NUEJ
Naresuan University
Engineering Journal

LCS-based Thai Trending Keyword Extraction from Online News

Kietikul Jearanaitanakij[1,*], Nattapong Kueakool[1,2], Puwadol Limwanichsin[1,2], Tiwat Kullawan[2] and Chankit Yongpiyakul[2]

[1] Department of Computer Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
[2] InfoQuest Limited, Bangkok, Thailand

* Corresponding author e-mail: kietikul.je@kmitl.ac.th

(Received: 23 August 2022, Revised: 4 November 2022, Accepted: 15 November 2022)

Abstract

A trending keyword is a common word or a phrase that is most frequently mentioned in the current period. Extracting trending keywords from Thai online news is not trivial. A too-short keyword may not have a specific meaning because it may be just a common word that does not have any significance to the interpretation. On the other hand, a long common keyword conveys a better meaning. However, the running time to extract the long keyword from a collection of documents may not be bounded within a reasonable time. A problem statement of this research is to find a varying-length trending keyword from Thai online news within a reasonable running time. We propose a novel method to extract trending keywords by applying the longest common substring (LCS) algorithm. The common keywords having high occurrence frequency are selected as the trending keywords. The proposed method inherits the advantage of the reasonable running time from the dynamic programming technique of the LCS algorithm. The experimental results on various sources of Thai online news agencies indicate a superior precision of the proposed method over char-N-gram and word-N-gram strategies.

Keywords: Longest common substring, Natural language processing, Online news, Thai trending keyword, Varying-length keyword.

1. INTRODUCTION

Among all sources of news in Thailand, it is undeniable that online news is the most popular platform. There is a lot of online news issued per day making the current trend extraction difficult. To find the trending keyword, two mechanisms are needed to implement. The first is the process of finding the common keyword from a collection of documents. The second is the ability to determine that the selected common keyword is the current news trend. There are many methods for extracting a common keyword from documents. Lee and Kim (2008) apply TF-IDF (Robertson, 2004) variants and filter keywords across domains to extract keywords from a huge pile of documents. The experimental results show that their approach can efficiently remove meaningless Korean words. Ma et al. (2008) identify the query and the topic-related features for each word that cooccurs in the same window length and then calculate the importance of the word from the combined features. Their experimental results are competitive with other candidates. Zhang et al. (2008) introduce the keyword extraction based on Conditional Random Fields (CRF) that use most of the features in the document. The CRF model outperforms other off-the-shelf machine learning algorithms such as SVM and multiple linear regression models in terms of F1 score. However, a computational cost of training CRF model is expensive.

Once the common keyword is detected, the next step is to confirm that the keyword represents the actual news trend. Various approaches are proposed to identify the trending keyword. Shimizu et al. (2005) extract the trending keyword from the press releases to analyze the competitors' strategies. Words that are matched with a description pattern, i.e., a product common phrase observed from press releases, are considered to be a product trending keywords. However, their method requires a user to specify a description pattern which may not be appropriate for an automatic system. Sutheebanjard and Premchaiswadi (2010) applied four string matching techniques, e.g., Guth, Levenshtein, Damerau-Levenshtein, longest common substring and longest common subsequence to disambiguate Thai personal names from a collection of articles. They found that the longest common subsequence was the best Thai personal name matching with 94.43% of F-Score. Aiello et al. (2013) identify the trending topics on Twitter by combing N-grams occurrence with a topic-ranking score representing the rapidly emerging topic. Their method outperforms other techniques on Twitter data. Madani et al. (2015) discover trending topics from Twitter in real-time by using topic clustering. Tweets thesaurus created in the form of a bag of words can semantically represent different terms corresponding to the same trending topic. Ousirimaneechai and Sinthupinyo (2018) detect the trending keywords using the simple character N-grams to tokenize document and find keywords. They identify the trending keyword by filtering a set of stop words generated by their algorithm. Their method requires neither word tokenization tools nor external stop words list. According to the experimental results, their method frequently takes too-short words as the keywords and depends mainly on the generated stop words. Indra et al. (2019) compare two methods, namely document pivot, and BN-grams, for detecting trending topics on Indonesian tweets. Their experimental results on ten topics indicate that BN-grams have better topic recall than the document pivot. Alzubi et al. (2020) extract trending scientific topics by scoring topics according to two factors; the number of citations and the number of accepted papers on that topic. In addition, their method can recommend topics customized to each user profile. However, they do not experiment on the real dataset but on the synthesized 3000 papers instead. Tanantong et al. (2020) extract trending keywords from Thai Twitter by using the N-gram-based word-combination technique. Words that are located in an adjacent position are considered to combine into one candidate keyword. The candidate keyword which has the highest rate of appearances within three consecutive days on Twitter is extracted as the trending keyword.

