
	HASBIRTHDATE	เกิด(เมื่อ)*	DATE
	BECOMEMONKIN	อุปสมบทเมื่อ	DATE
Cultural artifact	ISMADEBY	ผลิต(ขึ้น)*โดย ทำ(ขึ้น)*โดย ผลงานโดย	PER, ORG
	ISSOLDAT	จำหน่ายที่	LOC, ORG

Knowledge Map

ISBUILTIN(พระเจดีย์กลางน้ำ, พ.ศ.2403)

ISLOCATEDAT(พระเจดีย์กลางน้ำ, ตำบลปากน้ำ)

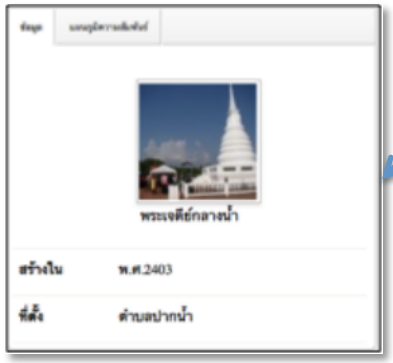
Infobox

Knowledge map

Knowledging

Semantically Enhanced
Cultural Database
[Place, Person, Artifact]

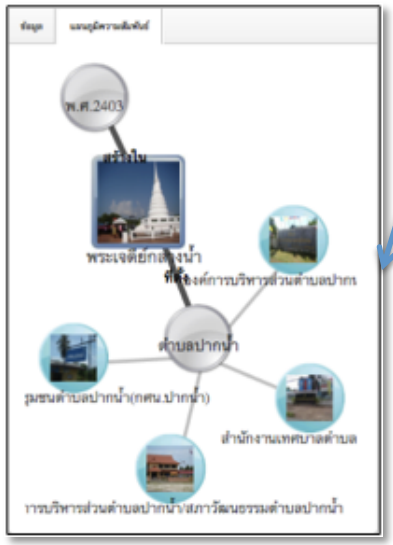
Infobox



Creator

Product

Knowledge map



Making

Shop

SOCIAL MOVEMENT UNDERSTANDING

Data Data Data!!!

- Drastically increase number of users on social network
- Keywords in the contents express the concepts of the talk
- Social media texts are input in a time sequence
- But, social media texts are normally short, incomplete and diverse

Word Article Matrix (WAM)

Creating WAM

Wikipedia Articles

Page Contents

Word seg. & Lemmatization

Thinking -> Think

Word list

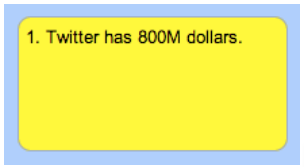
Page Title

Pages\Words	"Twitter"	"Tennis"	"Dollar"	"Google"	...
IT	2	0	1	4	
Sport	0	2	1	0	
Economics	0	0	2	0	
...					

Wikipedia WAM

Text Similarity

SIM Function: Dot Product $\sum_{t \in q \wedge t \in d} (w_{q,t} \cdot w_{d,t})$



"Twitter has 800M dollars"

