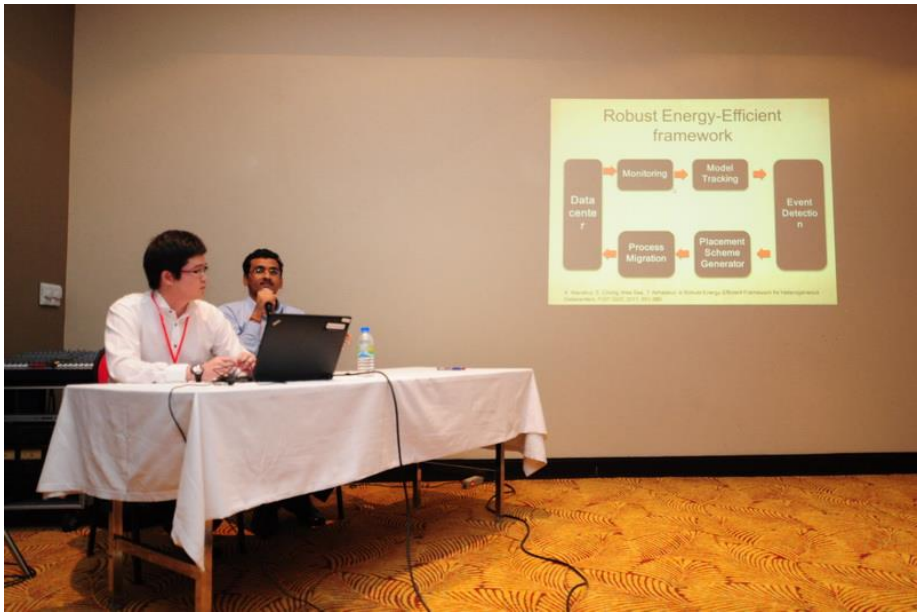
Overview of ECTI-CON 2012

- ECTI-CON 2012 is the ninth annual international conference.
- Organized by Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) Association, Thailand.
- May 16-18, 2012 Hua-Hin Thailand.
- Promote technological progress of thailand.
- The program of ECTI-CON 2012 will consist of 7 area.

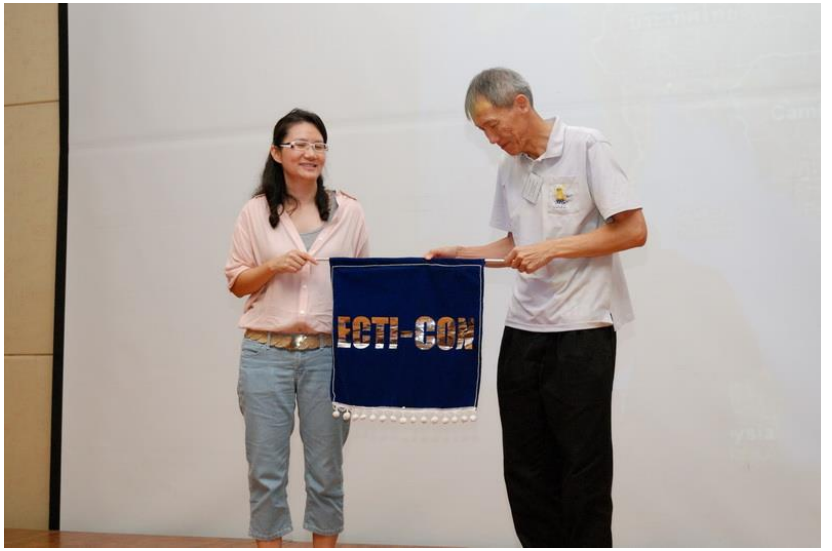
Overview of ECTI-CON 2012

- The program of ECTI-CON 2012 will consist of 7 area.
  - Device, Circuits and Systems
  - Computers and Information Technology
  - Communication Systems
  - Controls
  - Electrical Power Systems
  - Signal Processing
  - Other Related Fields

The atmosphere in ECTI-CON 2012



The atmosphere in ECTI-CON 2012



NECTEC
a member of NSTDA

Search Result Clustering for Thai Twitter Based on Suffix Tree Clustering

Santipong Thaiprayoon, Alisa Kongthon, Pornpimon Palingoon and Choochart Haruechaiyasak

Speech and Audio Technology Laboratory (SPT)
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park