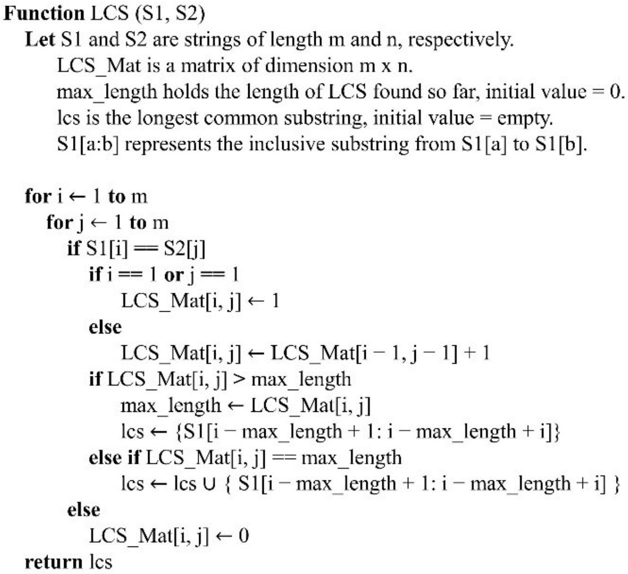
According to the above works, none of them extracts long and varying-length common keywords which tend to convey better meaning than short keywords. In this research, we propose Thai trending keyword extraction from online news by using the longest common substring (LCS) algorithm and the rescaled keyword frequency. LCS is a powerful tool that can find the longest common keyword from the collection of documents. It can be customized to search for varying-length keywords. While LCS helps us find the longest common keywords, the rescaled frequency of keyword occurrences identifies whether the common keyword is the current trend. The experimental results on the online Thai news collected from different sources show a superior performance of the proposed method compared to other algorithms.

2. RELATED WORKS

Two fundamental concepts, e.g., the longest common substring, Thai word segmentation, related to this research are described in the first two subsections. In addition, two previous works are also briefly explained since we compare their results with the proposed method. Let us abbreviate techniques in (Ousirimaneechai and Sinthupinyo, 2018) and (Tanantong et al., 2020) as Char-N-Gram and Word-N-Gram, respectively.

2.1 The Longest Common Substring Problem

The longest common substring (LCS) is a classical problem in computer science (Gusfield, 1997). The problem is about finding the longest substring of two or more strings. For instance, the LCS of “AABCDEFG” and “ABABCDEGH” is “ABCDE”. By using a dynamic programming technique, we can locate LCS between two strings of lengths m and n within O(mn) time complexity. Fig. 1 illustrates pseudocode for finding LCS between two strings (S1 and S2) by a dynamic programming technique. The algorithm processes the LCS matrix from left to right and top to bottom. For each element indexed by row i and column j, if S1[i] is equal to S2[j] then the value of LCS_Mat[i, j] equals the value from the previous upper left element plus one. Otherwise LCS_Mat[i, j] equals to zero. After all elements in the LCS matrix have been filled, the common substring which has the longest length will be chosen as the solution. The example for extracting the longest substring from the above two strings can be demonstrated in Fig. 2. Denote that the green diagonal elements represent the longest common substring which has the length of five alphabets, i.e., ABCDE. More details of the LCS algorithm including its variations (Mousavi et al., 2012; Beal et al., 2016; Charalampopoulos et al., 2021; Akmal et al., 2021) can be found in the reference section.



```
Function LCS (S1, S2)
  Let S1 and S2 are strings of length m and n, respectively.
      LCS_Mat is a matrix of dimension m x n.
      max_length holds the length of LCS found so far, initial value = 0.
      lcs is the longest common substring, initial value = empty.
      S1[a:b] represents the inclusive substring from S1[a] to S1[b].

  for i ← 1 to m
    for j ← 1 to m
      if S1[i] == S2[j]
        if i == 1 or j == 1
          LCS_Mat[i, j] ← 1
        else
          LCS_Mat[i, j] ← LCS_Mat[i − 1, j − 1] + 1
        if LCS_Mat[i, j] > max_length
          max_length ← LCS_Mat[i, j]
          lcs ← {S1[i − max_length + 1: i − max_length + i]}
        else if LCS_Mat[i, j] == max_length
          lcs ← lcs ∪ { S1[i − max_length + 1: i − max_length + i] }
      else
        LCS_Mat[i, j] ← 0
  return lcs
```

