Thai Word Segmentation on Social Networks with Time Sensitivity

Chirawan Ronran, Sayan Unankard, Wanvimol Nadee, Nongkran Khomwichai and Rangsit Sirirangsi

Maejo University, Thailand, {chirawan;sayan; wanvimon;nongkran_it, rangsit}@mju.ac.th

ABSTRACT

Social network service like Twitter is one of the important social networks that has had a huge impact on Thai culture. It has changed the behavior of many Thai people from using televisions to using computers or smart phones regularly. Thai people also share their experiences and get information such as news on social networks. With the increasing number of micro-blog messages that are originated and discussed over social networks, Thai word segmentation is becoming a compelling research issue as it is an important task in natural language processing. However, the existing Thai segmentation approaches are not designed to deal with short and noisy messages like Twitter. In this paper, we proposed Thai word segmentation on social networks approach by exploit both the local context (in tweets) and the global context from Thai Wikipedia. We evaluate our approach based on a real-world Twitter dataset. Our experiments show that the proposed approach can effectively segment Twitter messages over the baseline.

Keywords: Thai Segmentation, Tokenization, Social Network, Time Sensitivity.

I INTRODUCTION

In the present age, social networks have become the most popular way of communication for the current generation. The number of social network activities has increased dramatically, for example, information sharing, daily conversation and spreading news. Social network services provide a wealth of current topics which are discussed in social networks communities. Micro-blog like Twitter is being considered as a powerful means of communicating for Thai people looking to share and exchange information on a wide variety of topics. In 2015, the service rapidly gained worldwide popularity, with more than 4.5 million users who posted 3.4 million tweets in Thai[7].

The fast information sharing on Twitter from millions of users all over the world leads to almost real-time reporting of events or topics (Li et al., 2012; Mathioudakis and Koudas, 2010; Unankard et al., 2015). This strong temporal nature of shared information allows for the detection of significant events in the data stream. Therefore, before we can successfully identify events in social networks, we must understand how to segment Thai word from Twitter as it is an important task in natural language processing. Dealing with Thai language is more complex than English. Thai language does not have any explicit word boundary delimiters. Existing studies have focused on using a dictionary-based approach (Poowarawan, 1986; Sornlertlamvanich, 1993) however, the results rely on dictionary they have. On the other hand, using machine learning based approach has been studied in different ways (Haruechaiyasak et al., 2008; Limcharoen et al., 2009; Manning and Schutze, 1999). However, the approaches relies on having labelled data to train a classifier and it is not clear if retraining the classifier is needed.

Due to the characteristics of micro-blog messages in Thai, abbreviations and slang words are widely used in a message which cannot found in Thai dictionary (i.e., unknown word problem). Thus, we cannot rely on dictionary based approach. Also, labelled data is very expensive and time consuming process for training the model when supervised learning is applied. Therefore, the challenges of this paper are follows: (1) how to effectively segment Thai words in micro-blogs? (2) how to incorporate the local context (in Twitter) and the global context (from Thai Wikipedia) for Thai word segmentation task?

To our best knowledge, this paper is the first to fully focus on Thai word segmentation in social networks. The main contributions of this paper are as follows. (1) We present an approach to segment Thai word in micro-blogs (i.e., Twitter). (2) Local and global contexts are incorporated to improve the Thai segmentation. We evaluate our proposed approach with a real-world Twitter data posted by Thai-based users.

The rest of the paper is organized as follows. First, we describe the related work in Section II. Second, the proposed approach is presented in Section III. Third, we present the experimental setup and results in Section IV. Finally, the conclusions are given in Section V.

[7] http://syndacast.com/infographic-online-marketing-thailand-the-state-of-social-media/

II RELATED WORKS

Typically, word or text segmentation play importance roles in natural language processing (NLP). The concept of word segmentation is applied in different languages such as English, Thai, Chinese and Japanese. In this paper, we focus on finding a method to improve Thai word segmentation (Haruechaiyasak et al., 2008a). In general, a sequence of Thai words in sentence is written similar writing an English sentence. However, the processing text segmentation in Thai language is not easy as English Language. Due to Thai language does not use any delimiter to specify the explicit word boundary. It makes the word boundaries are ambiguous. Consequently, the meaning of words and phrases could be different from the meaning of its part. Other problems happen when a new word is formed by combining a few words into a compound word. This situation does not only generate the ambiguity problem, but it also generates the unknown word problem and the new word problem in input text (Limcharoen et al., 2009). The systems cannot find these words in dictionary, thereby the segmentation results may not be accurate.

