Social-based Traffic Information Extraction and Classification

Napong Wanichayapong, Wasawat Pruthipunyaskul, Wasan Pattara-Atikom, Pimwadee Chaovalit
National Electronics and Computer Technology Center (NECTEC)
National Science and Technology Development Agency (NSTDA)
Klong Luang, Pathumthani 12120 Thailand
plagad@gmail.com, wasawat.pruthipunyaskul@nectec.or.th, wasan@nectec.or.th, pimwadee@nstda.or.th

Abstract— **Social networks such as Twitter and Facebook are popular, personal, and real-time in nature. We found that there exists a significant number of traffic information such as traffic congestion, incidents, and weather in Twitter. However, an algorithm is needed to extract and classify the traffic information before publishing (re-tweeting) and becoming useful for others. Traffic information was extracted from Twitter using syntactic analysis and then further classified into two categories: point and link. This method can classify 2,942 traffic tweets into the point category with 76.85% accuracy and classify 331 traffic tweets into the link category with 93.23% accuracy. Our system can report traffic information real-time.**

Keywords-component; real-time; twitter; traffic information; classification; syntactic analysis;

I. INTRODUCTION

Twitter has become a very popular micro-blog social network. It has been used as real-time text information dissemination. Tweetple or Twitter users are tweeting (to send text to display on profile page) in average of 55 million tweets (text-based posts composed of up to 140 characters) a day and 37 percent of Twitter's active users use their phones to tweet. In Thailand, there were 1,191,760 tweets from 79,705 Twitter users in a single day (April 19, 2011), and 36.22% of them tweet from their phones (count form clearly Twitter mobile application only). These statistics show us that more than one-third of Twitter users are sending real-time messages of interesting events, such as weather reports, accidents, and traffic conditions. As for traffic-related messages alone, we found during a period of April 19, 2011-April 30, 2011 that there are about 2000 daily tweets that relay traffic information (e.g. traffic condition, accident) in Thailand.

Twitter is a sharing community. Tweetple usually tweet or retweet to share useful information of their interests with others, and follow other Tweetple who interested in same thing. Most of road-user Tweetple are interested in traffic events, e.g. road closures, traffic congestion, and accidents. They will tweet to let their followers know the traffic events that just happened. If we can capture that useful traffic information, we will be able to retweet them to a larger audience and make them more useful.

Tweetple can provide traffic reports in real-time. We used Traffy, the Twitter account for broadcasting traffic news in Bangkok, Thailand, to tweet and retweet traffic reports from Tweetple. The real-time traffic reports can help many people in Bangkok residents and travelers plan the routes and avoid traffic congestion.

The purpose of syntactic analysis is to determine the structure of the input text. This structure consists of a hierarchy of phrases. We classify traffic information into point and line categories via syntactic analysis concept. So, we must tokenize traffic information with Lexto, Thai Lexeme Tokenizer by Sansarn, before analyzing it.

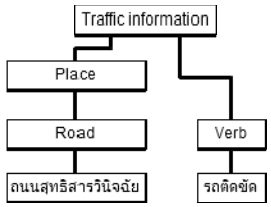
This paper is organized as follows. Section II outlines the extraction and classification using syntactic analysis techniques. Section III outlines the methodology in extracting and classifying traffic information. Section IV performance evaluation include data description and result discussion. Finally, section V concludes this paper with discussion and some future directions.

II. LITERATURE REVIEW

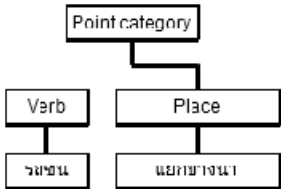
Twitter contains a lot of useful information. We can extract interesting information from Twitter such as Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo from The University of Tokyo who published "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors" in 2010 [3]. Divij Gupta and Chanh Nguyen from Stanford University published "Detecting Real-Time Messages of Public Interest in Tweets" in 2010 [4]. Real-time traffic information such as that obtained from social networks helps users avoid traffic congestion, better plan the routes, and potentially save fuel costs.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo used tweets from Twitter for detecting earthquake and typhoon events and estimating locations for those events in Japan. For detection, they regarded tweets as event sensors. The tweets were classified into 2 groups of positive class and negative class. They used Support Vector Machine (SVM) [5] which is a widely used machine-learning algorithm. For location estimation each tweet was associated with a location. Then, they applied Kalman filters and particle filters which are widely used in the field of location estimation. As an application, an earthquake reporting system was developed to notify people promptly of an earthquake event.