Word seg. &
Lemmatization
dollars -> dollar

"Twitter"	"Tennis"	"Dollar"	"Google"
1	0	1	0

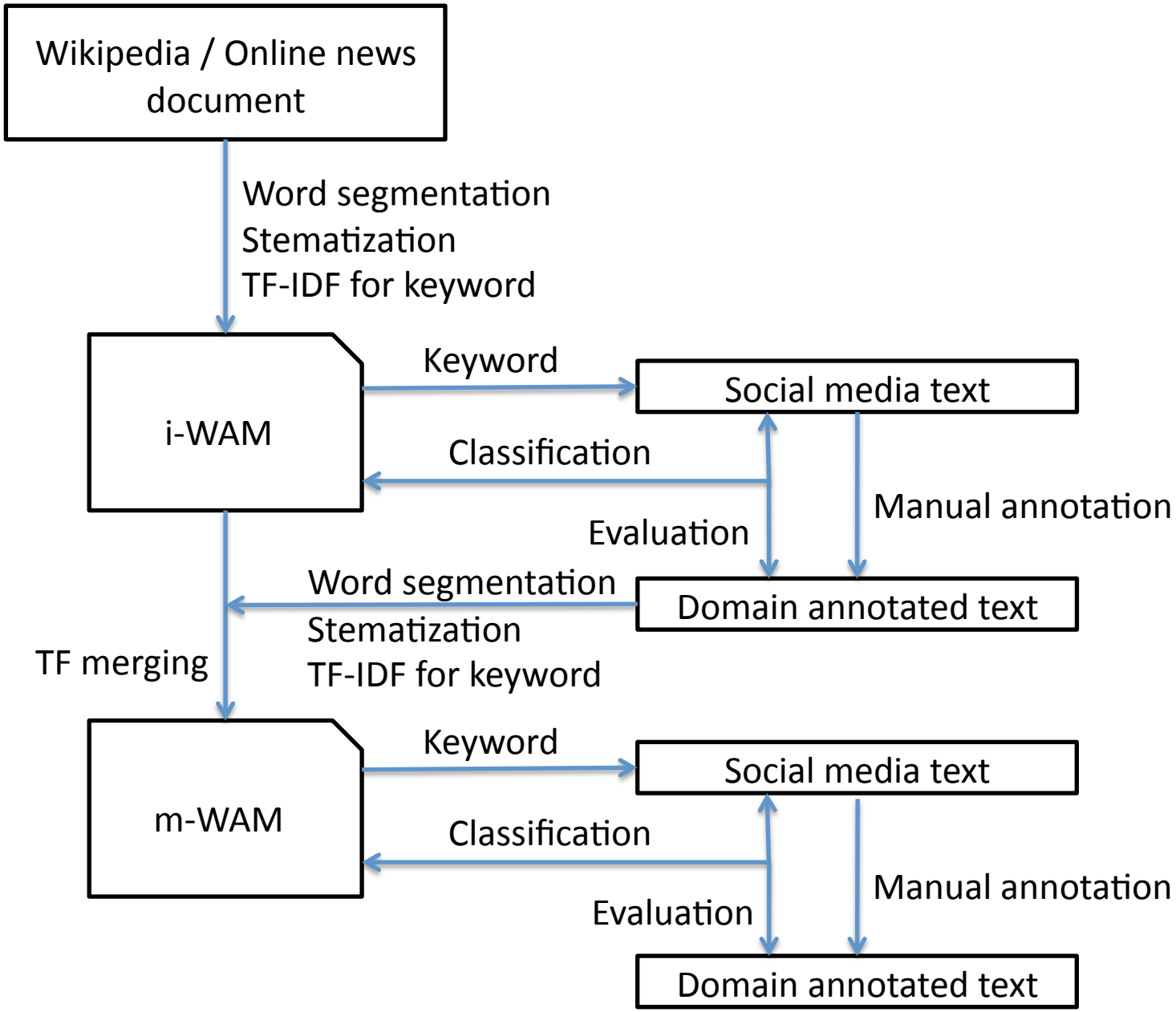
Pages\Words	"Twitter"	"Tennis"	"Dollar"	"Google"	...
IT	2	0	1	4	
Sport	0	2	1	0	
Economics	0	0	2	0	
...					

Wikipedia WAM

(n=2) Select Top-n
Most Associated Documents Ranking

Pages	Dot Product Score
IT	3
Sport	0
Economics	2
...	

Modified WAM for Social Media Text Classification



A Part of Modified WAM

i-WAM	อิเทล 'intel'	ศัลยกรรม 'surgery'	เมตตา 'mercy'	คลาวด์ 'cloud'	คลิป 'clip'	เน็ต 'network'
Life	0.052	0.205	0.160	0.168	0	0
Education	0	0.025	0.036	0	0	0
Technology	0.103	0	0	0.230	0	0

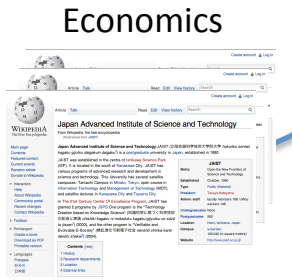
m-WAM	อิเทล 'intel'	ศัลยกรรม 'surgery'	เมตตา 'mercy'	คลาวด์ 'cloud'	คลิป 'clip'	เน็ต 'network'
Life	0.026	0.275	0.268	0.177	0.036	0.040
Education	0	0.013	0.018	0	0	0
Technology	0.164	0	0	0.281	0.009	0.187

m-WAM2	อิเทล 'intel'	ศัลยกรรม 'surgery'	เมตตา 'mercy'	คลาวด์ 'cloud'	คลิป 'clip'	เน็ต 'network'
Life	0.013	0.255	0.241	0.174	0.119	0.034
Education	0	0.009	0.011	0	0.009	0.019
Technology	0.107	0	0	0.238	0.019	0.143