Figure 1 Pseudocode of LCS algorithm

	A	B	A	B	C	D	E	G	H
A	1	0	1	0	0	0	0	0	0
A	1	0	1	0	0	0	0	0	0
B	0	2	0	2	0	0	0	0	0
C	0	1	0	0	3	0	0	0	0
D	0	0	0	0	0	4	0	0	0
E	0	0	0	0	0	0	5	0	0
F	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	1	0

Figure 2 LCS example for strings AABCDEFG and ABABCDEGH

2.2 Thai Word Segmentation

One of the hardest parts of Thai natural language processing is word segmentation. Due to the lack of space between Thai words, we need to apply subtle techniques to separate them into individual words. There are several well-known algorithms for Thai word segmentation, e.g., Newmm, Deepcut, and AttaCut. Newmm, implemented in PyThaiNLP (Phatthiyaphaibun et al., 2016), uses a maximal matching algorithm, based on words in a dictionary, and the Thai character cluster. It always breaks the same longest word no matter the position of the word in the context. For example, a string “ตากลม” is always segmented into “ตาก | ลม” (to gain air). On the other hand, DeepCut (Kittinaradorn et al., 2019) and its variance AttaCut (Chormai et al., 2019) apply a more advanced technique, i.e., a convolutional neural network, to segment a sentence into a list of words depending on the word surrounding context. From the previous example, the string “ตากลม” can be either segmented into “ตาก | ลม” (to gain air) or “ตา | กลม” (round eye) depending on its surrounding words. To conform with the LCS algorithm, we employ Newmm algorithm to consistently tokenize sentences into a list of the longest words.

2.3 Char-N-Gram

Char-N-Gram (Ousirimaneechai and Sinthupinyo, 2018) begins the process by tokenizing text into five-character grams and counting the number of occurrences of them. Afterward, the K-Means clustering algorithm is applied to the counted grams to group those grams into K clusters by using the elbow method. The lowest rank cluster is discarded since it has a low chance of containing keywords. The keyword clusters from other days are combined and clustered with the K-Means algorithm again. The cluster with the highest mean of occurrences will be assigned as the global keyword grams. For each news post, replace characters that are not in any part of elements in the global keyword grams with a blank symbol and split the resulting text by using a blank symbol. The global keywords are words that remained in the news post after removing stop words. Finally, to find the trending keywords, pick the global keywords which have a high frequency of occurrences within a day.

2.4 Word-N-Gram

Word-N-Gram (Tanantong et al., 2020) applies a simpler idea. It replaces stop words from the original post with a blank symbol and combines the adjacent non-blank words. These adjacent words are candidates for the trending keywords. The candidate which has the highest frequency of occurrences in all posts within a period is counted as a trending keyword.

3. Proposed Method

The following steps describe the proposed method along with the diagram in Fig. 3.

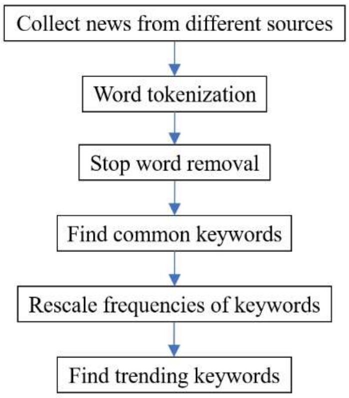
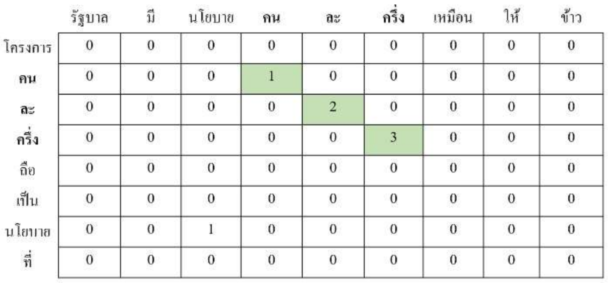


Figure 3 Proposed method

Step 1: Collect the current news posts from different sources and sort them by a chronological order.
Step 2: Tokenize words in news posts by using Newmm algorithm.
Step 3: Remove stop words from the tokenized news.
Step 4: Find the common keywords of all news posts by using the LCS algorithm.
Step 5: Rescale occurrence frequencies of all common keywords.
Step 6: Pick the rescaled LCS keywords which have high occurrence frequency as the current trends.

