WALAILAK JOURNAL
http://wjst.wu.ac.th Information Technology

Pattern-Sensitive Loanword Estimation for Thai Text Clustering*

Burhan WANGLEM* and Nattapong TONGTEP

Faculty of Technology and Environment, Prince of Songkla University, Phuket Campus,
Phuket 83120, Thailand

(*Corresponding author's e-mail: burhan.w@phuket.psu.ac.th, nattapong.t@phuket.psu.ac.th)

Received: 11 September 2016, Revised: 1 April 2017, Accepted: 1 May 2017

Abstract

Writing style and language usage vary depending on the purpose of the writers and change the readability. A good assessment of text readability helps readers find suitable texts with less effort. In the Thai language, text readability assessment is one of the challenging tasks in natural language processing, because the Thai texts are not segmented by words and have only ambiguous boundary markers for word and sentence segmentation. Furthermore, loanwords, words borrowed from other languages such as Pali and Sanskrit, play important roles in text readability. In this paper, we propose a method to cluster Thai texts according to their readability by detecting loanwords that can be used as features. First, loanwords in Thai are categorized into 7 types as different patterns. Then the set of loanword patterns is employed to detect those patterns in the set of documents retrieved from the search engine. The experimental result shows that the detection of the Thai words that are loaned not only from Pali but also from Sanskrit achieved the highest F-measure up to 100 and 98.29 % accuracy.

Keywords: Loanword detection, Pali word, Sanskrit word, Thai language, text clustering

Introduction

There are several documents in paper and digital formats at present [1]. The writing style and language are different depending on the purpose of the writers. For example, 2 authors want to write books about animals. The first author intends to write the book for children while another author wants to write the book for high school students. The book for children is easy to understand with fewer texts, and more pictures while the book for students has more technical terms, longer sentences, less pictures, and more examples. If readers would like to read a book that suits their favor and style, readers will have to read several books in order to find the proper one. If there are many books, readers will have to take more time to check. The level of readability based on the document content is proposed in order to help readers spend less time to find the books which suit their needs. To assess the difficulty level in English texts, there are 2 widely used formulas: Flesh and Flesh-Kincaid [2]. However, these 2 formulas are inapplicable for languages which their writing system is syllabic alphabet.

Thai, one of syllabic alphabetic languages, consists of consonants, vowels, and indistinctly boundary markers of word and sentences. There are few research works on assessing text readability in syllabic alphabetic languages, especially in the Thai language. Daowadung and Chen [3] proposed a text readability method for primary school student. Tongtep *et al*. [4] exploited text readability assessments on Thai texts with a small set of documents.

*Presented at 1st International Conference on Information Technology: October 27th - 28th, 2016

Loanwords, which were borrowed from other languages, make languages develop and create more vocabularies. Loanwords in the Thai language borrowed from several languages such as Pali, Sanskrit which can be used as religious vocabularies, royal words, or words in the literature [5]. Characteristics of loanwords are different from the original Thai words. Thai words are mostly single syllable with clear meaning such as แม่ (Mother) กิน (eat) ข้าว (rice). Spelling and writing a sentence using the original Thai words are not complicated in contrast to loanwords. The characteristics of loanwords can be used to assess the readability of texts.

In this work, extracting loanwords from the digital Thai texts are proposed. Patterns of Pali and Sanskrit loanwords are studied, constructed, and extracted based on linguistic knowledge. These patterns are applied to the Thai texts acquired from the digital documents. The extracted loanwords will be applied to evaluate the readability of the Thai texts.

The rest of the papers are related work, our proposed framework, followed by experimental setting sections. Results are discussed the next section while conclusion and future work are summarized at the ending section.

Related work

Loanword

A loanword is a word borrowed from other languages which can be found in all languages [6]. In English, loanwords borrowed from French, Greek, Spanish, and Italian. In Thai, loanwords mainly borrowed from Pali, Sanskrit, Khmer, Chinese, and English. Loanwords are different from original Thai words. Generally, original Thai words are one syllable and complete meaning. We can use loanword as features for ranking the readability of texts by studying loanword patterns in the focused language.

