Natural Language Processing Issues in Social Computing

Virach Sornlertlamvanich
Sirindhorn International Institute of Technology, Thammasat University, Thailand

Abstract : The three fundamental issues in natural language processing (NLP), namely word segmentation, named entity recognition or keyword extraction, and semantic relation extraction are formalized to serve the task of knowledge map generation and social media text understanding. The extreme growth of social media text can be a useful source for tracking the social interests. The results can be served not only to observe the social movement but also to prevent the undesired situation. The paper introduces a set of approaches and shows their effectiveness in creating a meaningful set of knowledge in terms of knowledge map, and analyzing the social media text collected from Twitter to understand the social movement at a specific moment.

Keywords : NLP, word segmentation, named entity recognition, semantic relation extraction, knowledge map, social media text understanding

1. Introduction

The paper formalizes the three fundamental issues in natural language processing (NLP), namely word segmentation, named entity recognition or keyword extraction, and semantic relation extraction, which are crucial in handling the explosively increasing social media text. In the flood of information today, we spend most of the time to grasp the essence of the information rather than to enjoy the reading. Many approaches have been proposed to handle these fundamental issues, however, there is still much room for improvement. The paper discusses the problematic issues in text processing and aim to make the problems well recognized. In a non-segmenting language such as Thai, Lao, Cambodian, Chinese, or Japanese, word segmentation to determine the word boundary in a sentence is an introductory process of an input text. In case of the Thai language, it is reported that the accuracy of word segmentation is around 92%, 93% and 96% differently performed by the approaches of longest matching, maximal matching, and POS probabilistic tri-gram, respectively. Mutual information and entropy are effective measures to uncover the possible word boundary for the non-segmenting languages such as the Thai language. It is remarkably to note that with the approach, the result has shown that about 30% of the extracted words are not defined in the Thai-Thai dictionary published by Thai Royal Institute in 1982. Though the current version of the dictionary has already increased the number of words but it takes almost a month to define a new word. Keyword labeling is also a task that we can effectively apply a machine learning approach such as MIRA (Margin Infused Relaxed Algorithm) to capture the word surrounding context. This can be done on the result from the word segmentation task. Undoubtedly, the accuracy of the annotated tag is ranked descendently from person (PER), date (DAT), location (LOC), and organization (ORG). This is because tag for person has the least ambiguity. The pattern for extracting the semantic relation between the type-annotated keywords is accordingly assigned to the word form of the disambiguated verb phrase. The experimental result shows that most of the distance between the keyword and the target verb phrase is not more than one word. Therefore, we can find the target verb phrase in the adjacent position or one word skipped position with the highest probability.

Based on the solution for the above discussed NLP fundamental issues, many more tasks are made possible on the current viable Internet connection. The paper demonstrates the two constructive applications on the huge generated data, i.e. linked data formation for knowledge map reasoning; and keyword tracking on social media to understand the online social movement.

The task of natural language processing today is not just only for the language itself any more, but it can bring along the possibilities on the advance of the Internet, big data, and machine learning technique.

.

2. C4.5 Learning Algorithm for Word Extraction Task [1,2]

The induction algorithm proceeds by evaluating content of a series of attributes and iteratively building a tree from the attribute values with the leaves of the decision tree being the value of the goal attribute. At each step of learning procedure, the evolving tree is branched on the attribute that partitions the data items with the highest information gain. Branches will be added until all items in the training set are classified. To reduce the effect of overfitting, C4.5 algorithm [3] prunes the entire decision tree constructed. It recursively examines each subtree to determine whether replacing it with a leaf or branch would reduce expected error rate. This pruning makes the decision tree better in dealing with the data different from the training data.

We treat the word extraction problem as the problem of word/non-word string disambiguation. The next step is to identify the attributes that are able to disambiguate word strings from non-word strings. The attributes used for the learning

algorithm are as follows.

2.1 Attributes

1. Left Mutual Information and Right Mutual Information

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

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

where

Lm is the left mutual information
Rm is the right mutual information
x is the leftmost character of *xyz*
y is the middle substring of *xyz*
z is the rightmost character of xyz
p( ) is the probability function.