Recently, there are many works related to Thai word segmentation tasks. They try to develop the algorithms or techniques to find Thai word boundaries to make the better segmentation results. In previous works, most word segmentation approaches rely on two main approaches: dictionary based and machine learning based. Dictionary based approaches use a set of words or terms from a dictionary for making word segmentation (Aroonmanakun et al., 2007; Phaholphinyo and Kosawat, 2011). Therefore, this approach requires making a list of words in advance. Poowarawan (1986) proposed the longest matching algorithm which based on dictionary based approach to solve the ambiguity of words (Poowarawan, 1986). In addition, the ambiguity can be solved by using the maximum matching algorithm which splits a sequence of characters prior to segmentation based on a word set (Haruechaiyasak et al., 2008a; Sornlertlamvanich, 1993). However, the dictionary based approaches cannot handle the unknown word, or new word and ambiguity problems without adding these words into dictionary. To address the problem of unknown words for Thai language, the rule based is employed to build the segmentation techniques for a new word (Kawtrakul et al., 1997; Mahatthanachai et al., 2015; Palmer, 1997). The new words were created by combing the rule-based of characters and the rule-based of unknown words, but this approach is unable to wrap words when there were verbs appear between two unknown words. However, the rule-based approach cannot provide the high accuracy and requires hand-crafted rules resource (Khankasikam and Muansuwan, 2005).

Some studies suggested that the machine learning based approaches can improve the performance of the dictionary based approaches when these two problems exist in the systems (Haruechaiyasak et al., 2008a, b, 2006; Peng et al., 2004). Most machine learning based approaches algorithms are built under the statistical language modeling (LM) such as n-gram model (Manning and Sch¨utze, 1999) and feature-based segmentation (Meknavin et al., 1997). N-gram model has been successful applied to many word segmentation problems (Silva et al., 1999). The models identify the word boundaries based on the feature of the characters surrounding the boundaries (Haruechaiyasak et al., 2008a; Limcharoen et al., 2009). Limcharoen et al. proposed a Thai word segmentation framework based on the combination of the concept of Thai Character Cluster (TCC) and word N-gram model to reduce the number of candidates and generate all possible word segmentation candidate (Limcharoen et al., 2009). This method does can be implemented without a dictionary for making word segmentation. Later, Theeramunknog et al. utilized the TCC to learn the word segmentation without dictionary by using the decision tree (Theeramunkong and Usanavasin, 2001).

However, some researcher argued that word segmentation is not the actual cause of the ambiguity problem, but this problem occurs from syllable segmentation (Aroonmanakun, 2002). Aroonmanakun proposed another word segmentation approach based on a syllable-based trigram model and maximum collocation. Author used a trigram model for syllable segmentation and determine word boundary and group syllables into words based on the idea of collocation (Aroonmanakun, 2002). Khankasikam et al. proposed another method to reduce the ambiguity problem by taking the semantics of words into consideration when making word segmentation (Khankasikam and Muansuwan, 2005).

To our best knowledge, this paper is the first to fully focus on Thai word segmentation in social networks. The main contributions of this paper are as follows. (1) We present an approach to segment Thai word in micro-blogs (i.e., Twitter). (2) Local and global contexts are incorporated to improve the Thai segmentation. We evaluate our proposed approach with a real-world Twitter data posted by Thai-based users.

III PROPOSED APPROACH

In order to provide a complete coverage of Thai word segmentation in social networks we proposed our system which has three stages presented in Figure 1. The following information provides detail of each stage.

A. Data Pre-processing

In particular dealing with micro-blog messages, the message is short and often noisy. In order to improve the quality of our dataset and the performance of the subsequent steps, the pre-processing was designed to remove irrelevant data e.g., re-tweet keyword, web address and message-mentioned username. A microblog loader is developed to collect the Twitter data from public users via the Twitter API service. The messages are removed by web addresses and the keyword RT(ReTweet) and the message-mentioned username such as “@username”.