F-Measure

	Life	Education	Technology
i-WAM on text1	86.79%	36.09%	55.79%
m-WAM on text1	86.64%	34.57%	56.01%
m-WAM on text2	91.33%	29.91%	43.17%
m-WAM2 on text2	94.93%	29.47%	43.72%

Tweeter Viewer



Word segmentation

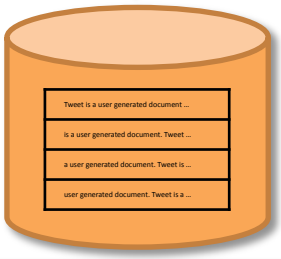
Article

Word

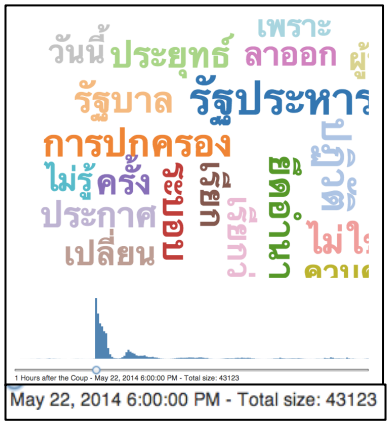
Pages \Words	"Twitter"	"Tennis"	"Dollar"	"Google"	...
IT	2	0	1	4	
Sport	0	2	1	0	
Economics	0	0	2	0	
...					

Word Article Matrix (WAM)

Tweet query by keyword



Domain specific tweets



Social movement timeline viewer

Coup on May 22, 2014

- ทหาร, คสช., ประเทศ, ประกาศ, สงบ, อำนาจ, รัฐบาล, รัฐประหาร, ชุมนุม, ตำรวจ, สถานการณ์, นายก, ควบคุม, ยึด, ประชุม, เศรษฐกิจ, กฎหมาย, ศึก, แกนนำ, รัฐมนตรี, เลือกตั้ง, ประชาธิปไตย, ปฏิวัติ. ยึดอำนาจ. เคอร์ฟิว. กฎอัยการศึก

- military, NCPO, country, announce, peace, power, government, coup d'etat, gathering, police, situation, PM, control, seize, meeting, economy, law, war, leader, minister, election, democracy, revolution, seize the power, curfew, martial law

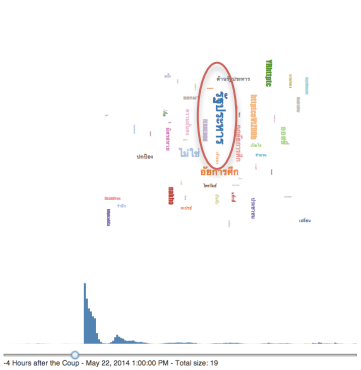
Tweet Query

- Search Tweets by using Restful API
  - GET search/tweets. Set q = the keyword set
  - 100 tweets/search limited
  - Repeatedly fetch data until all tweets in the coup periods are discovered
- Be able to search back to 7 days

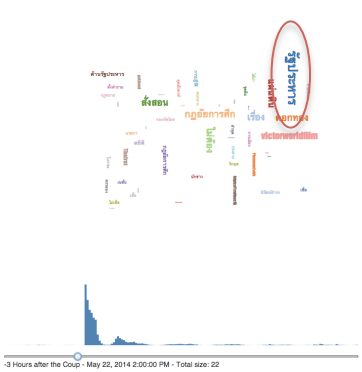
May 22, 2014 Coup-related tweet : 339,148 tweets

http://sn.iisilab.org/

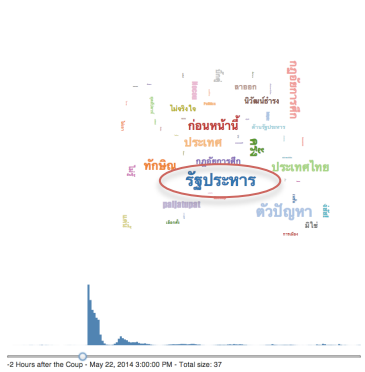
Timeline Word Cloud



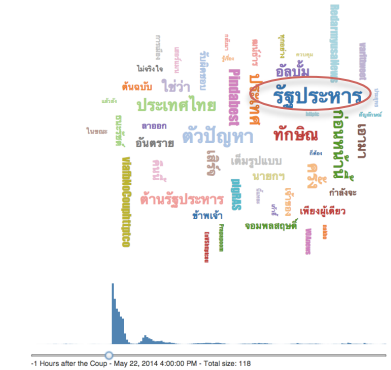
May 22, 2014 1:00:00 PM - Total size: 19