2. Left Entropy and Right Entropy

$$Le(\mathbf{y}) = -\sum_{\forall x \in A} p(x\mathbf{y} \mid \mathbf{y}) \cdot log_2 p(x\mathbf{y} \mid \mathbf{y})$$

$$Re(\mathbf{y}) = -\sum_{\forall z \in A} p(\mathbf{y}z \mid \mathbf{y}) \cdot log_2 p(\mathbf{y}z \mid \mathbf{y})$$

where

Le is the left entropy
Re is the right entropy
y is the considered string,
A is the set of all alphabets
x, z is any alphabets in *A*.

3. String Frequency

$$F(\boldsymbol{s}) = \frac{N(\boldsymbol{s})}{Sc}.Avl$$

where

s is the considered string
N(s) is the number of the occurrences of *s* in corpus
Sc is the size of corpus
Avl is the average Thai word length.

4. String Length
5. Function Words

Func(s) = *1* if string *s* contains a function word,
= *0* if otherwise

6. First and Last Two Characters

$$Fc(\boldsymbol{s}) = \frac{N(s_1 s_2 *)}{ND}$$

$$Lc(\boldsymbol{s}) = \frac{N(*s_{n-1} s_n)}{ND}$$

where

s is the considered string and s = $s_1 s_2 ... s_{n-1} s_n$

$N(s_1 s_2 *)$ is the number of words in the dictionary that begin with $s_1 s_2$

$N(*s_{n-1} s_n)$ is the number of words in the dictionary that end with $s_{n-1} s_n$

ND is the number of words in the dictionary.

2.2 Experimental Result

Table 1 shows that 31.5% of the extracted words are not found in the Thai Royal Institute dictionary (RID). These are the possible absent word entries from the Thai standard dictionary.

Table 1. Words extracted by the decision tree and RID

	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 (100.1%)	1046 (68.5%)	480 (31.5%)

3. Named Entity Recognition (NER) [4]

To evaluate our model, we obtained 33231, 20398, 8585, 2783 samples for PER, ORG, LOC, DATE, respectively. To ensure that our NE models work properly, we split samples into 90%/10% training/test sets and conducted some experiments. We trained our NE models using k-best MIRA (Margin Infused Relaxed Algorithm) [5]. We set k = 5 and the number of training iterations to 10. We denote the word by w, the k-character prefix and suffix of the word by $P_k(w)$ and $S_k(w)$, the POS tag by p and the NE tag by y. Table 2 summarizes all feature combinations used in our experiments. Our baseline features (I) include word unigrams/bigrams and NE tag bi-grams. Since we obtained the word boundaries and POS tags automatically, we introduced them gradually to our features (II, III, IV) to observe their effects.

Table 2. NE features

(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** $\langle P_k(w_0) \rangle, k \in [2, 3] \times y_0$ $\langle S_k(w_0) \rangle, k \in [2, 3] \times y_0$ $\langle P_k(w_0), S_k(w_0) \rangle, k \in [2, 3] \times y_0$

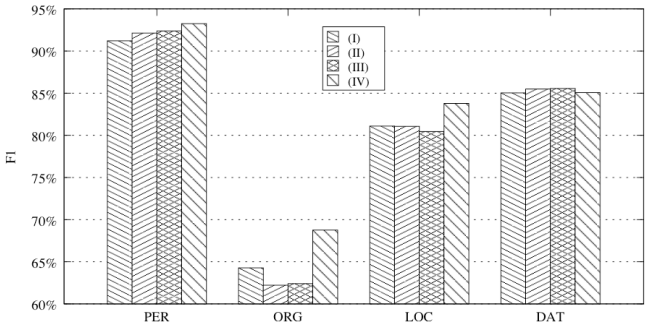


Figure 1. F1 evaluation results of the NE models

Figure 1 shows F1 results for the NE models. We observe that PER is easy to identify, while ORG is difficult. Pre-fix/suffix features dramatically improve performance on ORG. Using all features (IV) gives best performance on PER (93.24%), ORG (68.75%) and LOC (83.78%), while slightly

drops performance on DATE (85.06%). Thus, our final NE models used in relation extraction are based on all features (IV).