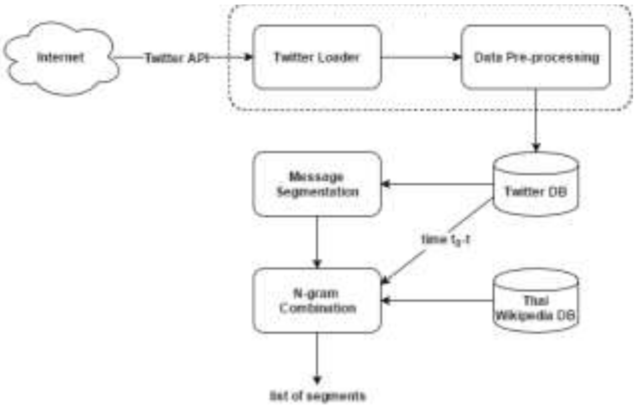
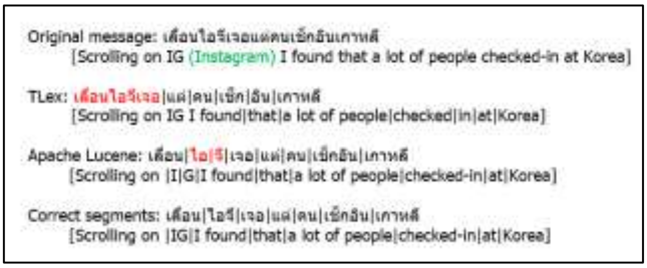


Fig. 1. Architecture Of Our System

B. Micro-blog Message Segmentation

The problem that we address in this section is how to segment Thai word in micro-blogs. Dealing with Thai language is more complex than English. Thai language does not have any explicit word boundary delimiters. The existing methods are unsuitable for dealing with Thai segmentation in micro-blog services. The example can be seen in Fig 2.



Original message: เลื่อนไอจีเจอแต่คนเช็กอินเกาหลี
[Scrolling on IG (Instagram) I found that a lot of people checked-in at Korea]

TLex: เลื่อนไอจีเจอ|แต่|คน|เช็ก|อิน|เกาหลี
[Scrolling on IG I found|that|a lot of people|checked|in|at|Korea]

Apache Lucene: เลื่อน|ไอ|จี|เจอ|แต่|คน|เช็กอิน|เกาหลี
[Scrolling on |I|G|I found|that|a lot of people|checked-in|at|Korea]

Correct segments: เลื่อน|ไอจี|เจอ|แต่|คน|เช็กอิน|เกาหลี
[Scrolling on |IG|I found|that|a lot of people|checked-in|at|Korea]

Fig. 2. The example of existing Thai segmentation methods

Thai Lexeme Analyser (*TLex*) is an approach using Conditional Random Fields (CRFs) (Haruechaiyasak and Kongyoung, 2009). *TLex* performs well when articles contain grammatical, syntactical, and stylistic standards where the writing used has a different style from that used in the micro-blog messages. Micro-blog messages like Twitter usually contain the form of a short description or keyword tags. Abbreviations are also widely used in a message. Moreover, the messages are often noisy. Therefore, *TLex* is not applicable for microblog messages due to its heavy dependence on local linguistic features.

In this stage, we aim to automatically segment micro-blog messages into words or phrases. Given an individual message m, the problem of message segmentation is to split m into k consecutive segments, m = {s_1, s_2, ..., s_k}. Each segment s_i contains one or more words. To obtain the optimal segmentation, we adopt *ThaiAnalyzer* method provided by *Apache Lucene*[8] for initial segmentation. Based on the observations, *Apache Lucene* sometime breaks a word apart incorrectly such as “โปรเจ็ค-Project” and “บางกอกแอร์เวย์-Bangkok Airway”. It splits “-Project” into two segments (i.e., “โปร-Pro”, “เจ็ค-ject”). Moreover, it sometime splits “บางกอกแอร์เวย์-Bangkok Airway” into two incorrect segments (i.e., “บาง-Bang”, “กอกแอร์เวย์-kok Airway”). In order to handle incorrect segmentation from Apache Lucene, we aggregate information in Twitter as local context and Thai Wikipedia as global context to compute the probability that a segment is a correct segment. By doing this, our approach is able to recognize new words or phrases, which may not appear in Thai dictionary.