May 22, 2014 2:00:00 PM - Total size: 22

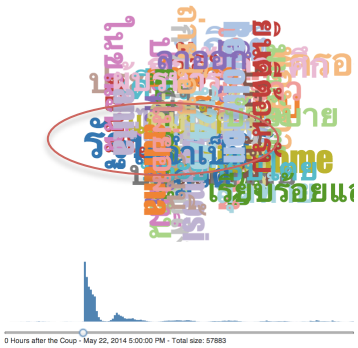


May 22, 2014 3:00:00 PM - Total size: 37

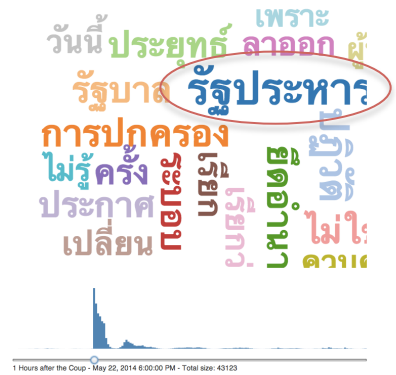


May 22, 2014 4:00:00 PM - Total size: 118

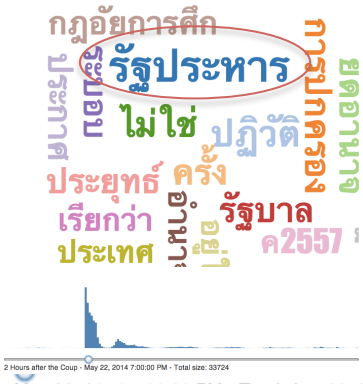
Coup D'etat



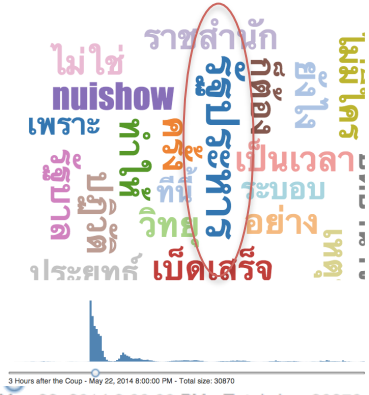
May 22, 2014 5:00:00 PM - Total size: 57883



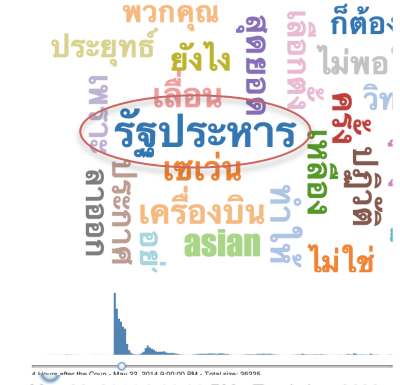
May 22, 2014 6:00:00 PM - Total size: 43123