After breaking news posts into a list of words by Newmm technique, we follow the process in (Tanantong et al., 2020) to remove stop words in step 3. Stop words tend to be insignificant to the main idea and frequently appear in news. To avoid selecting common keywords which are not necessary to be trending keywords, we monitor words that occur every day for one month and remove them from a list of tokenized words. It is worth noting that the row and column elements of the LCS matrix in step 4 are represented as the tokenized words (not alphabets). To avoid too-short and too-long keywords, we select only common phrases whose lengths are within three to eight words. As a result, the proposed method can produce the varying-length trending keywords. The example of LCS matrix calculation in step 4 can be illustrated in Fig. 4. The purpose of the rescaling process in step 5 is to ensure that all common keywords returned from LCS every day are on the same scale. The rescaled frequency of each LCS keyword is the ratio between the number of its occurrences and the total number of occurrences of all LCS common keywords within a certain period.



	รัฐบาล	มี	นโยบาย	คน	ละ	ครึ่ง	เหมือน	ให้	ข้าว
โครงการ	0	0	0	0	0	0	0	0	0
คน	0	0	0	1	0	0	0	0	0
ละ	0	0	0	0	2	0	0	0	0
ครึ่ง	0	0	0	0	0	3	0	0	0
ถือ	0	0	0	0	0	0	0	0	0
เป็น	0	0	0	0	0	0	0	0	0
นโยบาย	0	0	1	0	0	0	0	0	0
ที่	0	0	0	0	0	0	0	0	0

Figure 4 Example of the LCS matrix for a common keyword "คนละครึ่ง"

4. Experimental Results

We collect 155,661 online posts from eight different Thai news agencies; Thairath, Thai PBS, Spring News, Siamrath, Sanook, Post Today, Komchadluek, and Amarin, between 1st September 2021 and 31st December 2021. Table 1 shows the number of news instances from each agency. The ground truth of each day, i.e., the actual trending keyword, is manually assigned as a set of 1 – 3 keywords by our news experts. Each instance of news in the dataset composes of date, time, and news content.

To have a fair comparison, we conduct the experiments of all methods on the standard Google Colab (cpu: 2.2 GHz/single core/cache: 56320 KB, ram: 13.3 GB, disk: 108 GB). We begin the experiment with the quest for a suitable Thai word tokenization. Later in the section, the experimental results for top–10 precision, running time, and word cloud are provided.

Table 1 The number of posts collected from news agencies

News Agency	Number of news posts
Thairath	27544
Post Today	16988
Sanook	10743
Thai PBS	5895
Komchadluek	25450
Amarin	10999
Siamrath	51211
Spring	6831

The objective of the first experiment is to confirm that a suitable word tokenization algorithm for the proposed method is Newmm. We randomly pick 10000 news instances from the dataset and apply both Newmm and AttaCut to break news into words. The number of segmented words that appear in the LEXiTRON dictionary (Trakultaweekoon et al., 2007) and the running time of each algorithm are reported in Table 2. Newmm has better performance than AttaCut in terms of both percentages of words that appear in the LEXiTRON dictionary and running time. Moreover, words segmented by Newmm are consistent and do not depend on the surrounding context which is suitable for finding the trending keywords. As a result, we employ Newmm as the Thai word segmentation. To break English words which may embed in Thai news, we simply split text by using a blank character as a delimiter. As we will see in the experimental section, the Thai trending keyword tends to be a phrase that does not depend on the surrounding context. Therefore, the percentage of words that appear in the LEXiTRON dictionary is a suitable indicator to measure the consistency of word tokenization of the two algorithms.

Table 2 Word tokenization comparison between Newmm and AttaCut.