Twitter dataset is crawled from the messages sent by users in Thailand, from the dates 17 April 2016 to 18 April 2016 with 175,294 messages. Re-tweet messages are excluded from our dataset. However, it is not necessary to consider the complete usage history of data from Twitter because of the fast information sharing on social networks. The topic may change over time. New words emerge and old ones are disappear. Two words co-occurred at time t may not appear together at time t_0 where $t_0 < t$. Therefore, previous Twitter messages (24 hours in our experiments) will be used as a local context to compute the probability that two or more words co-occur together. Thai Wikipedia data generated in February 2014 consist of 86,269 articles[9] will be used as a global context.

Before we calculate n-grams probability in next step, we simply count the number of segments generated from Apache Lucene from both Twitter and Wikipedia. If the number of a segment found in both dataset less than the threshold, the more likely the segment is an incorrect segment. The threshold is defined as the minimum number of the segment is founded in the datasets. In this paper, our experiments shown that the threshold equals 4 give the best performance.

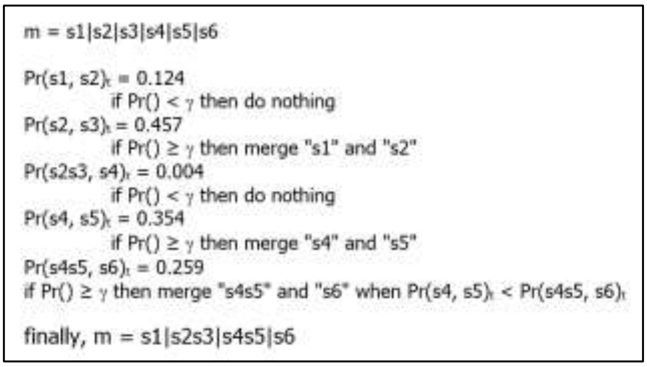
[8] https://lucene.apache.org

[9] http://www.e4thai.com/e4e/index.php?option=com content&view=article &id=1320:2014-02-15-11-11-14&catid=64&Itemid=122

The segments that have frequency less than or equals 4 will be split into two or more segments based on the positions of vowel nuclei and cluster. The algorithm is shown as Algorithm 1. A set of letters combination patterns can be seen in Thai Language article in Wikipedia[10].

C. N-Gram Combination

In this stage, all initial segments from previous step will be measured the probability of two or more segments co-occur together. The function that measures the likelihood ratio of an occurrence of segments is applied. We only compute segment occurrence up to 4-grams based on our observation from Apache Lucene results. The example of n-gram combination can be seen in Fig. 3.



```
m = s1|s2|s3|s4|s5|s6

Pr(s1, s2)t = 0.124
          if Pr() < γ then do nothing
Pr(s2, s3)t = 0.457
          if Pr() ≥ γ then merge "s1" and "s2"
Pr(s2s3, s4)t = 0.004
          if Pr() < γ then do nothing
Pr(s4, s5)t = 0.354
          if Pr() ≥ γ then merge "s4" and "s5"
Pr(s4s5, s6)t = 0.259
if Pr() ≥ γ then merge "s4s5" and "s6" when Pr(s4, s5)t < Pr(s4s5, s6)t

finally, m = s1|s2s3|s4s5|s6
```

Fig. 3. The example of n-gram combination

```
Algorithm 1 TokenSplitting(Token T)
  LC is a set of last consonant letters
  SL is a set of silence letters
  CL is a set of clusters
  TI is a set of tone indication
  RegEx is a set of letters combination patterns
  index is position that has been found using RegEX
  match_list is a list of indexes

  $RegEX$ = [Vowel][Any Thai letters][TI*][LC*][SL*] or
  [Any Thai letters][Vowel][TI*][LC*][SL*]
  for Each $RegEX$ do
    $matcher = pattern.matcher(T)$
    while $matcher$ is FOUND do
      add $index$ to $match\_list$
    end while
  end for

  $list\_of\_segments = SplitToken(T, match\_list)$
  // SplitToken is function that split token $T$ according to list
  of indexes.

  return $list\_of\_segments$
```