May 22, 2014 7:00:00 PM - Total size: 33724



May 22, 2014 8:00:00 PM - Total size: 30870



May 22, 2014 9:00:00 PM - Total size: 26225

URBAN-RURAL INFORMATION GAP FILLER

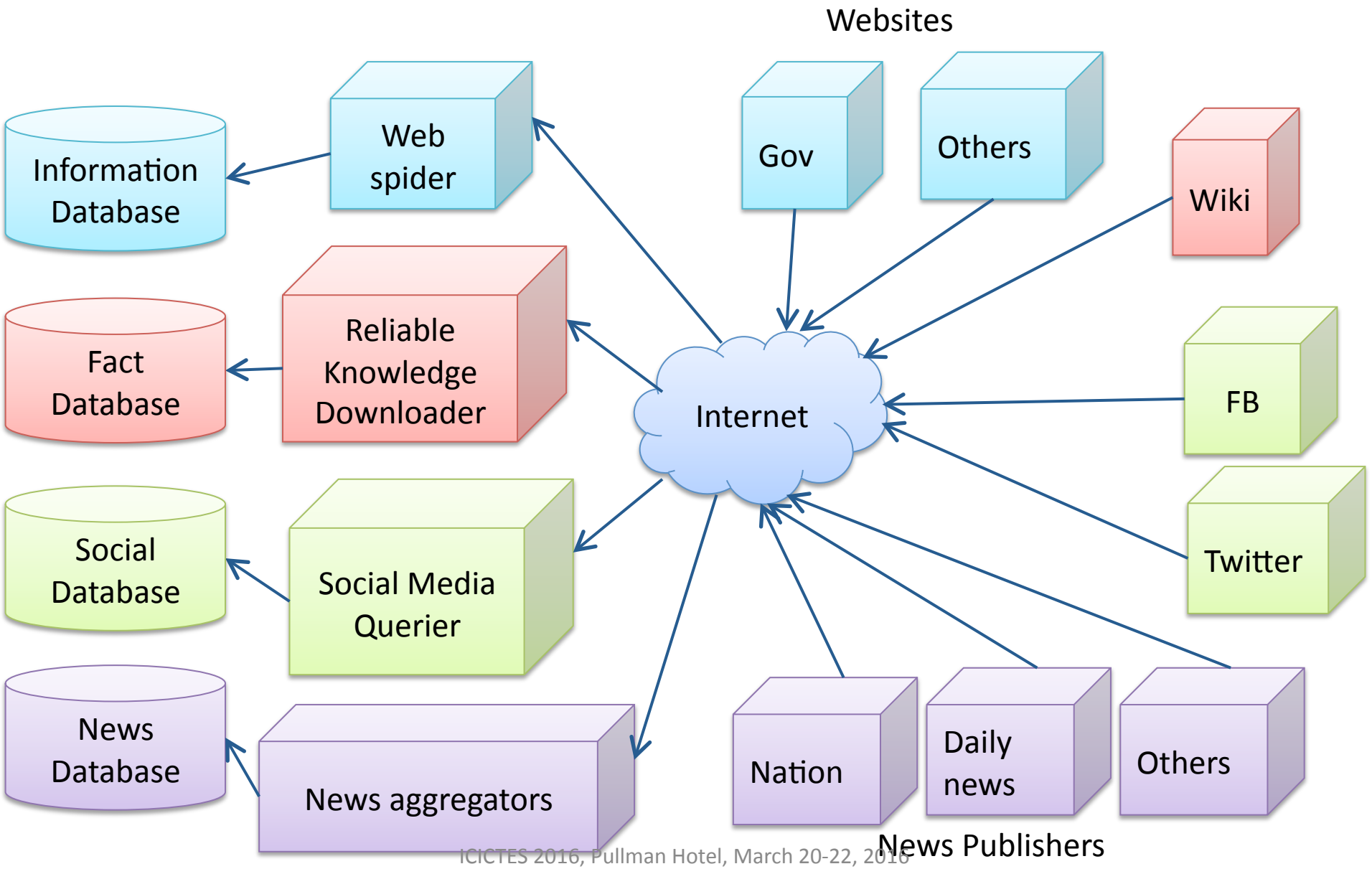
Purpose

- To deliver local news to complement the deteriorating local newspaper
- To promote data usage in rural area

To improving rural livelihood through its reputation extraction!