Presentation Outline

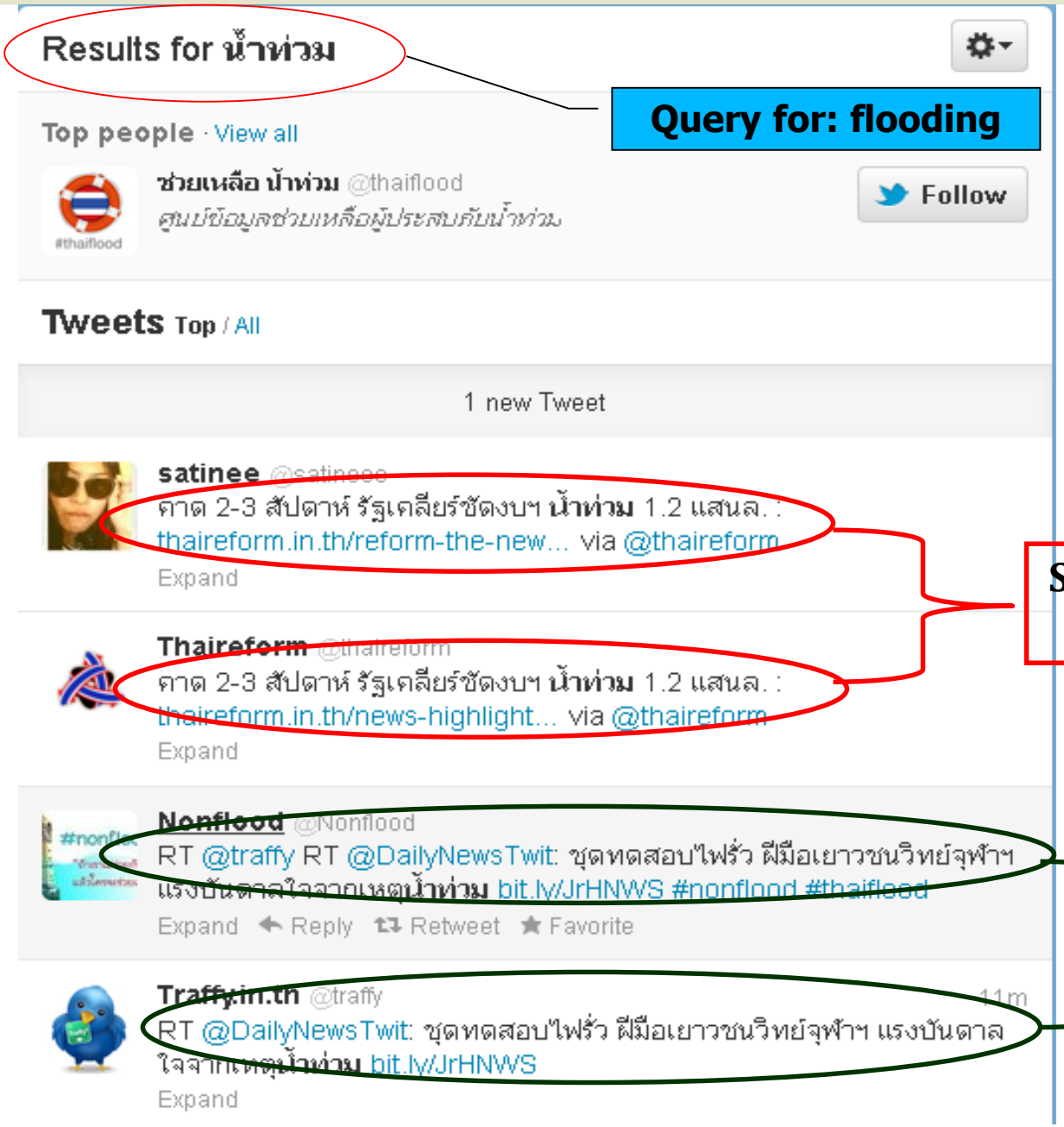
- Background and motivation
- Our proposed approach
  - Suffix Tree Clustering with Label Merging (STC-LM)
- Resource and Experiment
  - Corpus
  - Experiment
  - Evaluation results
- Conclusion
- Future works

Background and Motivation

NECTEC
a member of NSTDA

- Today Twitter has become a popular online medium for posting and sharing news and events.
- There are many Twitter posts containing similar messages on Twitter.
- There are some existing Twitter search engines such as TwitterSearch, TwitterScan and Topsy.
- Users have difficulties in finding relevant results from long list of search results.