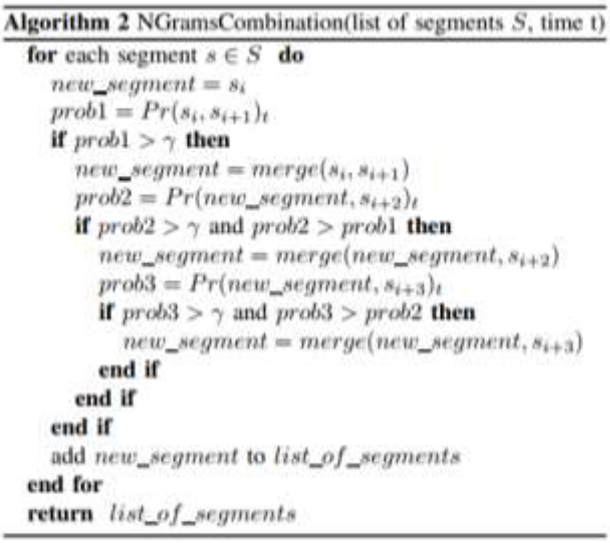
The probability of an occurrence of segments in the corpus is computed as follow:

$$P(s_1, s_2 \in D) = \frac{|s_1 s_2| \in D}{|s_1| \in D} \qquad (1)$$

where D is the corpus, $|s_1 s_2|$ is number of segment s_1 co-occur with s_2 in D, and $|s_1|$ is number of segment s_1 in D. By aggregating information in Twitter as local context and Thai Wikipedia as global context to compute the probability that a segment is a correct segment, the probability of an occurrence of segments s_1 and s_2 at time t can be computed as follow:

$$Pr(s_1, s_2)_t = \alpha P(s_1, s_2 \in TW_{t_0 - t}) + (1 - \alpha) P(s_1, s_2 \in W) \qquad (2)$$

where TW is the Twitter dataset from time t_0 to t and $t_0 < t$, W is the Thai Wikipedia dataset, α is a scaling of weights between Twitter and Wikipedia. For each message, we compute the probability of bi-grams, tri-grams and four-grams respectively. If the probability of bigram exist the merging threshold (γ), then we compute the probability of tri-grams and so on. However, if the probability of bi-grams greater than tri-grams, bi-grams will be selected as a correct segment in our approach. The algorithm of n-grams combination is shown as Algorithm 2.



```
Algorithm 2 NGramsCombination(list of segments $S$, time t)
  for each segment $s \in S$ do
    $new\_segment = s_i$
    $prob1 = Pr(s_i, s_{i+1})_t$
    if $prob1 > \gamma$ then
      $new\_segment = merge(s_i, s_{i+1})$
      $prob2 = Pr(new\_segment, s_{i+2})_t$
      if $prob2 > \gamma$ and $prob2 > prob1$ then
        $new\_segment = merge(new\_segment, s_{i+2})$
        $prob3 = Pr(new\_segment, s_{i+3})_t$
        if $prob3 > \gamma$ and $prob3 > prob2$ then
          $new\_segment = merge(new\_segment, s_{i+3})$
        end if
      end if
    end if
    add $new\_segment$ to $list\_of\_segments$
  end for
  return $list\_of\_segments$
```

IV EXPERIMENTS AND EVALUATION

In order to find the best solution of micro-blog message segmentation, we manually label 900 messages from Twitter as a test dataset. The experiments are repeated 10 times and 200 messages are random for each round. The average results of the experiments are given in Table III. We evaluate our algorithm by using Precision, Recall and F1-score.

[10] https://en.wikipedia.org/wiki/Thai language

$$Precision = \frac{|T \cap C|}{|C|} \qquad (3)$$

$$Recall = \frac{|T \cap C|}{|T|} \qquad (4)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (5)$$

where T is the true segments, C is system generated segments, |T | is number of segments in T, |C| is number of segments in C, and |T ∩ C| is number of segments that are in the same group in both T and C. We compare the performance of our approach with three baselines; ThaiAnalyzer approach by Apache Lucene, Using Thai Wikipedia as a global context approach, and Using previous Twitter messages as a local context approach.