4. Semantic Relation Extraction (SLR) [6]

Mapping text segments containing a given relation surface (e.g., "สร้างโดย" (is built by)) in a large database is not a trivial task. We applied surface-relation mapping as shown in Table 3. Here, we use Apache Solr for indexing and searching the database. Apache Solr works well with English and also has extensions for handling non-English languages. To process Thai text, one just enables Thai-WordFilterFactory module in schema.xml. This module invokes the Java BreakIterator and specifies the locale to Thai (TH). The Java BreakIterator uses a simple dictionary-based method, which does not tolerate word boundary ambiguities and unknown words. For example, the words "สร้าง" (build) and "ก่อสร้าง" (construct) occur in the Java's system dictionary. Both convey the same meaning (to build). We can see that the first word is a part of the second word. However, these two words are indexed differently. This means if our query is "สร้าง" (build), we cannot retrieve the records containing "ก่อสร้าง" (construct). In other words, the dictionary-based search returns results with high precision but low recall.

Table 3. Relation template (LOC: location; PER: person; ORG: organization; DATE: date)

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

Table 4. Numbers of relation instances when the distances are varied

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

Table 4 shows the numbers of relation instances when the distances are varied. For all relations, we observe that the numbers of relation instances do not significantly change after one word distance.

5. Knowledge Map Generation

Relations between NE (or keyword) are successfully extracted by NER and SLR approaches. The accuracy is acceptably high, ranging from 85% to 100% corresponding to the type of the relation. The tuples of relation are stored attaching to the record they belong to. Though the tuple of semantic relation is extracted from a part of the description, it determines the semantic modification to the title of the record. From the set of tuples of each record, the infobox of the record is generated to express the essence of the title we are looking for. NE's are used to modify the title which is also included in the set of NE. By mapping the NE found in the database, we can extensively trace the semantic modification of any target NE. Finally, the knowledge map, which is a network of the NE can be express to understand the relation among all NE's in the database.

Figure 2. shows the tuples of semantic relation extracted from the record of Phra Samut Chedi, i.e.

ISBUILDIN(พระเจดีย์กลางน้ำ, พ.ศ. 2403)

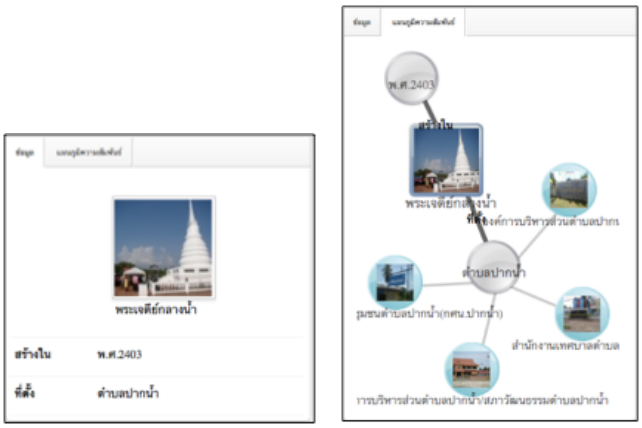
(Lit. ISBUILDIN(Phra Samut Chedi, BE 2403)), and

ISLOCATEDAT(พระเจดีย์กลางน้ำ, ตำบลปากน้ำ)

(Lit. ISLOCATEDIN(Phra Samut Chedi, Tambon Paknam)).

In the infobox as shown in Figure 3(1), it notifies when and where the Phra Samut Chedi was constructed. The summary information about the record in the form of infobox can help the audience to grasp the information about the record in quick. By knowing that the pagoda (Chedi) was founded in Tambon Paknam, we can trace further for what else are related to the NE of Tambon Paknam. The example of the knowledge map expression is shown in Figure 3(2). The audience can traverse for other related information about the focus topic and understand the relation among the records. Further level of relation can be expended as far as they are connected with the extracted tuples of semantic relation.

Figure 2. Tuples of semantic relation extracted from the record of Phra Samut Chedi.



(1) Infobox (2) Knowledge map

Figure 3. Infobox and knowledge map extracted from the cultural database for the record of Phra Samut Chedi

6. Social Movement Understanding [7]

The initial Word-Article matrix (i-WAM) is generated from online news document. It cannot precisely reflect the context of social media text such as tweets from Twitter. The words and the value of TF/IDF of the words from social media text can be different from the ones from general text. Figure 4 shows how to modify i-WAM for social media text classification.

