PAPER *Special Section/Issue on ******

Statistical-based Approach to Non-segmented Language Processing

Virach Sornlertlamvanich[†], Thatsanee Charoenporn[†], Shisanu Tongchim[†], Canasai Kruengkrai[††] and Hitoshi Isahara[††]

Summary
Several approaches have been studied to cope with the exceptional features of non-segmented languages. When there is no explicit information about the boundary of a word, segmenting an input text is a formidable task in language processing. Not only the contemporary word list, but also usages of the words have to be maintained to cover the use in the current texts. The accuracy and efficiency in higher processing do heavily rely on this word boundary identification task. In this paper, we introduce some statistical based approaches to tackle the problem due to the ambiguity in word segmentation. The word boundary identification problem is then defined as a part of others for performing the unified language processing in total. To exhibit the ability in conducting the unified language processing, we selectively study the tasks of language identification, word extraction, and dictionary-less search engine.

Key words:
non-segmented language, unified language processing, statistical approach, probability, language identification, word extraction, search engine

1. Introduction

Disambiguation is our major concern in natural language processing. Though the morphological and syntactic information play an important role in assisting the disambiguation process, degree of the allowing information for taking into account can vary according to the type of the language. For instance, a space character between words reduces the task in identifying word boundary. Similarly, the grammatical markers, inflection and punctuation marks are quite meaningful information for identifying the role of each word and structural relations between words in a sentence. Consequently, it is natural to say that many approaches in language processing have been studied in terms of word units. Definitively, a word unit is a unit of language that native speakers can identify. Based on the classical approaches, without knowing the entity of word, it is not so efficient to develop the language model. For the language that does not allow to use a space character (or any special markers) to separate words in natural text, so-called a non-segmented language, i.e. Thai, Chinese, Japanese, Korean, etc., a high performance word segmentation algorithm to identify word boundary is crucial. The performance of language processing totally relies on the efficiency and the accuracy of the word segmentation algorithm.

In our recent research, we proposed a language interpretation model to deal with an input text as a byte sequence rather than a sequence of words. It is an approach to unify the language processing model to cope with the ambiguities in word determination problem. The approach takes an input text in the earliest stage of language processing when the exhaustive recognition of total word identity is not necessary.

In this paper, we present the achievements in identifying language based on the study of 20 different languages, word candidates extraction based on the unified input byte sequence, and indexing algorithm for full text retrieval applied in a search engine. Our experiments also show promising results for overcoming the drawbacks of the non-segmented language.

2. Language Identification

Language identification is yet another challenging task when it is going to be conducted without any grammatical knowledge. Byte sequence is the only magic key in our approach to determine the language of the input text. We introduce string kernel for this language identification task. We conducted experiments using 2 kernelized versions of centroid-based method [11] and support vector machine (SVM) [12]. The accuracies of identification are acceptable for both methods. The accuracies reach 95 percent with only 10 training sets (2 KB per set). It is also found that the simple centroid-based classifier is comparable to the SVM classifier based on the string kernel.

The centroid-based method is comparably simpler than SVM method. [11]'s algorithm constructs the language profile from training data, but we use the centroid (or mean) vector as the profile and apply the squared Euclidean distance instead of the out-of-space measure [13]. Training samples are pre-categorized according to their languages, forming the disjoint subsets

Manuscript received January 2001.
Manuscript revised March 1, 2001.
[†] The author is with TCL, NICT Asia Research Center, Pathumthani, Thailand.
[††] The author is NICT, Kyoto, Japan.

$C_1, \ldots, C_m$. Using the mapping function Φ, the centroid vector for each C_j can be calculated by:

$$m_j = \frac{1}{|C_j|}\sum_{v \in C_j} \Phi(v) \quad (1)$$

We can classify a new string $\boldsymbol{u}$ as a member of the centroid C^* with the minimum squared Euclidean distance:

$$C^* = \arg\min_j \left\| \Phi(u) - m_j \right\|^2 \quad (2)$$

By replacing (1) in (2), we obtain:

$$\left\| \Phi(u) - \frac{1}{|C_j|}\sum_{v \in C_j} \Phi(v) \right\|^2 \quad (3)$$

The equation (3) can then be expanded and rewritten in terms of $K_r(\cdot,\cdot)$ for inner product terms:

$$K_r(u,u) - \frac{2}{|C_j|}\sum_{v \in C_j} K_r(u,v) + \frac{1}{|C_j|^2}\sum_{v,w \in C_j} K_r(u,w) \quad (4)$$

For the SVM method, we apply it based on the structural risk minimization principle from the statistical learning theory [12]. Given training samples (v_1,y_1), …, (v_l,y_l), in order to obtain the hyper plane having the maximum margin with the closest training samples, we consider the following primal optimization problem:

$$\textit{minimize:} \quad V(w,b,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \quad (5)$$

where ξ_i is a slack variable, and C is a parameter for controlling the tradeoff between the training error and the maximum margin.

Given a new string u, we classify it by using the following decision function:

$$f(u) = \mathrm{sgn}\left(\sum_i \alpha_i K_r(v_i,u) + b\right) \quad (6)$$

String kernel is introduced to compute the common subsequences of two strings. A kernel can be simply thought of as the inner product function between two vectors, $K(x, y) = \langle \Phi(x), \Phi(y) \rangle$. With the recent advance in kernel methods, the kernel computation is not just limited to vectorial objects, but can be performed on sequences based on the so-called string kernels [2][9].