Measure	Newmm	AttaCut
Percentage of words appear in LEXiTRON dictionary (%)	82.60	44.47
Running time (second)	9.18	88.27

The experiments of different methods are conducted on the dataset described in the previous section. The ground truth for each day is manually assigned as a set of 1 – 3 trending keywords by our volunteers. News instances are fed to the algorithms in chronological order starting from the oldest to the latest. Each algorithm predicts ten trending keywords for each day. If the volunteers found that at least one of ten predictions semantically matches the ground truth of that day, the prediction is counted as a hit, i.e., top-10 precision. The task of precision judging is not trivial to execute without human intervention since the predicted trending keywords may not exactly match the ground truth keywords. In other words, they implicitly match the meaning of the ground truths. For example, the keyword "แอนชิลี สก็อต-เคมมิส" (Anchilee Scott-Kemmis) implicitly matches the ground truth "มิสยูนิเวิร์สไทยแลนด์" (Miss Universe Thailand), "แอพเป๋าตัง" (Paotang application) implicitly matches the ground truth "โครงการคนละครึ่ง" (The half-half co-payment project). Therefore, we need human to judge the top-10 precision. The top-10 precision percentages of the proposed method (LCS-based) and the other two techniques; Char-N-Gram (Ousirimaneechai and Sinthupinyo, 2018) and Word-N-Gram (Tanantong et al., 2020), are shown in Table 3.

LCS-based method gains the best top-10 precision of all three months. Since the number of words in a common keyword can be varied between three and eight, the LCS-based method can produce meaningful keywords. In contrast, Char-N-Gram limits the length of the keyword to a gram of five characters and keeps only the cluster that has the highest mean of occurrences. Although this cluster usually contains keywords that have the highest mean of occurrences, those keywords tend to be short and less meaning, e.g., "การ" (task), "เพราะ" (because), "จำนวน"

(amount). Therefore, it frequently misses long keywords like “น้อม รำลึก ถึง พระ มหา กรุณา ธิ คุณ” (bow in remembrance of His Majesty the King), “เจ้า พนักงาน ละเว้น โดย มิ ชอบ” (staff wrongful refrain), “ฮู แต ยือ ลอ เระ บา เจาะ” (Hutaeyelor village, Bacho district), etc. On the other hand, Word-N-Gram does not suffer from short keywords. After stop word removals, it concatenates the remaining adjacent words resulting in keywords with different lengths. However, Word-N-Gram does not try to produce the longest keyword like the LCS-based method. The keywords from Word-N-Gram which have the highest rate of occurrences may be long, e.g., “มหาวิทยาลัย” (university), “อาจารย์เจ้าหน้าที่” (staff and faculty), but do not have specific meanings. As a result, more than half of trending keywords from Word-N-Gram do not semantically match any of the three ground-truth keywords.

The next experiment is to measure the running time among three methods. Table 4 compares the running time (in seconds) of trending keyword extraction for all news instances within three months. Word-N-Gram spends the least running time because of its simple procedure. In contrast, Char-N-Gram takes so much time in finding all possible 3-gram, 5-gram, and 8-gram words and running two rounds of K-Means clustering. The LCS-based method inherits a moderate running time of O(mn) from the dynamic programming technique of the LCS algorithm (Gusfield, 1997); where m and n are news lengths. Although we can further optimize the running time of LCS by creating a generalized suffix tree for strings, there is a tradeoff in memory space for storing a large tree.

Table 3 Top-10 Precisions of three methods

Measure	Char-N-Gram (N = 3 / 5 / 8)	Word-N-Gram	LCS-based
October	11.45/13.04/12.94	26.09	63.04
November	11.98/14.13/14.03	39.13	57.60
December	16.54/19.56/18.99	45.65	60.86
Average	13.32/15.58/15.32	36.96	60.50

Table 4 Running time in second of three methods

Measure	Char-N-Gram (N = 3 / 5 / 8)	Word-N-Gram	LCS-based
October	5211/6318/7410	320	955
November	5345/6655/7833	322	885
December	5572/6845/7942	330	856
Average	5376/6606/7728	324	898.67