Pali

Words borrowed from Pali have final consonant or final pronounced letter of a word and following consonant. The following consonant is adjacent to the final consonant based on principles [7] as shown in **Table 1**.

Table 1 Relationship between final consonants and following consonants in Pali.

Group (วรรค)	Row 1	Row 2	Row 3	Row 4	Row 5	Row 6 or other group (เศษวรรค)
วรรค กะ	ก	ข	ค	ฆ	ง	ย ร ล ว ส ห ฬ
วรรค จะ	จ	ฉ	ช	ฌ	ญ	
วรรค ฎะ	ฎ	ฐ	ฑ	ฒ	ณ	
วรรค ตะ	ต	ถ	ท	ธ	น	
วรรค ปะ	ป	ผ	พ	ภ	ม	

From the **Table 1**, characteristics of a Thai word borrowed from Pali can be described as follows.

1. Each group, a consonant in Row 1 is followed by a consonant in Row 1 or Row 2. For example, a consonant ก in Row 1 is followed by a consonant ข in Row 2 such as ทุกข์.

2. Each group, a consonant in Row 3 is followed by a consonant in Row 3 or Row 4. For example, a

consonant ค in Row 3 is followed by a consonant ค in Row 3 such as อัคคี.

3. Each group, a consonant in Row 5 is followed by a consonant in Rows 1 - 4. For example คงคา.

4. A consonant in Row 6 is followed by a consonant in Rows 1 - 5, then a consonant in Row 6 is diphthong as follows;

- A consonant ย is diphthong if ย is followed by a consonant in Rows 1 - 5.
- A consonant ร is diphthong if ร is followed by a consonant in Row 1 or Row 3.
- A consonant ล is diphthong if ล is followed by a consonant in Row 1.
- A consonant ว is diphthong if ว is followed by a consonant in Rows 1 - 5.
- A consonant ห is diphthong if ห is followed by a consonant in Row 5.

Sanskrit

Thai words borrowed from Sanskrit have final consonant or final pronounced letter of a word and following consonant like Pali with additional consonants and vowels i.e. ศ ษ ฤ ฤๅ ฦ ฦๅ ไอ เอา [8]. Thai words borrowed from Sanskrit can be observed from these additional consonants and vowels.

Loanwords and their orthography

Original Thai words do not have orthography because Thai is isolating language. Thai words borrowed from other languages are likely to have Karan, a consonant with a muter such as ร์ in คอนเสิร์ต (Concert) [7]. Karan can be used as a clue to detect loanwords.

Framework

Our framework has mainly 6 processes i.e. data collection, data preprocessing, model construction, loanword extraction by expert, loanword extraction by model, and model evaluation as shown in **Figure 1**.

Data collection

This process is required for collecting the relevant documents based on an interesting topic. First, keywords are determined and then applied for searching related documents (Document Search). A set of related documents and their links can be retrieved using search engine algorithm with specified keywords.

Data preprocessing

This process prepares a set of related texts used in the loanword extraction process by model and experts. Texts from the related documents collected during the data collection process are retrieved (Text Retrieval). Images, videos, and other medias including programming language scripts are discarded. The number of characters in each set of related documents can cause inappropriate results related to document processing or ranking techniques, then limiting the number of characters in a document is optionally adapted (Character Limitation). The final step in the data preprocessing is segmenting characters into words (Word Segmentation) using word segmentation techniques such as longest word matching [9], or character cluster-based word segmentation [10].

Model construction

This process constructs the model for detecting loanword positions. Loanword writing style in the focused language is observed and analysed (Loanword Study). In this work, the theories of loanword related to the Thai language are studied. Loanwords mostly found in the Thai language are Pali and Sanskrit words. Loanword patterns are then computationally constructed as the detection model (Loanword Pattern Construction). The position of each loanword is observed and constructed as rules or patterns as shown in **Figure 2**. For example, the adjacent consonants of ก and ข are part of loanword pattern observed in Pali, so กข is constructed as a rule for detecting Pali. In this paper, 47 patterns of Pali loanwords, 25 patterns of Sanskrit loanwords, and one pattern of orthography loanword are constructed as shown in **Table 2**.