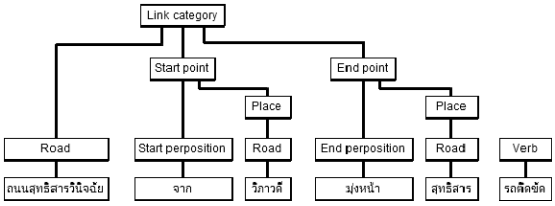
Divij Gupta and Chanh Nguyen received their inspiration from helping people who are new in town look for promotions or event updates. They compared various classification concepts, e.g. Naive Bayes, Logistic Regression, SVM, and Boosted Decision Trees.



Fix.1 Traffic information parse tree



Fix.2 Point Traffic information parse tree



Fix.3 Link Traffic information parse tree

III. METHODOLOGY

We classified traffic information into 2 types: (1) point and (2) link. First, point information associates with only one point (e.g. a car crash at a crossroad), as shown in Fig.4, while link information associates with a road start point and an end point (e.g. a traffic jam between two squares), as shown in Fig.5.

We want to capture and extract traffic information from all accounts from social networks. After the information was extracted it can be made available for others, for example in this case, via re-tweeting. The information should be complete enough in order to be useful; hence it should state two important pieces of information of the traffic events: "what" and "where". The extracted traffic events should be informative, factual, and definitive. In other words, they inform and do not request for information from others. Moreover, the traffic information should by no means contain any vulgarity or profanity.

Traffic information is usually consisted of various word categories. First, Place is set of places, roads, crossroads, alleys and zones name. Second, Verbs is set of traffic problems, vehicle incidents/accidents, street closures, lane restrictions, road works, obstruction hazards, road conditions, weather conditions, traffic activities, traffic light statuses, and traffic regulations. Third, Preposition come before place and are used to indicate spatial relationships in traffic. In this paper, Preposition is divided into two categories: Start preposition and End preposition. When combined with Place, Start prepositions and End prepositions indicate the start locations and end locations of traffic events. Finally, Ban, natural language texts like those found in tweet messages, at times contain words with profanity or vulgarity, which could lead to the banning of those tweets from being broadcasted again. Traffic tweets which are truly traffic information will have places, roads, crossroads, alleys or zones and traffic problems (e.g. accidents, road works) but by no means to question or to have vulgarity or profanity. The concept above leads to important definitions, as described below.

Definition 1: **Place** is word set of building, **road**, crossroads, alley and zone name.

Definition 2: **Road** is subset of **Place**.

Definition 3: **Verb** is word set of Traffic Event, e.g. traffic problem, accidents, road works.

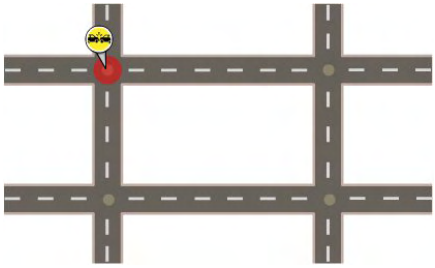
Definition 3: **Ban** is word set of vulgarity/profanity and word that means to question

Definition 5: **Start and End preposition** is set of words that means Place after it is start or end point of traffic event.

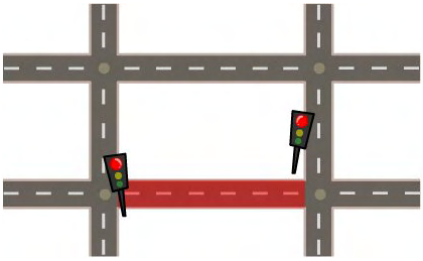
Definition 6: **Traffic information** is information that has Place and Verb word but not have Ban word.

Definition 7: **Link category traffic information** is traffic information that clearly has road, start and end point (e.g. building, crossroads, alley or zone).

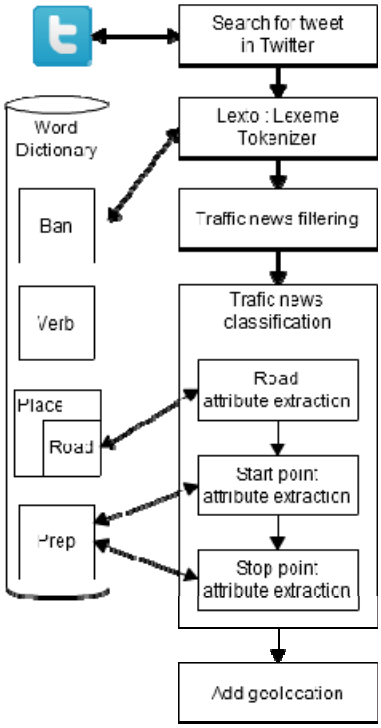
Definition 8: **Point category traffic information** is traffic information that has road and start point or road and end point. And, **Point category traffic information** is traffic information that has only one **Place**.