To have a quick and visual insight of the trending keywords from different methods, we generate word clouds to highlight popular keywords based on their frequencies of occurrences. The larger font of the word, the more frequently it occurs in news posts. Fig. 5 illustrates word clouds on 25th October 2021 which has two ground truths; “พิธีอัญเชิญพระเกี้ยว” (Phra Kiew coronet parade) and “มิสยูนิเวิร์สไทยแลนด์” (Miss Universe Thailand). Keywords in red ellipses are trends that semantically match the trending ground truths. Trending keywords from Char-N-Gram and Word-N-Gram do not partially match any keywords in the ground truths. Keywords from Char-N-Gram are too short while those from Word-N-Gram are not related to the ground truths. In contrast, two top trending keywords, i.e., “นิสิตอัญเชิญเกี้ยว” (students who participate in Phra Kiew coronet parade), and “ชิลีสก๊อตมิส” (Anchilee Scott), from LCS-based method semantically match the target trending keywords. Notice that many long keywords from Word-N-Gram have a small frequency of occurrences, e.g., “มันสำปะหลังข้าวโพดเลี้ยงสัตว์” (cassava, corn, animal husbandry), “หลอมรวมความสมัครสมาน” (gather unity), “อนึ่งหย่อมความกดอากาศ” (by the way, the air pressure patch) In addition, the air pressure patch. After combining words leftover from the stop word removals, those long keywords of Word-N-Gram do not reflect any trend in the current news posts. On the other hand, the LCS-based method intends to find the longest common keywords to help promote the collocation degree of keywords; making them cover more meanings in the ground truths.

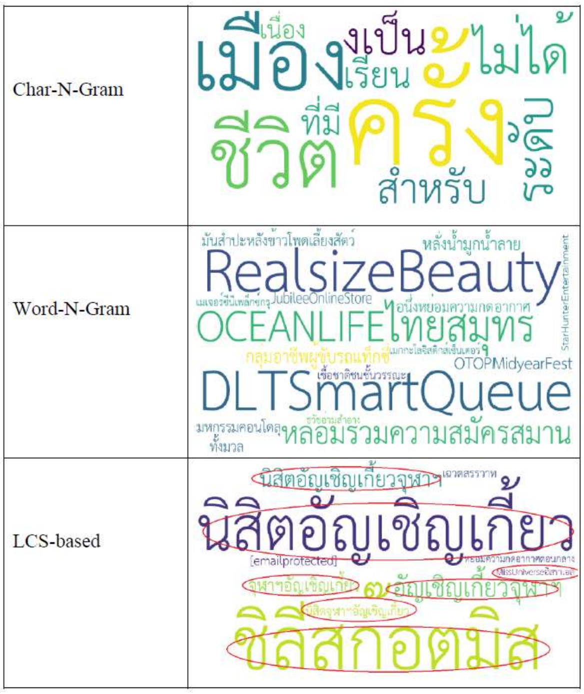


Figure 5 Word clouds for trending keywords on 25 October 2021

Fig. 6 expresses a similar phenomenon as Fig. 5. It demonstrates the word cloud on 23rd November 2021 which has the ground truths “พัคชินฮเย ชเวแทจุน” (Park Shin-hye & Choi Tae-joon), “ดราม่า แอนชิลี สก็อต” (Anchilee Scott accused of Thailand flag abuse). Neither Char-N-Gram nor Word-N-Gram captures any trend in the ground truths whereas the LCS-based method covers all current trends. Fig.7 shows another word cloud on 24th December 2021 which has the ground truths “เลือกตั้งซ่อม กทม.” (Bangkok election), “ฟุตบอลไทย-เวียดนาม AFF” (Thailand-Vietnam AFF football), “โอมิครอน” (Omicron). Both Char-N-Gram and Word-N-Gram implicitly capture only one keyword “โอมิครอน” (Omicron) while the LCS-based method semantically hits all trending ground truths.

Interestingly, Word-N-Gram contains more English words than other methods in Fig. 5-7. These are situations where Word-N-Gram lost important Thai stop words while replacing them with blank symbols. After combining non-blank words, its keywords may look strange and have a low frequency of occurrences. As a result, it misses the ground truth trending keywords and continues extracting lower-frequency keywords which may include English keywords.