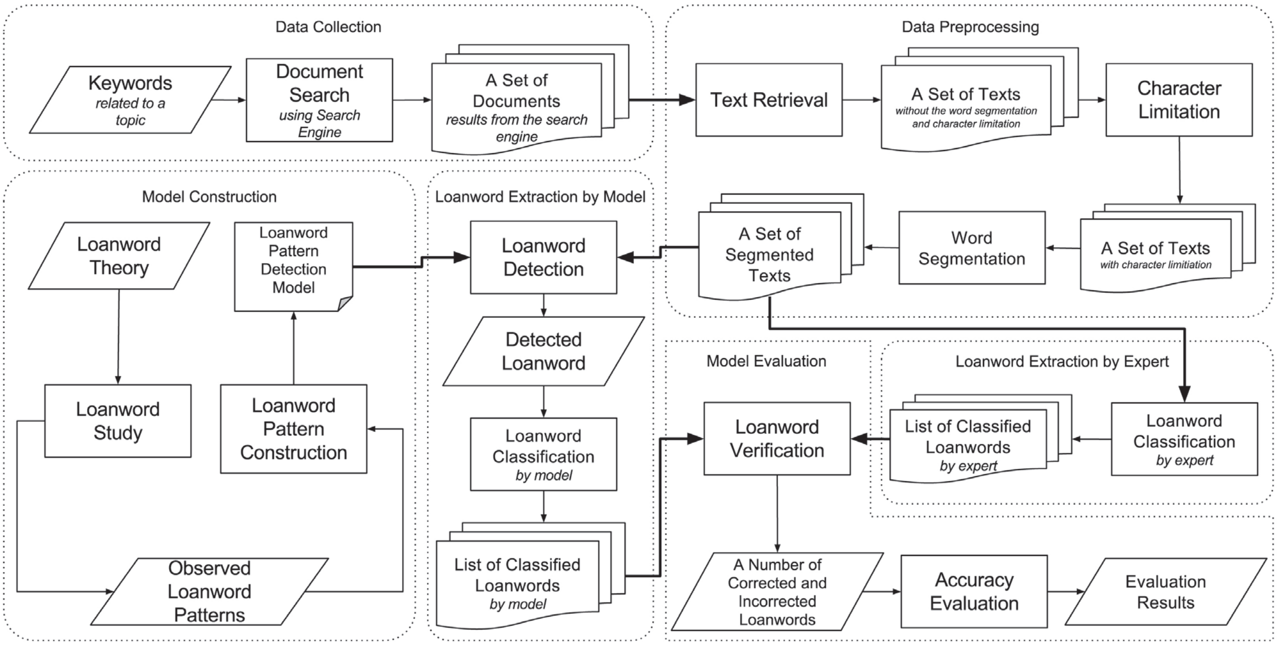

Data Collection
Keywords related to a topic
Document Search using Search Engine
A Set of Documents results from the search engine
Data Preprocessing
Text Retrieval
A Set of Texts without the word segmentation and character limitation
Character Limitation
Model Construction
Loanword Theory
Loanword Pattern Detection Model
Loanword Extraction by Model
Loanword Detection
A Set of Segmented Texts
Word Segmentation
A Set of Texts with character limitation
Detected Loanword
Loanword Study
Loanword Pattern Construction
Model Evaluation
Loanword Extraction by Expert
Loanword Classification by model
Loanword Verification
List of Classified Loanwords by expert
Loanword Classification by expert
Observed Loanword Patterns
List of Classified Loanwords by model
A Number of Corrected and Incorrected Loanwords
Accuracy Evaluation
Evaluation Results

Figure 1 Framework.

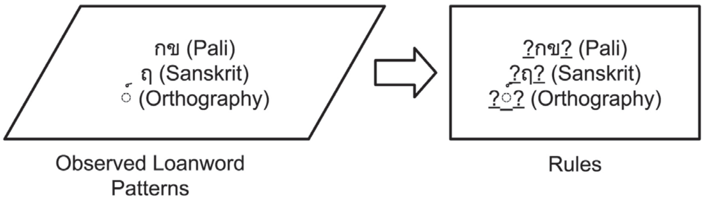


Figure 2 Loanword pattern construction.