Fix.4 Point Category



Fix.5 Link Category



Fix.6 Process flow

A. *Search for traffic tweets in Twitter*

First, we limited the domain of tweets from which we were extracting and classifying traffic information. We chose Thai words that describe traffic conditions, such as "รถติด" (traffic congestion) and "อุบัติเหตุ" (accident) as search keywords. We used keywords to search in Twitter via Twitter search API at http://search.twitter.com/search.atom?q=X, where **X** is a keyword. Results from the search are public tweets that are related to the keywords. To automate the searching of Twitter messages, a Cron job was run at 5-minute intervals. We performed incremental searches by checking with Twitter message identifiers to prevent duplicate messages. On average, we obtained about 20 messages for each incremental search.

B. *Text Tokenization using Lexto and dictionary*

1) *Dictionary*

Dictionary is a very important tool for extracting information. It works in cooperation with Lexto [6] such that it provides words for Lexto to tokenize tweet messages. We built a special traffic word dictionary of four categories: (1) Place, (2) Verb, (3) Ban, and (4) Preposition.

Place dictionary consists of 46,241 names of roads, places, crossroads, and alleys. Verb dictionary have 1,093 of words or phrases of traffic conditions, such as "รถติด" (traffic jam). Ban dictionary have 149 words of vulgarity, profanity, and question words. Preposition dictionary have 192 words of direction of roads, which indicates the starting and ending of roads. All numbers of words in the dictionary were as of April 28, 2011

2) *Lexto*

Lexto is a dictionary-based tokenizer with lexical analysis for Thai language, provided by Sansarn [7]. We tokenized traffic information in order to further classify it using a syntactic analysis concept.

C. *Traffic news filtering*

During this step, we needed to decide on which Twitter messages are considered real traffic tweets. Even though the domain of tweets were once refined by the previous search for only tweets that contain traffic keywords such as "รถติด" (traffic congestion) and "อุบัติเหตุ" (accident), they were only traffic news candidates until further proof. We tokenized the messages and parsed the tokens into the same four categories of words, i.e. Place, Verb, and Ban, and Preposition, as the categories from the dictionary. Then, a simple heuristics was applied for filtering text messages. The messages which will be considered traffic information must pass 2 filtering rules that are described below:

a) The messages must contain at least 2 words, one of Place and other one of Verb categories. Both types of words are essential for being traffic news, otherwise we cannot know where the incident is or what did just happen.

b) The messages must not contain any Ban words. Some text messages could have been traffic tweets if not for the rudeness displayed or question words.

A few examples of text messages which are considered traffic news are listed in table I. The first text message means “traffic jam”. This information is obviously traffic-related but we can never know from the little information where this event happened. Therefore, the first message will not serve as a good piece of traffic news. The second message means “Is there a traffic jam at Din-dang road?”. The sentence has both Place and Verb words, so we know what happened with the place mentioned in the message. However, the message has a question word thus the sentence must be filtered out. Finally, the last text message means “There is a collision between a van and a tricycle at Kalai intersection with some injury”. This text message lets us know what happened and where. Therefore, it is a perfect message to be considered traffic information.

TABLE I. TRAFFIC NEWS FILTERING SAMPLE

Messages	Place	Verb	Ban
รถติด	-	รถติด	-
ถนนดินแดงรถติดหรือเปล่าครับ	ถนนดินแดง	รถติด	หรือเปล่า
แยกแครายมีอุบัติเหตุรถตู้ชนกับรถซาเล้งมีคนเจ็บ	แยกแคราย	ชน	-

D. Traffic news classification

The next step in our methodology was to extract Road, Start point, and End point to classify traffic information.

1) Road attribute extraction

We compared each token with the Road word of Place dictionary. If that token matches a road entry in database, that token will be marked as a road attribute for this traffic information. Then, we parsed the rest of traffic information tokens to the next step: start point attribute extraction.

2) Start point attribute extraction

We searched for a start point attribute by comparing each word with the Start preposition and Place dictionary. If that word matches a Start preposition in the dictionary and its following word matches a Place in the dictionary, that word that matches a Place will be marked as a start point attribute of this traffic information. Then, we parsed the rest traffic information tokens to the next step: end point attribute extraction.

3) End point attribute extraction

Similarly to start point attribute extraction, we searched for an end point attribute by comparing each word with the End preposition and Place dictionary. If that word matches an End preposition in the dictionary and its following word matches a Place in the dictionary, that word that matches a Place will be marked as an end point attribute of this traffic information.