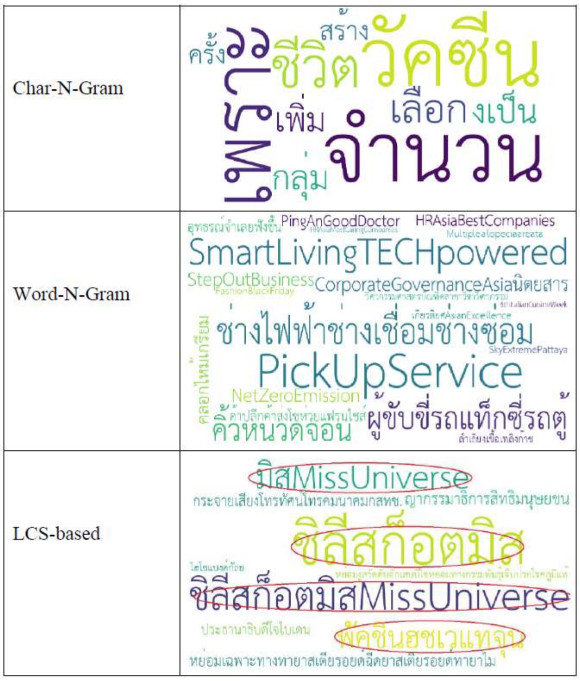


Figure 6 Word clouds for trending keywords on 23 November 2021

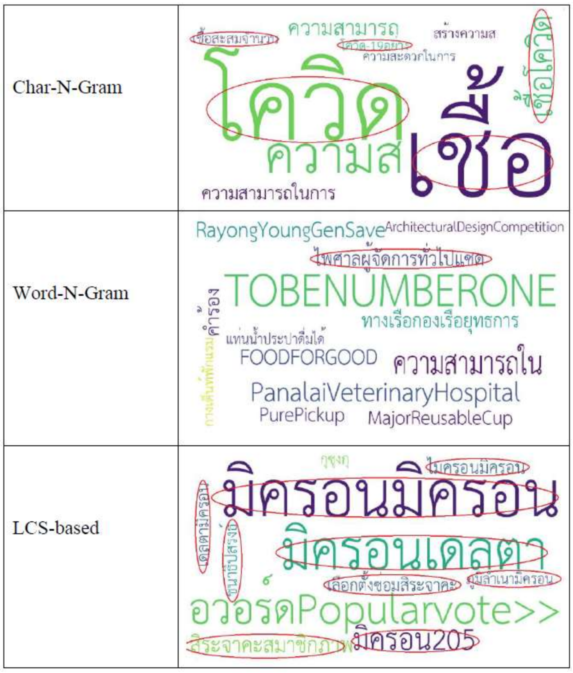


Figure 7 Word clouds for trending keywords on 24 December 2021

5. Conclusion

We propose a novel method to extract Thai trending keywords from online news by using the longest common substring (LCS) algorithm to find a set of the longest common keywords. The longest common keywords that have a high frequency of occurrences will be selected as the trending keywords. The experimental results on 155,661 online news posts collected from eight Thai news agencies between September 2021 and December 2021 indicate the superior performance of the LCS-based method over other techniques. The contribution of the proposed method is the effort to find the longest trending keyword from the current collection of news. The LCS-based method can find the keywords with varying ranges, e.g., 3 to 8 words, depending on the context of the news. In practice, we recommend a user to find a suitable range from the sample news posts. We use the range between 3 and 8 words because of two factors. First, a glance from our dataset does not show a keyword with more than eight words. Second, we need to control the running time of the LCS algorithm within a reasonable bound. The advantage of long keywords is the ability to express trends more meaningfully. The LCS-based method possesses the highest top-10 precision, on average, 60.50% and the moderate running time at 898.67 seconds compared to other methods. In addition, it is possible to apply the LCS-based method to extract trending keywords from other languages by simply changing a word tokenization module. One of the possible future works of the proposed method is to apply the cosine similarity to group similar trending keywords together and express them as one keyword. This improvement can help limit the trending keywords to a smaller set of keywords and make it easy to interpret news insight. Despite its moderate running time, the LCS-based method may suffer from the O(NxN) operation for coupling a pair of news to find LCS. It is recommended to preliminary run the LCS-based method to find a suitable value of N that produces an acceptable total running time or apply multiple threads to gain the concurrency of LCS operations.

6. Acknowledgment

We gratefully acknowledge support from eight Thai news agencies, including Thairath, Post Today, Sanook, Thai PBS, Komchadluek, Amarin, Siamrath, and Spring, for their valuable online news posts.

7. References

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., G ̈oker, A., Kompatsiaris, Y., & Jaimes, A. (2013). Sensing trending topics in Twitter. *IEEE Transaction on Multimedia*, *15(6)*, 1268–1282. https://doi.org/10.1109/TMM.2013.2265080

Akmal, S., & Williams, V. V. (2021). Improved approximation for longest common subsequence over small alphabets. *48th International Colloquium on Automata, Languages, and Programming* (pp. 1-19). ArXiv. https://doi.org/10.48550/arXiv.2105.03028