Problem of Twitter search engines

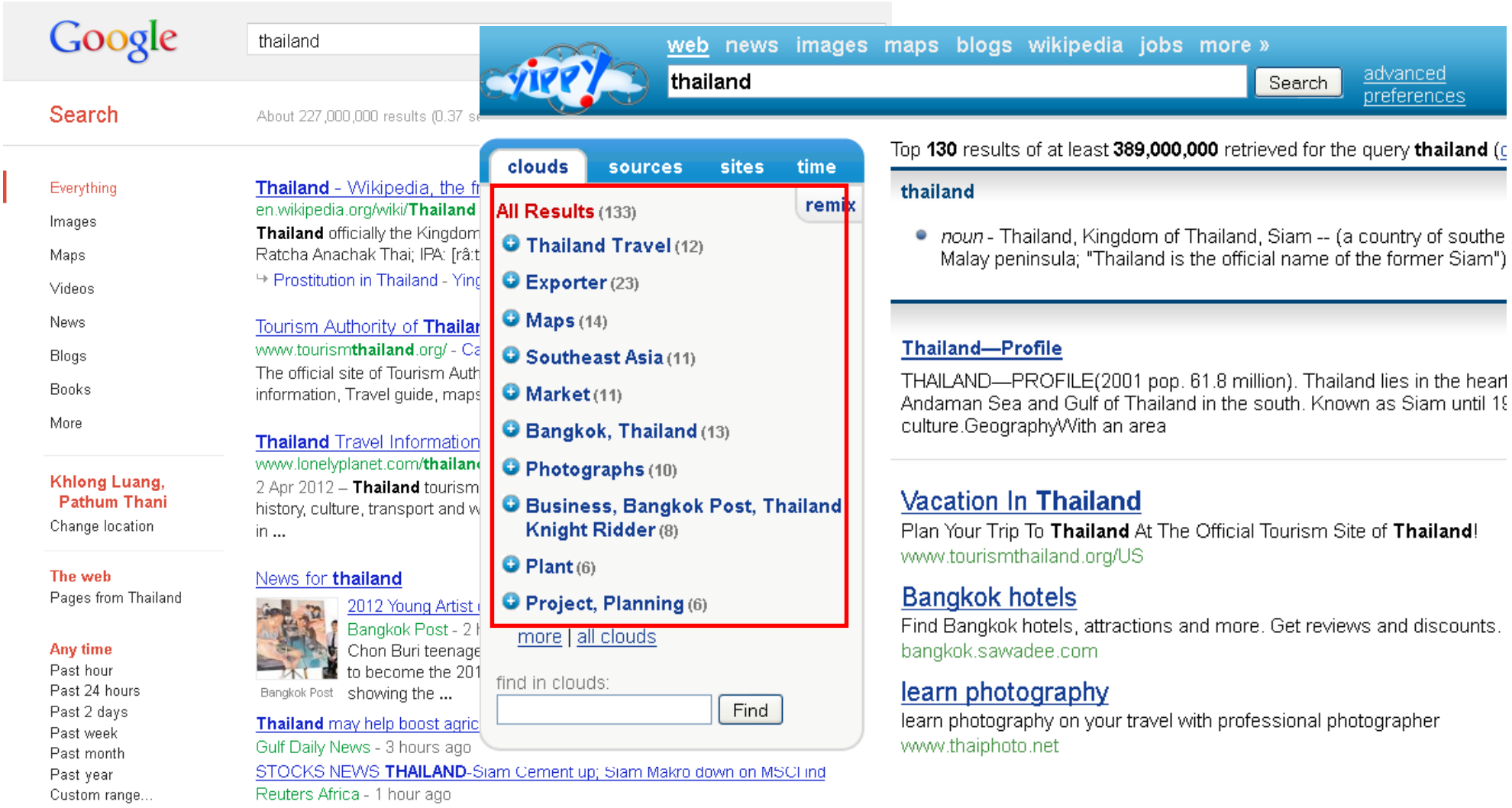


Similar messages on Twitter

Similar messages on Twitter

What is Search Result Clustering (SRC)

- Clustering small collections of documents, search results or document abstract into thematic categories.