4) Classification

Link information is the traffic information that has all the attributes: Road, Start point and End point. Point information is the traffic information that has Road and Start or Stop point attribute or only Road or Start point attribute or Stop attribute, as shown in table II.

TABLE II. TRAFFIC NEWS FILTERING SAMPLE

Message	Road	Start	End	Category
สุขุมวิทขาออกข้ามแยกบางนามาหน้ากรมอุตุฯติด	สุขุมวิท	แยกบางนา	กรมอุตุฯ	Link
แจ้งวัฒนะฝนตกแล้วจร้า	แจ้งวัฒนะ	-	-	Point
ส.กรุงธนฯเข้าท้ายชะลอตัวในถนนสิรินทร	ส. กรุงธนฯ	-	ถนนสิรินทร	Point
ส.กรุงธนฯ มุ่งหน้าอนุสาวรีย์ฯ รถไล่ง	ส. กรุงธนฯ	อนุสาวรีย์ฯ	-	Point

E. Add geo-location

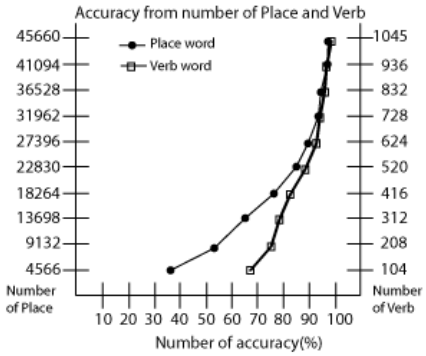
The last step in our methodology is adding geo-location (latitude and longitude) of Start point attribute and End point attribute in the text messages. Geo-location will help visualizing the traffic information on the map. To find latitude and longitude of each place, we searched in the place dictionary, which contains latitude and longitude obtained from the Ministry of transportation, Thailand (MOT). If the place was not found in the place database, we then used Google geocoding to find that. Google geocoding is a Google API that can search for latitude and longitude of a place. Sometimes Google geocoding may return multiple results. We were able to filter results by country (selecting only Thailand) and chose the first result, which by default was the most relevant from Google. If Google geocoding was able to find that, we update our place database for that record. Using this approach, our place database has latitude and longitude information from both Google and MOT.

IV. EXPERIMENTS

In this section, we present experiment results of: (1) Lexto tokenization accuracy, (2) Traffic text message filtering accuracy, (3) Traffic news classification accuracy, and (4) Polar coordinate finding Accuracy.

A. *Lexto tokenization accuracy*

First, our text tokenization method was measured for its performance. In order to evaluate the accuracy of text tokenization by the number of words added into our dictionary, we chose 986 text messages from Twitter and 45,660 words of Place dictionary and 1,045 of Verb dictionary, as of April 28, 2011. We randomly split both Place and Verb words from dictionary into 10 parts, ranging from 4,566 words to 45,660 words for Place and from 104 words to 1,045 for Verb. Then, we used Lexto to tokenize the text message to measure its tokenization accuracy. As there can be several cut points within one message, we only counted a message as being accurately tokenized if and only if all cut points were accurate, otherwise we considered it inaccurately tokenized.

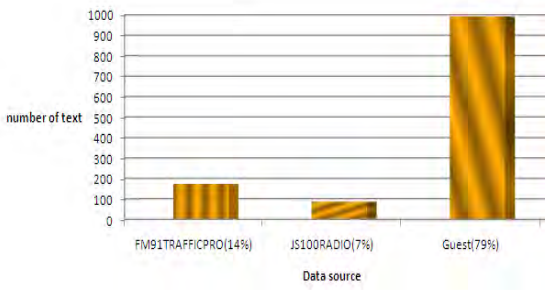


Fix.7 Lexto tokenization accuracy

We found from Fig.7 that adding the first 10% of place words (4,566 words) gave the accuracy of LEXTO tokenization of 38%. By adding each 10% more, the accuracy of text tokenization increased to 52%, 65%, 76%, and finally reaching 97%. Simialrly, adding the first 10% of verb words (104 words) gave the accuracy of LEXTO tokenization of 68%. By adding each 10% more, the accuracy of text tokenization increased to 75%, 77%, 82%, and finally reaching 98%. The result shows that adding words into the dictionary at the beginning is especially useful for text tokenization but when the words in dictionary are enough, adding words into the dictionary hardly increased the performance of text tokenization. Finally, verb words have a bigger impact on the performance of text tokenization than place words.