http://nation.iisilab.org

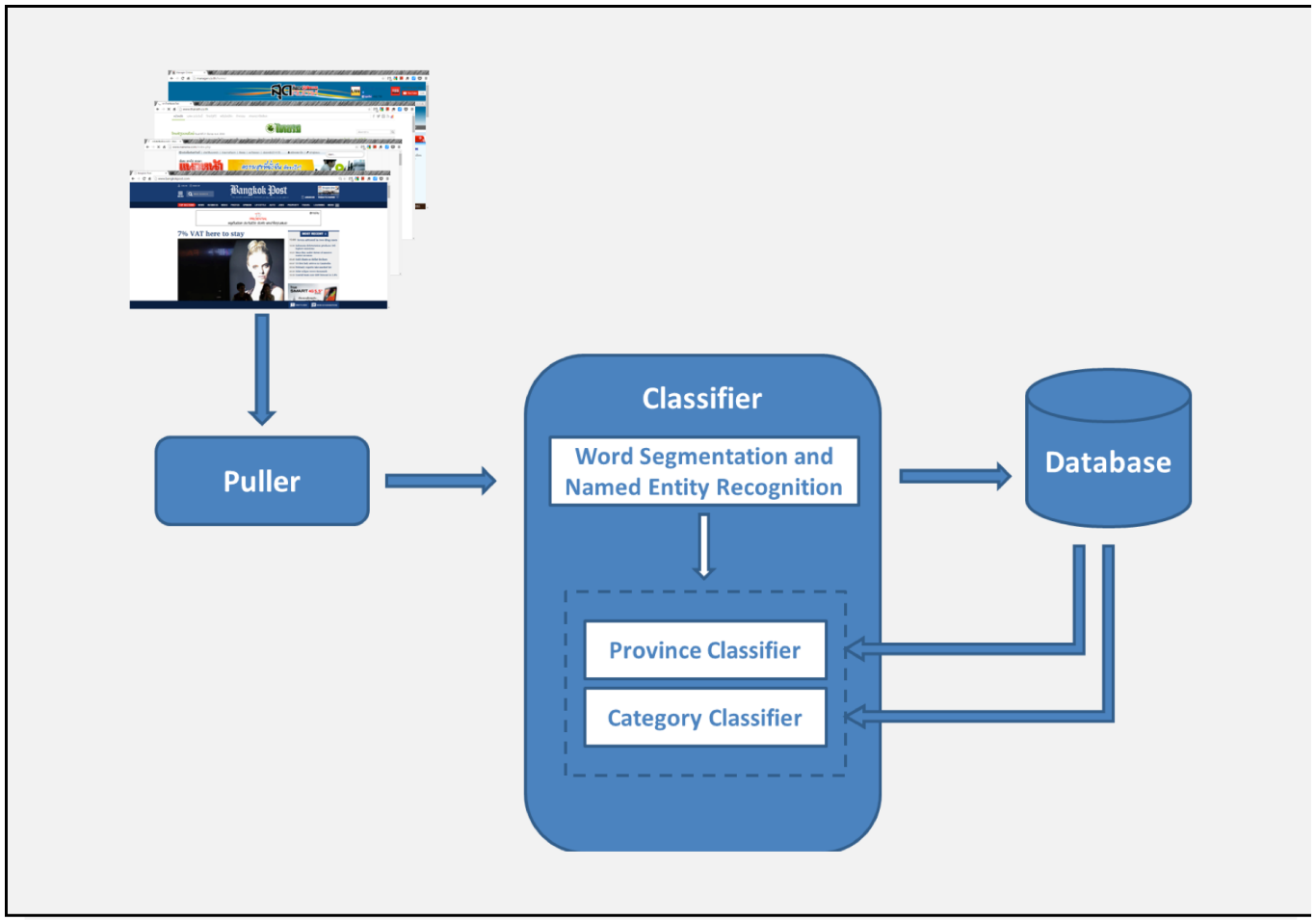
Collecting information

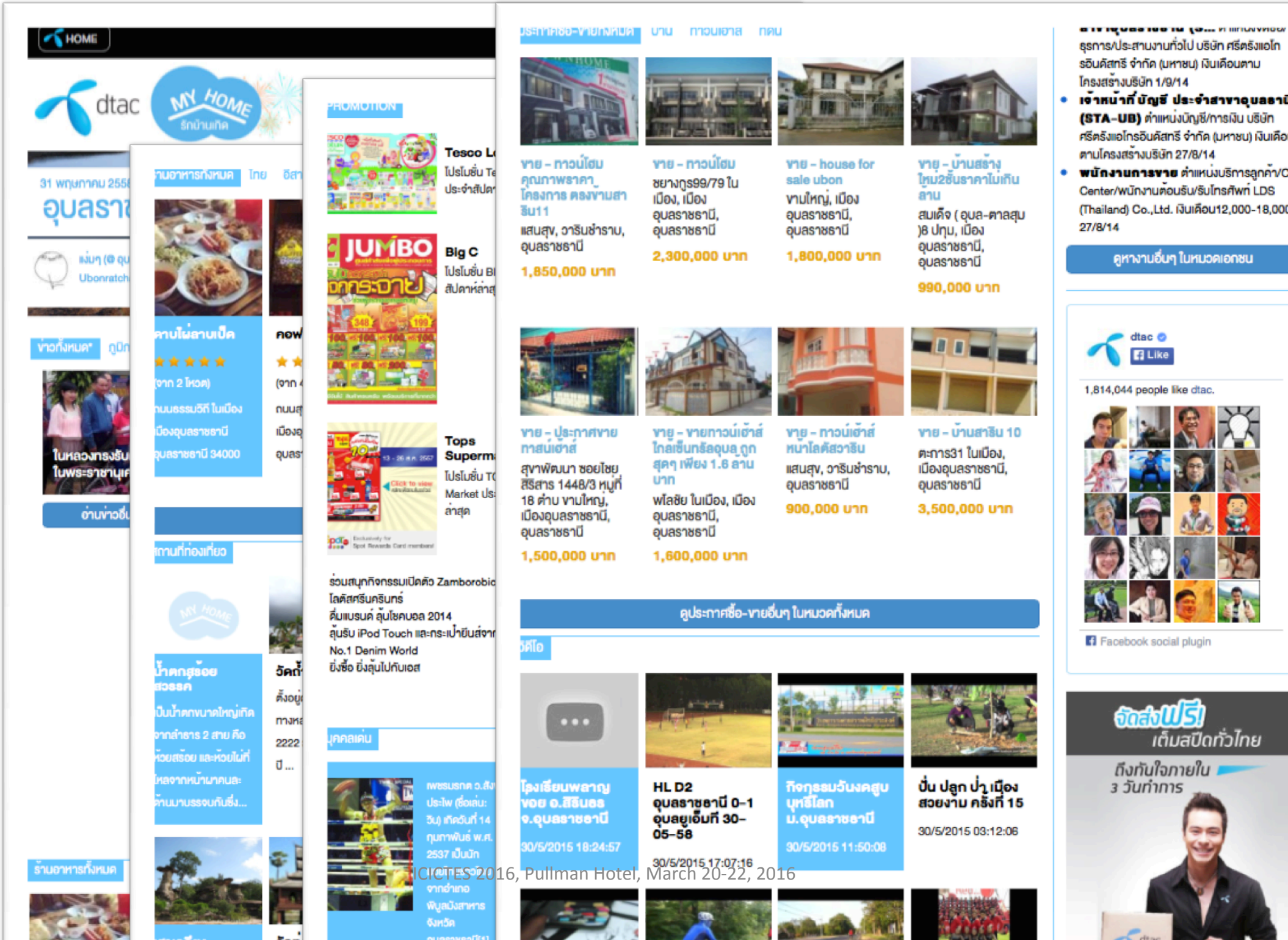


Analyze

- **NEWS Classification**
  - Word segmentation
  - Named Entity Recognition
  - NEWS domain and province classification and ranking based on TF/IDF technique
- **Information Extraction**
  - Template based extraction
- **Infobox**
  - Celebrity i.e. politician, idol
  - Province
- **Social Media**
  - Trending topic
  - Leader and follower
  - Opinion polarity

News Classification




HOME
dtac
MY HOME
รักบ้านเกิด
PROMOTION
Tesco L
โปรโมชั่น Te
ประจำสัปดา
JUMBO
Big C
โปรโมชั่น BI
สัปดาห์ล่าสุ
Tops
Superm
โปรโมชั่น TO
Market ประ
ล่าสุด
ร่วมสนุกกิจกรรมเปิดตัว Zamborobic
โลตัสศรีนครินทร์
ดื่มแบรนด์ ลุ้นโชคบอล 2014
ลุ้นรับ iPod Touch และกระเป๋าเป้ยีนส์จาก
No.1 Denim World
ยิ่งซื้อ ยิ่งลุ้นไปกับเอส
ร้านอาหารทั้งหมด
ไทย
สถานที่ท่องเที่ยว
บุคคลเด่น
ประกาศซื้อ-ขายทั้งหมด
ขาย – ทาวน์โฮม คุณภาพราคา โครงการ ตองขามสา รินi1
แสนสุข, วารินชำราบ, อุบลราชธานี
1,850,000 บาท
ขาย – ทาวน์โฮม
ชยางกูร99/79 ใน เมือง, เมือง อุบลราชธานี, อุบลราชธานี
2,300,000 บาท
ขาย – house for sale ubon
ขามใหญ่, เมือง อุบลราชธานี, อุบลราชธานี
1,800,000 บาท
ขาย – บ้านสร้างใหม่2ชั้นราคาไม่เกินล้าน
สบเด็จ (อุบล-ตาลสุม)8 ปทุม, เมือง อุบลราชธานี, อุบลราชธานี