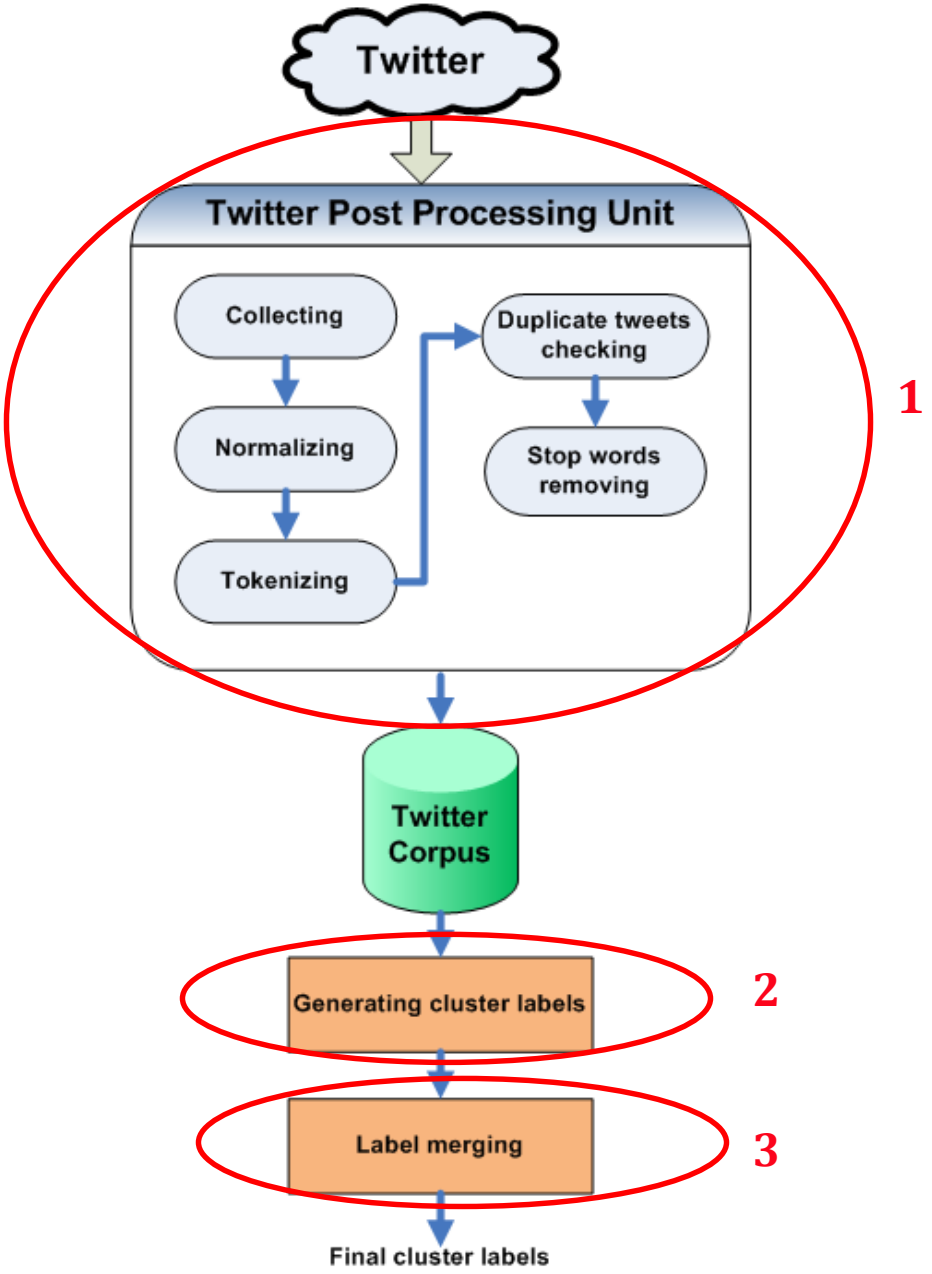


Background and Motivation (cont'd)

NECTEC
a member of NSTDA

- Many previous works use Suffix Tree Clustering (STC) algorithm for clustering search results.
- We apply the STC algorithm which is a part of Carrot2 framework.
- However, the main drawbacks of Carrot2 are.
  - Some of the returned cluster labels are unmeaningful.
  - It is unable to create a hierarchical structure of cluster labels.