B. *Traffic text message filtering accuracy*

For the evaluation of traffic text message, first we prepared a dataset of 1,249 Twitter messages. The data came from two sources: traffic information center (21%) and individual users called "guest" (79%), as shown in Fig.8. The sources of traffic information center were divided into FM91TRAFFICPRO (14%) and JS100RADIO (7%). In our test, we used human interpretation as the benchmark. A team of three research assistants was asked to manually extract the traffic messages and make the decision as to whether the messages are related to traffic. Then, the human-generated output is compared the result from our method.



Fix.8 The number of text messages for filtering, shown by users

TABLE III. TRAFFIC TEXT MESSAGE FILTERING CONFUSION MATRIX

	Predicted		
Actual	True	False	Total
True	435	62	497
False	41	711	752
Total	476	773	1249
Accuracy	91.75%		
Precision	91.39%		
Recall	87.53%		

Table III shows the accuracy of our filtering method for traffic information from text messages. Messages were flagged True if related to traffic and False if not related to traffic. The method can filter the messages with the accuracy of 91.75%. We note that most errors occured because there were some English words in the text messages that the proposed method currently was not designed to capture, and the algorithm did not understand those words as much as humans did.

C. *Traffic news classification accuracy*

To measure point and link classify algorithm accuracy, Two research assistants were asked to manually categorize 3,311 news (point, link, unclassified). Their decisions were made by referring to the definitions provided by the authors. The classification results obtained from the proposed algorithm were calculated for accuracy using the research assistants' results as human judgment baseline. Note that the total input news for point classification algorithm were news that were not classified into link category by link classification algorithm.

This method can classify traffic information into the point category with 76.85% accuracy of 2,942 test messages and into the link category with 93.23% accuracy of 3,311 test messages.

TABLE IV. LINK CATEGORY CONFUSION MATRIX

	Predicted		
Actual	True	False	Total
True	349	207	556
False	17	2738	2755
Total	366	2945	3311
Accuracy	93.23%		
Precision	62.77%		
Recall	95.36%		

TABLE V. POINT CATEGORY CONFUSION MATRIX

	Predicted		
Actual	True	False	Total
True	2209	494	2703
False	187	52	239
Total	2396	546	2942
Accuracy	76.85%		
Precision	81.72%		
Recall	92.20%		

Most of the errors occurred because there were more than one point in a single news, in which case the algorithm classified it into the link category. Moreover, unknown places in the dictionary and user-mistyping caused the misclassification of link category to point category and the misclassification of point category to unclassified messages.

D. Polar coordinate finding Accuracy

Geocoding accuracy was measured by pinning the latitude-longitude coordinates on the Google map. Those pin locations were then checked by two research assistants. For this experiment, traffic news data came from the classification phase.

All news contains 2,857 places in total, but only 295 unique places. The accuracy for geocoding traffic news is 92.20%. Most errors occur because some places have the same name.

V. CONCLUSION

Twitter contains a lot of useful information. We extracted traffic information from Twitter and classified it. Then, we broadcasted that useful information to help many people plan their routes, avoiding traffic congestion in real-time. This effective classification algorithm can be applied to traffic information from other social networks, not only for Twitter.

Dictionary is important for this algorithm because the number of words in dictionary affects the performance of the tokenizer. More dictionary words gave more accuracy to the tokenizer performance. For classification, more place words gave more accurate classification.

We have applied syntactic analysis techniques to perform the extraction and classification of traffic news. This method can be improved by enhancing dictionary and geocoding technique. We can also augment the algorithm's ability to handle misspelt place names, and to guess probable place words in order to add those words to dictionary. Finally, the geocoding can be adjusted to handle places that have the same name. For example, the word “monument” may refer to a few monuments in Bangkok, but can be further refined by surrounding contexts.

REFERENCES

[1] Twitter User Statistics REVEALED, http://www.huffingtonpost.com/2010/04/14/twitter-user-statistics-r_n_537992.html

[2] Thailand Trending (Thailand twitter trending), http://www.lab.in.th/thaitrend/

[3] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, 2010.

[4] Divij Gupta and Chanh Nguyen. Detecting Real-Time Messages of Public Interest in Tweets, 2010.

[5] Thorsten Joachims, Text categorization with support vector machine:Learning with many relevant features.LS-8 Report 23, Computer Science Departnebt. University of Dortmund, 1997.

[6] Lexto, Thai Lexeme Tokenizer by Sansarn, http://www.sansarn.com/lexto/

[7] Sansan, Search engine with Thai lexeme tokenization by NECTEC, http://www.sansarn.com/

[8] Paul Klint, Syntax analysis, 2007. http://homepages.cwi.nl/~daybuild/daily-books/learning-about/syntax-analysis/syntax-analysis.htm