Table 2 Statistics of patterns.

Loanword	No. of patterns	Examples
Pali word (P)	47	กก, กข, จจ, จฉ, ฏฏ, ฏฐ, ตต, ตถ, ปป, ปผ, คค, คฆ, ชช, ทธ, พพ, พภ, งก, งข, งค, งฆ, ญจ, ญช, ญญ, ณฏ, ณฐ, ณฑ, นต, นถ, นท, นธ, นน, มป, มผ, มพ, มภ, มม, ลล, ลย, ลห, วห, ฬห, สส, สต, สน, สม, สย, สว
Sanskrit word (S)	25	ฤ, ฦ, ศ, ษ, รร, ชญ, กร, รค, คร, ตย, ทย, นย, ธย, ณย, สก, รถ, ตร, ปต, ตว, ทร, ปร, ลป, รป, รม, ฑ
Orthography (O)	1	◌์

Loanword extraction by model

This process extracts the position of loanwords from the set of segmented texts using the loanword pattern detection model (Loanword Detection). The detected positions of loanword then are classified as type of loanword using the classification model (Loanword Classification). In this work, we propose 7 types of possible loanwords and another type as non-loanword as shown in **Table 3**.

Table 3 Types of loanword and non-loanword.

Type (Tag)	Definition	Example
Pali (P)	A Thai word is loaned from Pali.	ปัจจุบัน (Current)
Sanskrit (S)	A Thai word is loaned from Sanskrit.	กันยายน (September)
Orthography (O)	A Thai word which has its orthography.	ออนไลน์ (Online)
Pali & Sanskrit (PS)	A Thai word is loaned from Pali and Sanskrit.	ประสิทธิภาพ (Efficiency)
Pali & Orthography (PO)	A Thai word which has its orthography and is loaned from Pali.	กลยุทธ์ (Strategy)
Sanskrit & Orthography (SO)	A Thai word which has its orthography and is loaned from Sanskrit.	วิเคราะห์ (Analyze)
Pali, Sanskrit, Orthography (PSO)	A Thai word which has its orthography and is loaned from Pali and Sanskrit.	ประสิทธิ์ (Success)
Other	A Thai word which has not its orthography and is not loaned from Pali and Sanskrit.	ข้อมูล (Data)

Loanword extraction by expert

This process prepares a set of corrected loanwords for model evaluation. A set of texts which are segmented from the data preprocessing process are used for loanword classification by experts. Experts mark loanwords' position and assign loanwords' type (Loanword Classification by expert). This process results are the position and type of loanwords that are tagged.

Model evaluation

This process evaluates the results of loanword extraction. List of classified loanwords by model and list of classified loanwords by experts are compared and the number of corrected and uncorrected loanwords are calculated (Loanword Verification). Accuracy and the F-measure values are computed for assessing the loanword extraction model using a confusion matrix as shown in **Table 4** [11].

Table 4 Confusion matrix.

Actual class / Predict class	C1	¬ C1	Total
C1	True Positives (TP)	False Negatives (FN)	P
¬ C1	False Positives (FP)	True Negatives (TN)	N
Total	P'	N'	P+N

$$Accuracy = \frac{TP + TN}{P + N} \tag{1}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

To evaluate our model, accuracy, the F-measure, precision, and recall can be calculated by using Eqs. (1) - (4), respectively. From the confusion matrix, positive value (P) is the total number of loanwords provided by the experts. Negative value (N) is the total number of non-loanwords answered by the experts. True positive value (TP) is the number of corrected loanwords detected by the model when compared to P. False positive value (FP) is the number of uncorrected loanwords detected by the model when compared to P. True negative value (TN) is the number of corrected non-loanwords that detected by the model when compared to N. False negative value (FN) is the number of uncorrected non-loanwords detected by the model when compared to N.

Experimental setting

In this section, all processes from the proposed framework are implemented. In data collection, 2 topics are defined i.e., data mining, and iOS operating systems. Two sets of 11 and 19 related documents in regard to each topic are respectively collected from the Google search engine [12,13] using keywords as shown in **Table 5**.