Our Proposed Approach



Twitter Post Processing Unit

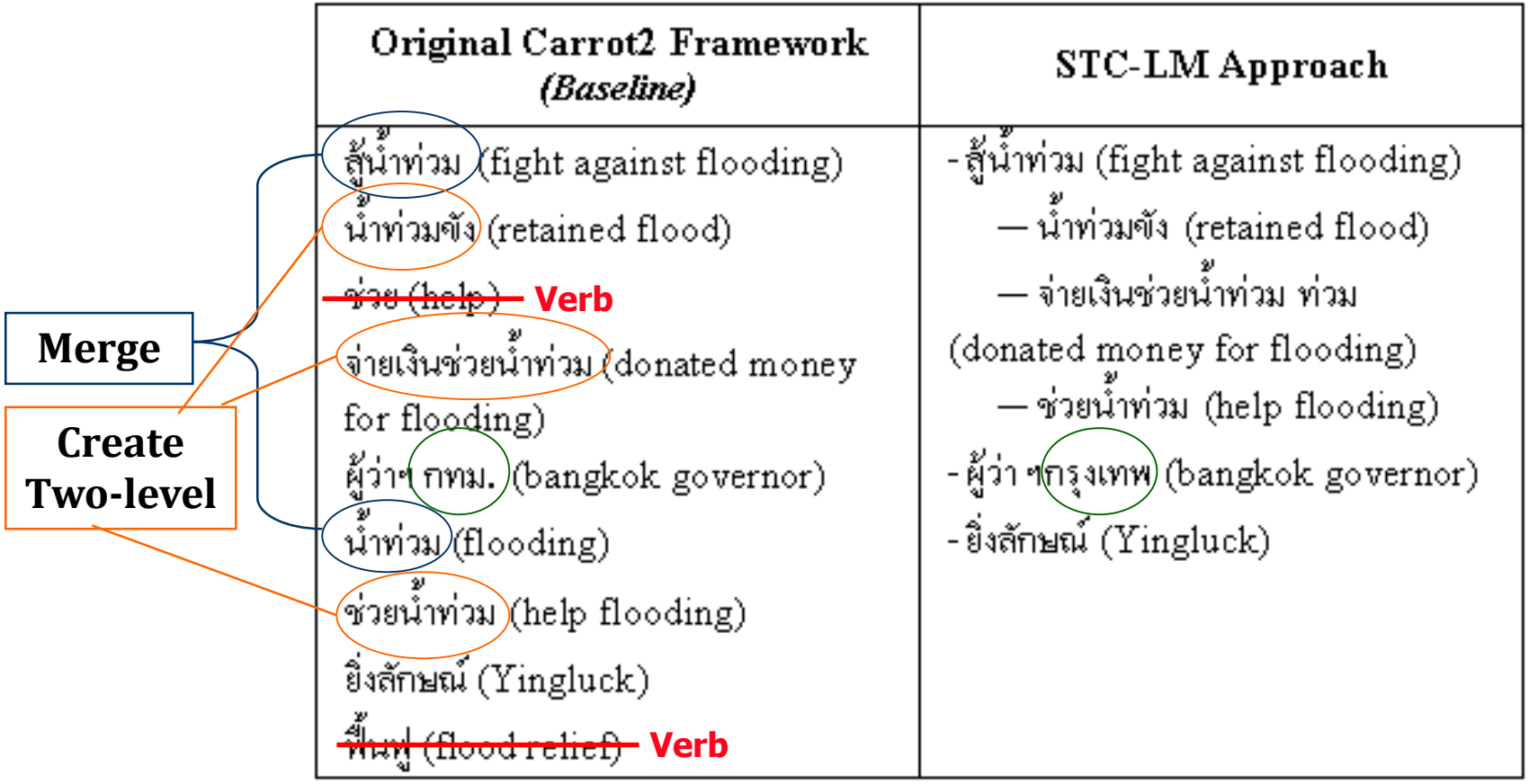
- Collecting: using the twitter4j library
- Normalizing: strip the HTML tags, number and special symbols.
- Tokenizing: segment a given text into word tokens
- Duplicate tweet checking: filter out the near-duplicate tweets by using Jaccard coefficient.
- Stopword: remove the Thai and English stopword

Generating Cluster labels

- This component will return the top-N tweets results from the Twitter corpus.
- The tweet results are cluster labels generated.

Label Merging

- Each cluster label is converted from abbreviated form to full word form.
- We automatically tag the POS of each word and remove the cluster labels which do not contain any noun.



Thai Tweet Corpus

NECTEC
a member of NSTDA

- We collected tweets from many Twitter profile pages.
- 160,000 tweets.
- We selected popular keywords under 12 topics such as flooding, survival kits and oil prices.
- Each tweet is cleaned and tokenized by using the Twitter post processing unit.