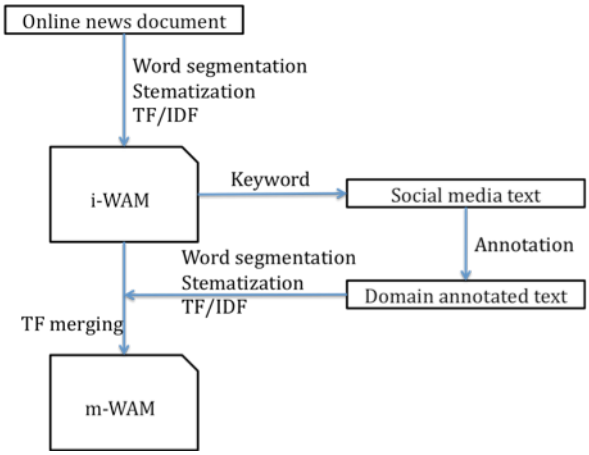


Figure 4: Generating m-WAM

In the initial step, we use the list of word from the i-WAM to collect the tweets by expecting the tweets of the domain in question, i.e. life, education, and technology. The additional training set for tuning the i-WAM is then generated from the tweets manually. Additional terms are selected according to their TF/IDF value. In merging process into the i-WAM, for the existing words in the i-WAM, the additional count is added to recomputed for the term frequency. For the newly found words, the term frequency among the tweets are computed and added into the i-WAM.

As a result, the modified WAM (m-WAM) to fit the social media text can be generated.

We conducted the experiment on the topic of Thailand coup d'etat declaration on May 22, 2014. We acquired the related tweets using the keywords from WAM and then classified the tweets to the topic in question. As a result, 339,148 tweets centering on the date of coup d'etat declaration on May 22, 2014 are collected. Figure 5 shows the procedure in extracting keywords from related tweets to represent in word cloud manner.

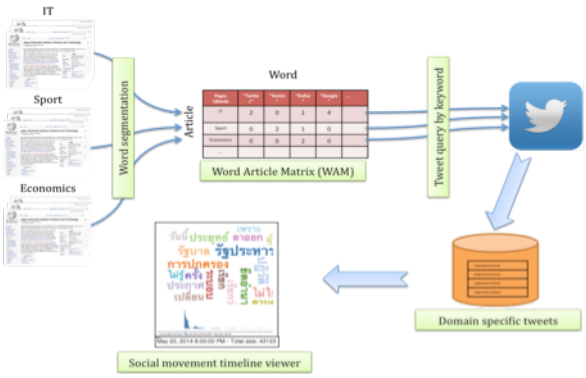


Figure 5: Keyword extraction from related tweets

The figure shows a significant expression of the keywords related to the context of coup d'etat. The keywords are growing to the peek time of the announcement and then keep stable for some hours. This means that the people on the social network were already aware of the coup d'etat before the peek time of the coup d'etat declaration.

7. Conclusion

Under the situation of extreme growth of social media text data, NLP fundamental tools, namely word segmentation, named entity recognition, and semantic relation extraction are crucial. High performance algorithms are necessary to understand the social media text in real-time. The results are not only used to observe the social movement but they also can be served to prevent the undesired development of a specific situation.

References

[1] V. Sornlertlamvanich, and H. Tanaka, "The Automatic Extraction of Open Compounds from Text," Proceedings of COLING, 1996.

[2] V. Sornlertlamvanich, T. Potipiti and T. Charoenporn, "Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm," Proceedings of COLING, 2000.

[3] J. R. Quinlan, "C4.5 Programs for Machine Learning," Morgan Publishers San Mated, California, 1993.

[4] V. Sornlertlamvanich and C. Kruengkrai, "Effectiveness of Keyword and Semantic Relation Extraction for Knowledge Map Generation," Proceedings of The Second International Workshop on Worldwide Language Service Infrastructure (WLSI), 2015.

[5] K. Crammer, R. McDonald, and F. Pereira, "Scalable large-margin online learning for structured classification," Proceedings of NIPS Workshop on Learning With Structured Outputs, 2005.

[6] C. Kruengkrai, V. Sornlertlamvanich, W. Buranasing, and T. Charoenporn, "Semantic Relation Extraction from a Cultural Database," Proceedings of Workshop on SANLP, COLING, 2012.

[7] V. Sornlertlamvanich, E. Pacharawongsakda and T. Charoenporn, "Understanding Social Movement by Tracking the Keyword in Social Media," Proceedings of Multiple Approaches Lexicon (MAPLEX), 2015.