Table 5 Statistics of collected documents.

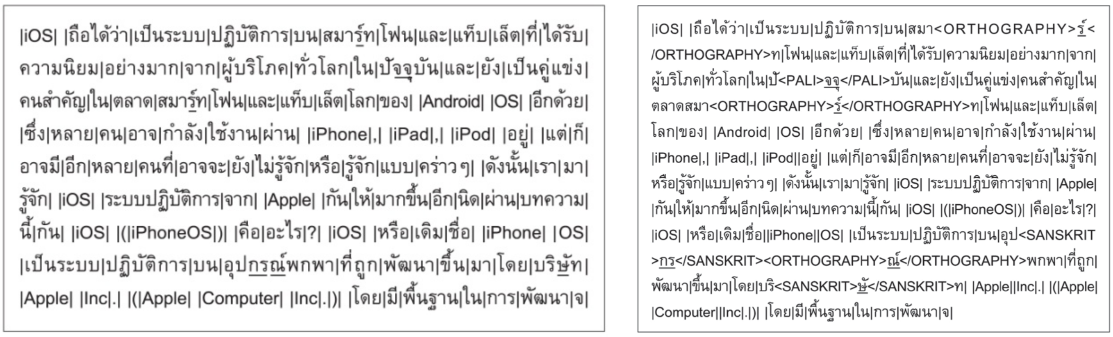
Topic	Keywords	No. of documents	No. of characters (Without character limitation)	No. of characters per document (With character limitation)
Data mining	datamining คือ, data mining คือ, การทำเหมืองข้อมูล, ดาต้าไมนิ่ง คืออะไร	11	27,852	2,532
iOS operating system	ios คือ, ระบบปฏิบัติการ ios	19	10,336	544

Data preprocessing

Only texts from the queried documents in prior process are extracted (Text Retrieval). For text clustering purpose, number of characters in each document in the document set should be equal. The number of characters is equal to the minimum number of characters in a document (Character Limitation). Word segmentation is then applied and used in the next step (Word Segmentation). In this paper, longest matching technique is exploited for segmenting words [14].

Loanwords extraction by expert

In this work, 3 Thai native readers as experts are asked to identify and tag loanwords found in the Thai texts (Loanword Classification by expert) using the similar theory of loanword pattern construction from the model construction process. An example list of loanwords detected by experts has shown in **Figure 3**.



|iOS| |ถือได้ว่า|เป็นระบบ|ปฏิบัติการ|บน|สมาร์ท|โฟน|และ|แท็บ|เล็ต|ที่|ได้รับ|
ความนิยม|อย่างมาก|จาก|ผู้บริโภค|ทั่วโลก|ใน|ปัจจุบัน|และ|ยัง|เป็นคู่แข่ง|
คนสำคัญ|ใน|ตลาด|สมาร์ท|โฟน|และ|แท็บ|เล็ต|โลก|ของ| |Android| |OS| |อีกด้วย|
|ซึ่ง|หลาย|คน|อาจ|กำลัง|ใช้งาน|ผ่าน| |iPhone|,| |iPad|,| |iPod| |อยู่| |แต่|ก็|
อาจมี|อีก|หลาย|คนที่|อาจจะ|ยัง|ไม่รู้จัก|หรือ|รู้จัก|แบบ|คร่าว ๆ| |ดังนั้น|เรา|มา|
รู้จัก| |iOS| |ระบบปฏิบัติการ|จาก| |Apple| |กัน|ให้|มากขึ้น|อีก|นิด|ผ่าน|บทความ|
นี้|กัน| |iOS| |(|iPhoneOS|)| |คือ|อะไร|?| |iOS| |หรือ|เดิม|ชื่อ| |iPhone| |OS|
|เป็นระบบ|ปฏิบัติการ|บน|อุปกรณ์พกพา|ที่ถูก|พัฒนา|ขึ้น|มา|โดย|บริษัท|
|Apple| |Inc|.| |(|Apple| |Computer| |Inc|.|)| |โดย|มี|พื้นฐาน|ใน|การ|พัฒนา|จ|