Alzubi, S., Hawashin, B., Mughaid, A., & Jararweh, Y. (2020). Whats trending? an efficient trending research topics extractor and recommender. *11th International Conference on Information and Communication Systems* (pp. 191-196). IEEE. https://doi.org/10.1109/ICICS49469.2020.239519

Beal, R., Afrin, T., Farheen, A., & Adjeroh, R. (2016). A new algorithm for "the LCS problem" with application in compressing genome resequencing data. *BMC Genomics*, *17(4)*, 369–381. https://doi.org/10.1186/s12864-016-2793-0

Charalampopoulos, P., Kociumaka, T., Pissis, S. P., & Radoszewski, J. (2021). Faster algorithms for longest common substring. *29th Annual European Symposium on Algorithms* (pp. 1-30). ArXiv. https://doi.org/10.48550/arXiv.2105.03106

Chormai, P., Prasertsom, P., & Rutherford, A. (2019). AttaCut: a fast and accurate neural Thai word segmenter. *ArXiv*, *1911(7056)*, 1–13. https://doi.org/10.48550/arXiv.1911.07056

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511574931

Indra, S. K., Winarko, E., & Pulungan, R. (2019). Trending topics detection of Indonesian tweets using BN-grams and Doc-p. *Journal of King Saud University - Computer and Information Sciences*, *31(2)*, 266–274. https://doi.org/10.1016/j.jksuci.2018.01.005

Kittinaradorn, R., Chaovavanich, K., Achakulvisut, T., Srithaworn, K., Chormai, P., Kaewkasi, C., Ruangrong, T., & Oparad, K. (2019, September 23). DeepCut: A Thai word tokenization library using Deep Neural Network. Retrieved from https://doi.org/10.5281/zenodo.3457707

Lee, S., & Kim, H. (2008). News keyword extraction for topic tracking. *Fourth International Conference on Networked Computing and Advanced Information Management* (pp. 554-559). IEEE. https://doi.org/10.1109/NCM.2008.199

Ma, L., He, T., Li, F., Guil, Z., & Chen, J. (2008). Query-focused multi-document summarization using keyword extraction. *International Conference on Computer Science and Software Engineering* (pp. 20-23). IEEE. https://doi.org/10.1109/CSSE.2008.1323

Madani, A., Boussaid, O., & Zegour, D. E. (2015). Real-time trending topics detection and description from Twitter content. *Social Network Analysis and Mining*, *5(59)*, 1–13. https://doi.org/10.1007/s13278-015-0298-5

Mousavi, S. R., & Tabataba, F. (2012). An improved algorithm for the longest common subsequence problem. *Computers & Operations Research*, *39(3)*, 512–520. https://doi.org/10.1016/j.cor.2011.02.026

Ousirimaneechai, N., & Sinthupinyo, S. (2018). Extraction of trend keywords and stop words from Thai Facebook pages using character n-grams. *International Journal of Machine Learning and Computing*, *8(6)*, 589–594. http://www.ijmlc.org/vol8/750-ML0015.pdf

Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., & Chormai, P. (2016, June 27). PyThaiNLP: Thai Natural Language Processing in Python. Retrieved from http://doi.org/10.5281/zenodo.3519354

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520. https://doi.org/10.1108/00220410410560582

Shimizu, Y., Akiyoshi, M., & Komoda, N. (2005). A method of extracting product trend keywords from press releases to analyze product strategy of competitors. *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce* (pp. 631-635). IEEE. https://doi.org/10.1109/CIMCA.2005.1631539

Sutheebanjard, P., & Premchaiswadi, W. (2010). Disambiguation of Thai personal name from online news articles. *International Conference on Computer Engineering and Technology* (pp. V3-302-V3-306). IEEE. https://doi.org/10.1109/ICCET.2010.5485879

Tanantong, T., Kreangkriwanich, S., & Laosen, N. (2020). Extraction of trend keywords from Thai Twitters using n-gram word combination. *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (pp. 320-323). IEEE. https://doi.org/10.1109/ECTI-CON49241.2020.9158061

Trakultaweekoon, K., Porkaew, P., & Supnithi, T. (2007). LEXiTRON vocabulary suggestion system with recommendation and vote mechanism. *Proceedings of Symposium of Natural Language Processing* (pp. 43-48). National Electronics and Computer Technology Center. lexitron.nectec.or.th/2009_1/paper/paper_3.pdf

Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y., & Wang, B. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computer Information Systems*, *4(3)*, 1169–1180. https://core.ac.uk/download/pdf/11884499.pdf