990,000 บาท
ขาย – ประกาศขาย ทาสน์เฮาส์
สุขาฟัฒนา ซอยโชย สิริสาร 1448/3 หมู่ที่ 18 ตำบ ขามใหญ่, เมืองอุบลราชธานี, อุบลราชธานี
1,500,000 บาท
ขาย – ขายทาวน์เฮ้าส์ ไกล้เซ็นทรัลอุบล ถูกสุดๆ เพียง 1.6 ล้าน บาท
พโลชัย ในเมือง, เมือง อุบลราชธานี, อุบลราชธานี
1,600,000 บาท
ขาย – ทาวน์เฮ้าส์ หน้าโลตัสวารินชำราบ
แสนสุข, วารินชำราบ, อุบลราชธานี
900,000 บาท
ขาย – บ้านสารินชำราบ 10
ตะการ31 ในเมือง, เมืองอุบลราชธานี, อุบลราชธานี
3,500,000 บาท
ดูประกาศซื้อ-ขายอื่นๆ ในหมวดทั้งหมด
วีดีโอ
โรงเรียนพลาญขอย อ.สิรินธร จ.อุบลราชธานี
30/5/2015 18:24:57
HL D2 อุบลราชธานี 0-1 อุบลยูเอ็มที่ 30-05-58
30/5/2015 17:07:16
กิจกรรมวันงดสูบบุหรี่โลก ม.อุบลราชธานี
30/5/2015 11:50:08
ปั่น ปลูก ป่า เมือง สวยงาม ครั้งที่ 15
30/5/2015 03:12:06
ธุรการ/ประสานงานทั่วไป บริษัท ศรีตรังแอโกรอินดัสทรี จำกัด (มหาชน) เงินเดือนตามโครงสร้างบริษัท 1/9/14
เจ้าหน้าที่บัญชี ประจำสาขาอุบลราชธานี (STA-UB) ตำแหน่งบัญชี/การเงิน บริษัท ศรีตรังแอโกรอินดัสทรี จำกัด (มหาชน) เงินเดือนตามโครงสร้างบริษัท 27/8/14
พนักงานการขาย ตำแหน่งบริการลูกค้า/Call Center/พนักงานต้อนรับ/รับโทรศัพท์ LDS (Thailand) Co.,Ltd. เงินเดือน12,000-18,000 27/8/14
ดูหางานอื่นๆ ในหมวดเอกชน
dtac
Like
1,814,044 people like dtac.
Facebook social plugin
จัดส่งฟรี
เต็มสปีดทั่วไทย
ถึงทันใจภายใน 3 วันทำการ