(a)

|iOS| |ถือได้ว่า|เป็นระบบ|ปฏิบัติการ|บน|สมา<ORTHOGRAPHY>ร์<
/ORTHOGRAPHY>ท|โฟน|และ|แท็บ|เล็ต|ที่|ได้รับ|ความนิยม|อย่างมาก|จาก|
ผู้บริโภค|ทั่วโลก|ใน|ปั<PALI>จจุ</PALI>บัน|และ|ยัง|เป็นคู่แข่ง|คนสำคัญ|ใน|
ตลาดสมา<ORTHOGRAPHY>ร์</ORTHOGRAPHY>ท|โฟน|และ|แท็บ|เล็ต|
โลก|ของ| |Android| |OS| |อีกด้วย| |ซึ่ง|หลาย|คน|อาจ|กำลัง|ใช้งาน|ผ่าน|
|iPhone|,| |iPad|,| |iPod||อยู่| |แต่|ก็|อาจมี|อีก|หลาย|คนที่|อาจจะ|ยัง|ไม่รู้จัก|
หรือ|รู้จัก|แบบ|คร่าว ๆ| |ดังนั้น|เรา|มา|รู้จัก| |iOS| |ระบบปฏิบัติการ|จาก| |Apple|
|กัน|ให้|มากขึ้น|อีก|นิด|ผ่าน|บทความ|นี้|กัน| |iOS| |(|iPhoneOS|)| |คือ|อะไร|?|
|iOS| |หรือ|เดิม|ชื่อ||iPhone||OS| |เป็นระบบ|ปฏิบัติการ|บน|อุป<SANSKRIT
>กร</SANSKRIT><ORTHOGRAPHY>ณ์</ORTHOGRAPHY>พกพา|ที่ถูก|
พัฒนา|ขึ้น|มา|โดย|บริ<SANSKRIT>ษั</SANSKRIT>ท| |Apple||Inc|.| |(|Apple|
|Computer||Inc|.|)| |โดย|มี|พื้นฐาน|ใน|การ|พัฒนา|จ|

(b)

Figure 3 Loanword extraction (a) loanword detection and (b) loanword classification.

We constructed 73 patterns (47 patterns for Pali, 25 patterns for Sanskrit, and one pattern for orthography). In this experiment, there were 26 patterns found in the collected documents as shown in **Table 6**. The number of Pali and Sanskrit patterns are similar (13 and 12 patterns for Pali and Sanskrit, respectively). One pattern can match several loanwords and a loanword can be extracted by several patterns. Loanword types extracted by experts in this experiment are described as shown in **Table 7**.

Table 6 Statistics of loanword patterns from the collected documents.

Loanword	No. constructed patterns	No. of matched patterns
Pali (P)	47	13
Sanskrit (S)	25	12
Orthography (O)	1	1

Table 7 Statistics of loanwords extracted by experts.

Loanwords	Topic		The average number of loanword types
	iOS operating system	Data mining	
Pali (P)	1	5	3.00
Sanskrit (S)	11	14	12.50
Orthography (O)	14	4	9.00
Pali & Sanskrit (PS)	1	5	3.00
Pali & Orthography (PO)	0	11	5.50
Sanskrit & Orthography (SO)	6	24	15.00
Pali, Sanskrit, Orthography (PSO)	0	0	0.00
The average number of loanword types based on topic	4.71	9.00	

Results and discussion

The results of loanword extraction, which are the F-measure, precision, and recall, are shown in **Tables 8 - 10**.

Table 8 Average precision of loanword extraction.