To our best knowledge, this paper is the first to fully focus on Thai word segmentation in social networks. Three baselines are described as follows:

- Baseline 1 is an open-source library (i.e., ThaiAnalyzer) written in Java provides by Apache Lucene.
- Baseline 2 is an approach which use only Thai Wikipedia articles as a global context to compute the likelihood ratio of n-gram words. Thai Wikipedia data consist of 86,269 articles.
- Baseline 3 is an approach which use only previous one day Twitter data as a local context to calculate the likelihood ratio of n-gram words. Twitter dataset is crawled from the messages sent by users in Thailand, from the dates 17 April 2016 to 18 April 2016 with 175,294 messages. Re-tweet messages are excluded from our dataset.

According to our approach, we need to find the merging threshold (γ) and α which is a scaling of weights between Twitter and Wikipedia. In order to find the most suitable value for merging threshold (γ), we carry out segmentation on wikipedia and tweets from the dataset with different γ values. Our tests show that when γ = 0.20 it renders the best performance (as shown in Table II). Parameter α has been learned from statistics computed from the Twitter and Wikipedia datasets. Our experiments show that when α = 0.70 it renders the best performance for incorporate between Twitter and Wikipedia (as shown in Table I).

We present the results of the experiments in Table III. The baseline 2 and 3 gave the best results when merging threshold (γ) are 0.45 and 0.50 respectively. It can be seen that our approach can effectively segment Thai word with a F1-score of 64.90% which is significantly larger than the baselines. In other words, the incorporation between Twitter data and Thai Wikipedia can improve our segmentation performance.

V CONCLUSION

In this paper, an approach to automatically segment Thai words with time sensitivity over micro-blogs is developed. The goal of our approach is to effectively segment Thai word by utilizing real-time micro-blog messages and Wikipedia information. Our contributions are summarized as follows:

- An approach to segment Thai word in micro-blogs (i.e., Twitter) is presented.
- Local (i.e., Twitter) and global (i.e., Wikipedia) contexts are incorporated to improve the Thai segmentation.
- We evaluate our proposed approach with a real-world Twitter data posted by Thai-based users.

Our experiments are performed against three baseline approaches. The results show that our approach is effective in segmenting Thai words in social networks. In future work, Hybrid algorithms and Name Entity Recognition will be further studied to improve the performance of Thai word segmentation.

Table I Segmentation Results Of Our Approach With Different Weight Scale (A) And Γ = 0.20

α	Precision	Recall	F1-Score
0.5	60.00	70.01	64.62
0.6	60.83	69.51	64.88
0.7	**60.90**	**69.46**	**64.90**
0.8	60.65	69.62	64.82

Table Ii Segmentation Results Of Our Approach With Different Merging Thresholds (Γ) And A = 0.70.

γ	Precision	Recall	F1-Score
5	60.38	62.06	61.21
10	61.22	67.08	64.02
15	60.93	68.52	64.50
20	**60.90**	**69.46**	**64.90**
25	60.31	69.63	64.63
30	59.47	69.29	64.00
35	59.74	70.10	64.51
40	59.57	70.28	64.48
45	59.37	70.36	64.40
50	59.06	70.44	64.25
55	59.01	70.49	64.24
60	59.00	70.53	64.25
65	58.90	70.51	64.18
70	58.87	70.54	64.18
75	58.81	70.50	64.13
80	58.79	70.49	64.11
85	58.81	70.52	64.13
90	58.72	70.46	64.06

Table Iii Segmentation Results Of Our Approach Againt Baseline Methods.

Method	γ	α	Precision	Recall	F1-Score
Baseline 1 (Lucene)	-	-	58.63	69.30	63.52
Baseline 2 (Wiki)	0.45	-	59.07	70.92	64.46
Baseline 3 (Twitter)	0.30	-	60.52	68.68	64.34
Our Approach	**0.20**	**0.70**	**60.90**	**69.46**	**64.90**

REFERENCES

Aroonmanakun, W. (2002). Collocation and thai word segmentation. In Proceedings of the 5th SNLP & 5th Oriental COCOSDA Workshop, pages 68–75. Citeseer.