Summary

- **Word segmentation, Keyword extraction, NER, WSD, Semantic relation extraction etc.** are the basic issues of NLP in the current language environment.
- **Efficient computational scenario** is crucial under the drastic growth of data and coverage of the Internet.
- **User-generated contents** are the unlimited language resources.
- NLP technology helps in filling the **information gap** between the haves and the have-nots, filling the **technology gap** between the resource-rich and under-resourced languages.

References

- Virach Sornlertlamvanich, Tanapong Potipiti and Thatsanee Charoenporn. **Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm**. Proceedings of the 18th International Conference on Computational Linguistics (COLING2000), Saarbrucken, Germany, pp 802-807, July-August 2000.
- Virach Sornlertlamvanich, Thatsanee Charoenporn, Canasai Kruengkraii, and Hitoshi Isahara. **Statistical-based Approach to Non-segmented Language Processing**, IEICE Transactions on Information and Systems, Vol. E90-D, No.10, pp.1565-1573, October 2007.
- Canasai Kruengkrai, Virach Sornlertlamvanich, Watchira Buranasing, and Thatsanee Charoenporn. **Semantic Relation Extraction from a Cultural Database**, Proceedings of Workshop on South and Southeast Asian NLP, COLING2012, Mumbai, India, December 8-15, 2012.
- Virach Sornlertlamvanich and Kobkrit Viriyayudhakorn. **Social Movement Understanding by Keyword Tracking**, In Book: Information Modelling and Knowledge Bases XXVI, Volume 272, Eds. Thalheim, B. and et al., IOS Press, 2014.
- Virach Sornlertlamvanich and Canasai Kruengkrai. **Effectiveness of Keyword and Semantic Relation Extraction for Knowledge Map Generation**, Proceedings of The Second International Workshop on Worldwide Language Service Infrastructure (WLSI), Kyoto University, Kyoto, Japan, January 22-23, 2015.
- Virach Sornlertlamvanich. **Hyper Local NEWS Publishing. ---Collect, Analyze, Visualize---**, International Conference on Information Modelling and Knowledge Bases (EJC2015), Maribor, Slovenia, June 8-12, 2015.

Acknowledgement

- Eakasit Pacharawongsakda
- Kobkrit Viriyayudhakorn
- Canasai Kruengkrai
- Thatsanee Charoenporn
- Tanapong Potipiti
- Sitdhibhong Laokok
- Watchira Buranasing
- Thanasan Tanhermhong
- Wirat Chinnan
- Ministry of Culture
- Claudia Soria (hint of images)