Loanwords	**The average precision of related documents in each topic (%)**		**The average precision based on loanwords (%)**
	iOS operating system	**Data mining**	
Pali (P)	60.00	25.71	42.86
Sanskrit (S)	49.29	43.68	46.48
Orthography (O)	80.00	40.00	60.00
Pali & Sanskrit (PS)	100.00	100.00	100.00
Pali & Orthography (PO)	20.00	100.00	60.00
Sanskrit & Orthography (SO)	76.00	100.00	88.00
Pali, Sanskrit, Orthography (PSO)	100.00	40.00	70.00
The average precision based on topic (%)	69.33	64.20	

Table 8 shows the average precision for extracting loanwords from 2 topics i.e., iOS operating system and data mining. Thai words which are loaned from Pali and Sanskrit can be extracted with 100 % precision from both topics. Moreover, extracting Thai words which have their orthography and are loaned from Pali (PO) or Sanskrit (SO) from the data mining topic achieved up to 100 % of the precision. However extracting Thai words which are loaned from Pali (P) or Sanskrit (S) only gained the lowest precision (42.86 % from both topics).

Table 9 Average recall of loanword extract.

	The average recall of related documents in each topic (%)		**The average recall based on loanwords (%)**
	iOS operating system	**Data mining**	
Pali (P)	60.00	60.00	60.00
Sanskrit (S)	100.00	80.00	90.00
Orthography (O)	48.00	40.00	44.00
Pali & Sanskrit (PS)	100.00	100.00	100.00
Pali & Orthography (PO)	20.00	100.00	60.00
Sanskrit & Orthography (SO)	80.00	96.00	88.00
Pali, Sanskrit, Orthography (PSO)	100.00	40.00	70.00
The average recall based on topic (%)	72.57	73.71	

Table 9 shows the average recall for loanword extraction from 2 topics. On the average, extracting loanwords gained 72.57 and 73.71 % recall from iOS operating system and data mining topics, respectively. Thai words which are loaned from Pali and Sanskrit (PS) can be extracted with 100 % recall from both topics, followed by extracting Thai words which are loaned from Sanskrit (S) (90 % recall). However, extracting Thai words which have their orthography achieved the lowest recall up to 44 % on average from both topics.

Table 10 Average F-measure of loanword extraction.

Loanwords	**The average F-measure of related documents in each topic (%)**		**The average F-measure based on loanwords (%)**
	iOS operating system	**Data mining**	
Pali (P)	60.00	35.56	47.78
Sanskrit (S)	64.67	54.19	59.43
Orthography (O)	58.10	40.00	49.05
Pali & Sanskrit (PS)	100.00	100.00	100.00
Pali & Orthography (PO)	20.00	100.00	60.00
Sanskrit & Orthography (SO)	77.78	97.78	87.78
Pali, Sanskrit, Orthography (PSO)	100.00	40.00	70.00
The average F-measure based on topic (%)	68.65	66.79	

Table 10 shows the average F-measure for extracting loanwords from 2 topics. For iOS operating system topic, extracting Thai words which have their orthography and are loaned from Pali and Sanskrit (PSO) achieved the highest F-measure up to 100 % while extracting Thai words which have their orthography and are loaned from Pali achieved the lowest F-measure up to 20 %. However, for the data mining topic, extracting Thai words which are loaned from Pali and Sanskrit (PS) or Pali and orthography (PO) gained 100 % the F-measure. On average of 2 topics, extracting Thai words which are loaned from Pali and Sanskrit achieved the highest F-measure (100 %), followed by extracting Thai words which have their orthography and are loaned from Sanskrit (87.78 %).

Table 11 Average accuracy of loanword extraction.

Topic	**Average of accuracy (%)**
iOS operating system	96.28
Data mining	96.30
The average accuracy based on all topics	96.29

From **Table 11**, the average accuracy from extracting loanword using our extraction model is 96.29 %. From **Figure 4**, we can conclude that extracting a Thai word which is loaned from Pali and Sanskrit (PS) can achieve 100 % of the F-measure. The F-measure of loanword extraction from both topics gains more than 66 %. The lowest F-measure of obtained from extracting Thai words which are loaned from Pali (P) and Thai words which only have their orthography (O).