Experiment

- We apply the carrot2 framework to perform all the experiments.
- We used the default setting of carrot2 framework for generating cluster labels.
- We limited the maximum final cluster labels to 20
- We compare cluster label results between our approach and original Carrot2 by using precision, recall and F1 measure.

Performance Evaluation Metric

$$P = \frac{\text{number of returned cluster labels that are relevant}}{\text{number of returned cluster labels}}$$

$$R = \frac{\text{number of returned cluster labels that are relevant}}{\text{number of relevant number of cluster labels}}$$

$$F = 2 \times \frac{P \times R}{P + R}$$

Evaluation Example

Human	Carrot2 (Baseline)	STC-LM
ระดับ คัน กั้น น้ำ	อพยพ	ระดับ คัน กั้น น้ำ
คัน กั้น น้ำ บึก	ระดับ คัน กั้น น้ำ	คัน กั้น น้ำ บึก
กระสอบ ทราย	คัน กั้น น้ำ บึก	แนว คัน กั้น น้ำ
แนว คัน กั้น น้ำ	กระสอบ ทราย	คัน กั้น น้ำ คลอง
คลอง มหา สวัสดิ์	นนทบุรี	เหนือ คัน กั้น น้ำ
คัน กั้น น้ำ คลอง	แนว คัน กั้น น้ำ	ร็อ คัน กั้น น้ำ
เหนือ คัน กั้น น้ำ	คลอง	คัน กั้น น้ำ ริม
น้ำ กระสอบ ทราย	คลอง มหา สวัสดิ์	น้ำ กระสอบ ทราย
คลอง ประปา	คัน กั้น น้ำ คลอง	กระสอบ ทราย คัน
พัง คัน กั้น น้ำ	เหนือ คัน กั้น น้ำ	คลอง มหา สวัสดิ์
กระสอบ ทราย คัน		คลอง ประปา
ร็อ คัน กั้น น้ำ		เจ้า พระยา
คัน กั้น น้ำ ริม		นนทบุรี จังหวัด

P = 7/10 = 0.7
R = 7/16 = 0.43
F1 = 2*(0.7*0.43)/(0.7+0.43) = (

P = 12/13 = 0.92
R = 12/16 = 0.75
F1 = 2*(0.92*0.75)/(0.92+0.75) = 0.82

Precision	**0.7**	**0.92**
Recall	**0.43**	**0.75**
F-Measure	**0.53**	**0.82**

Cluster Label Evaluation Results

Topics	Baseline			STC-LM		
	P	R	F1	P	R	F1
ฟอปลาปู *(plaboo's father)*	0.900	0.642	0.750	1.000	0.785	0.880
น้ำท่วม *(flood)*	0.700	0.538	0.608	0.615	0.615	0.615
งูแมมบ้า *(mambas)*	0.900	0.642	0.750	0.900	0.642	0.750
ศปภ. *(flood center)*	0.700	0.583	0.636	0.666	0.666	0.666
กระสอบทราย *(sandbag)*	0.900	0.692	0.782	0.909	0.769	0.833
ถุงยังชีพ *(survival kits)*	0.600	0.461	0.521	0.777	0.538	0.636
บิ๊กแบ๊ก *(big bag)*	0.800	0.571	0.666	1.000	0.714	0.833
คันกั้นน้ำ *(dike)*	0.700	0.437	0.538	0.857	0.750	0.800
ซี๗ ตบป้องหู *(C7 assault)*	0.600	0.352	0.444	0.687	0.647	0.666
ราคาน้ำมัน *(oil prices)*	0.900	0.642	0.750	1.000	0.642	0.782
ปรับ ครม. *(cabinet shuffle)*	0.600	0.461	0.521	0.777	0.538	0.636
บริหารจัดการน้ำ *(water management)*	0.500	0.294	0.370	0.428	0.352	0.387
Average	**0.733**	**0.527**	**0.612**	**0.802**	**0.639**	**0.707**

Conclusion

- We present a new approach called Suffix Tree Clustering with Label Merging (STC-LM).
- STC-LM help generating more informative cluster labels.
- The main task of the STC-LM is to merge and create two-level label structure.
- 70% F1 measurement (improved 9%)

Future Works

- Considering timelines for cluster labels.
- Using incorporate probabilistic and machine learning into our label merging algorithm

Thank you for your attention

Questions / Comments ?