Aroonmanakun, W. et al. (2007). Thoughts on word and sentence segmentation in thai. In Proceedings of the Seventh Symposium on Natural language Processing, Pattaya, Thailand, December 13–15, pages 85–90.

Haruechaiyasak, C. and Kongyoung, S. (2009). Tlex: Thai lexeme analyser based on the conditional random fields. In Proceedings of 8th International Symposium on Natural Language Processing.

Haruechaiyasak, C., Kongyoung, S., and Dailey, M. (2008). A comparative study on thai word segmentation approaches. In Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on, volume 1, pages 125–128. IEEE.

Haruechaiyasak, C., Kongyoung, S., and Damrongrat, C. (2008b). Learnlexto: a machine-learning based word segmentation for indexing thai texts. In Proceedings of the 2nd ACM workshop on Improving non english web searching, pages 85–88. ACM.

Haruechaiyasak, C., Sangkeettrakarn, C., Palingoon, P., Kongyoung, S., and Damrongrat, C. (2006). A collaborative framework for collecting thai unknown words from the web. In Proceedings of the COLING/ACL on Main conference poster sessions, pages 345–352. Association for Computational Linguistics.

Kawtrakul, A., Thumkanon, C., Poovorawan, Y., Varasrai, P., and Suktarachan, M. (1997). Automatic thai unknown word recognition. In Proceedings of the Natural Language Processing Pacific Rim Symposium, volume 1997.

Khankasikam, K. and Muansuwan, N. (2005). Thai word segmentation a lexical semantic approach. the Proceedings of the Tenth Machine Translation Summit, pages 331–338.

Li, C., Sun, A., and Datta, A. (2012). Twevent: segment-based event detection from tweets. In 21st ACM International Conference on Information and Knowledge Management, CIKM’12, Maui, HI, USA, October 29 - November 02, 2012, pages 155–164.

Limcharoen, P., Nattee, C., and Theeramunkong, T. (2009). Thai word segmentation based-on glr parsing technique and word n-gram model. In Eighth International Symposium on Natural Lanugage Processing.

Mahatthanachai, C., Malaivongs, K., Tantranont, N., and Boonchieng, E. (2015). Development of thai word segmentation technique for solving problems with unknown words. In 2015 International Computer Science and Engineering Conference (ICSEC), pages 1–6. IEEE.

Manning, C. D. and Sch¨utze, H. (1999). Foundations of statistical natural language processing, volume 999. MIT Press.

Mathioudakis, M. and Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010, pages 1155–1158.

Meknavin, S., Charoenpornsawat, P., and Kijsirikul, B. (1997). Feature-based thai word segmentation. Proceedings of of NLPRS.

Palmer, D. D. (1997). A trainable rule-based algorithm for word segmentation. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pages 321–328. Association for Computational Linguistics.

Peng, F., Feng, F., and McCallum, A. (2004). Chinese segmentation and new word detection using conditional random fields. In Proceedings of the 20th International Conference on Computational Linguistics, COLING ’04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Phaholphinyo, S. K. K. K. S. and Kosawat, K. (2011). Thai word segmentation verification tool. In Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2011), page 16. Citeseer.

Poowarawan, Y. (1986). Dictionary-based thai syllable separation. In Proceedings of the Ninth Electronics Engineering Conference, pages 409–418.

Silva, J. F., Lopes, G. P., et al. (1999). A local maxima method and a fair dispersion normalization for extracting multiword units. In Proceedings of the 6th Meeting on the Mathematics of Language, volume 381.

Sornlertlamvanich, V. (1993). Word segmentation for thai in machine translation system. Machine Translation, National Electronics and Computer Technology Center, Bangkok, pages 50–56.

Theeramunkong, T. and Usanavasin, S. (2001). Nondictionary-based thai word segmentation using decision trees. In Proceedings of the first international conference on Human language technology research, pages 1–5. Association for Computational Linguistics.

Unankard, S., Li, X., and Sharaf, M. A. (2015). Emerging event detection in social networks with location sensitivity. World Wide Web,18(5):1393–1417.