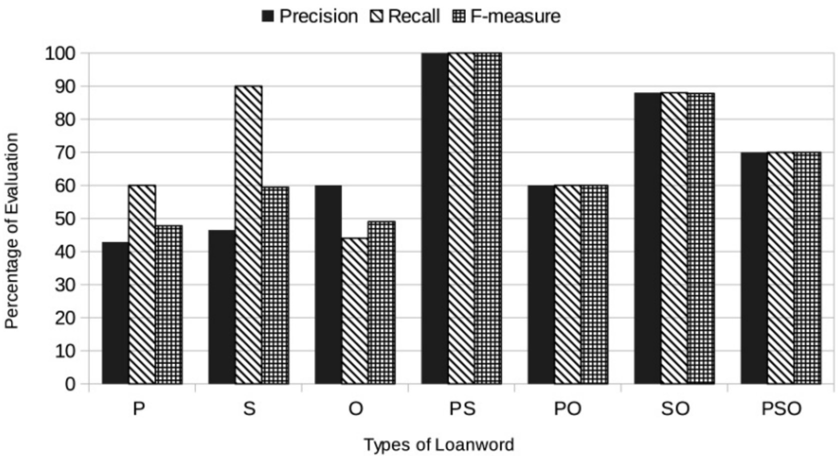


Figure 4 Evaluation summary of loanword extraction.

The loanword extraction model in this work is originally constructed by analyzing loanword theory in the Thai language. However, from the experimental results, the loanword theory of Pali or Sanskrit may not comply with the Thai language at present. There are some transliterated words and unknown words which decrease the correctness of extracting loanwords. Transliterated words are words from one language transliterated into another language, while unknown words or out of vocabulary are undefined words or misspelling words. These words affect to those loanword rules. In order to obtain higher performance of loanword extraction, recognizing transliterated words and unknown words should take into account. Our proposed framework can be applied to extract loanwords in other syllabic alphabetic languages such as Khmer, Lao, and Burmese since these languages have common writing structure.

Conclusions and future work

In this paper, the framework of loanword extraction, which has 6 processes, is proposed. Loanwords in the Thai language are studied and analyzed based on the loanword theory. Seven types of loanword are classified. The experimental result shows that the Thai words which are loaned not only from Pali but also Sanskrit achieved the highest F-measure up to 100 and 96.29 % accuracy. However, extracting other types of loanword needs to be improved by considering transliterated words and unknown words. The loanword estimation can be used for clustering the readability of texts.

References

[1] National Statistical Office, Available at: http://www.nso.go.th, accessed August 2016.

[2] K Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL Int. J. Appl. Linguist*. 2014; **165**, 97-135.

[3] P Daowadung and YH Chen. Using word segmentation and SVM to assess readability of Thai text for primary school students. *In*: Proceedings of the 8th International Joint Conference on Computer Science and Software Engineering. Nakhon Pathom, Thailand, 2011, p. 170-4.

[4] N Tongtep, F Coenen and T Theeramunkong. Content-based readability assessment: A study using a syllabic alphabetic language (*in Thai*). *In*: Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence. Gold Coast, Australia, 2014, p. 863-70.

[5] Samut Prakan School, Available at: http://www.prakan.ac.th, accessed May 2016.

[6] S Phongphaiboon. *Principles of Thai Language.* Thai Watana Panich, Bangkok, 1991, p. 1-14.

[7] K Tonglo. *Principles of Thai Language*. Ruam Sarn, Bangkok, 2007, p. 88-140.

[8] S Makjeng. *Pali and Sanskrit Language in Thai Language*. Odeon Store, Bangkok, 1992, p. 12-4.

[9] YH Chen and P Daowadung. Assessing readability of Thai text using support vector machines. *Maejo Int. J. Sci. Tech*. 2015; **9**, 355-69.

[10] N Tongtep and T Theeramunkong. Simultaneous character-cluster-based word segmentation and named entity recognition in Thai language. *In*: Proceedings of the 5th International Conference on Knowledge, Information, and Creativity Support Systems. Chiang Mai, Thailand, 2011, p. 216-25.

[11] J Han, M Kamber and J Pei. *Data Mining: Concepts and Techniques*. Elsevier, MA, 2011, p. 364-9.

[12] Google, Available at: https://www.google.co.th, accessed January 2016.

[13] Statista, Available at: http://www.statista.com, accessed June 2016.

[14] National Electronics and Computer Technology Center, Available at: http://www.sansarn.com, accessed June 2016.