To describe how to compute the string kernel based on our proposed efficient method of using fast matching with suffix tree [10], we consider two strings of $\boldsymbol{u}$ = yzxxz and $\boldsymbol{v}$ = xyzxxxy, where the range of considering substring length (r) is $1 \le r \le 2$ and the set of characters (Σ) is $\Sigma = \{x,y,z\}$. Fig. 1 demonstrates the suffix tree for the string $\boldsymbol{v}$ = xyzxxxy.

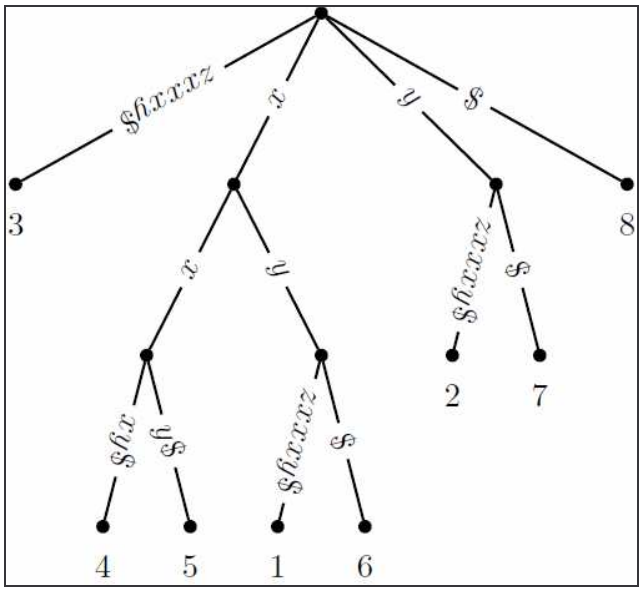

Fig. 1 Suffix tree for the string $\boldsymbol{v}$ = *xyzxxxy*$.

Focusing on all substrings of the maximum length r, the string kernel for the dynamic alignment can be expressed as [14]:

$$K_r(u,v) = \sum_{i=1}^{r} \lambda^{2i} \sum_{s \in \sum^i} \sum_{i \in \Gamma_{s,u}} \sum_{j \in \Gamma_{s,v}} I(s = u[i] = v[j]) \quad (7)$$

where

Γ is the set of lexicon
s is a word in Γ
I() is the indicator function
i, *j* are index vectors of $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively

The following table lists all the matched between $\boldsymbol{u}$ and $\boldsymbol{v}$, where the number of all occurrences of substrings in common is given in parentheses. The scores are computed by using the above equation. By summing all the scores, we get the value of kernel, $K_r(u,v) = 12\lambda^2 + 4\lambda^4$.

u	v[i]	Score
y	y(2), yz(1)	$2\cdot\lambda^{2\cdot1} + 1\cdot\lambda^{2\cdot2} = 2\lambda^2 + \lambda^4$
z	z(1), zx(1)	$1\cdot\lambda^{2\cdot1} + 1\cdot\lambda^{2\cdot2} = \lambda^2 + \lambda^4$
x	x(4), xx(2)	$4\cdot\lambda^{2\cdot1} + 2\cdot\lambda^{2\cdot2} = 4\lambda^2 + 2\lambda^4$
x	x(4), xz(0)	$4\cdot\lambda^{2\cdot1} = 4\lambda^2$
z	z(1)	$1\cdot\lambda^{2\cdot1} = \lambda^2$

As a result of applying suffix tree to the substring matching, we can compute the string kernels with the computational complexity of $O(c|\boldsymbol{u}|+|\boldsymbol{v}|)$, where $c=r+k$, *r* is the length of substring (length of the predefined character gram) and *k* is the occurrence of a substring of $\boldsymbol{u}$ in $\boldsymbol{v}$.

We conducted experiments on 4 groups of languages i.e. Multi4 (Thai, Chinese, Japanese, Korean), Multi8-1 (English, French, Italian, Portuguese, Spanish, Swedish, German, Hungarian), Multi8-2 (Czech, Polish, Croatian, Slovak, Slovenian, Bulgarian, Russian, Greek) and Multi20 (all the 20 languages). The results are shown in Fig. 2-5.

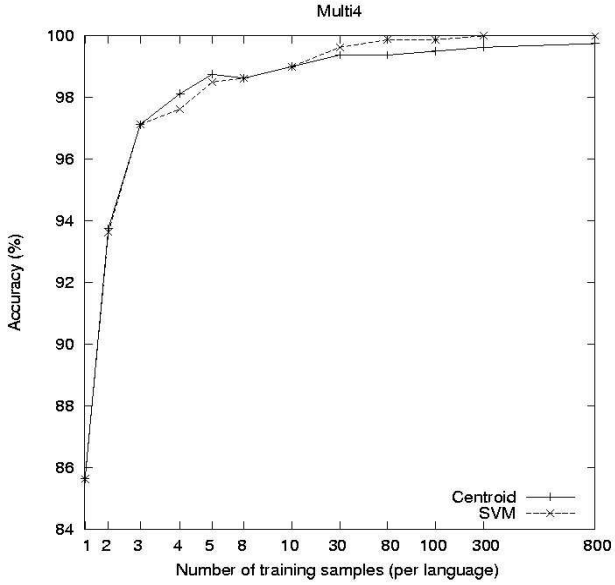

Fig. 2 Classification accuracy on Multi4.

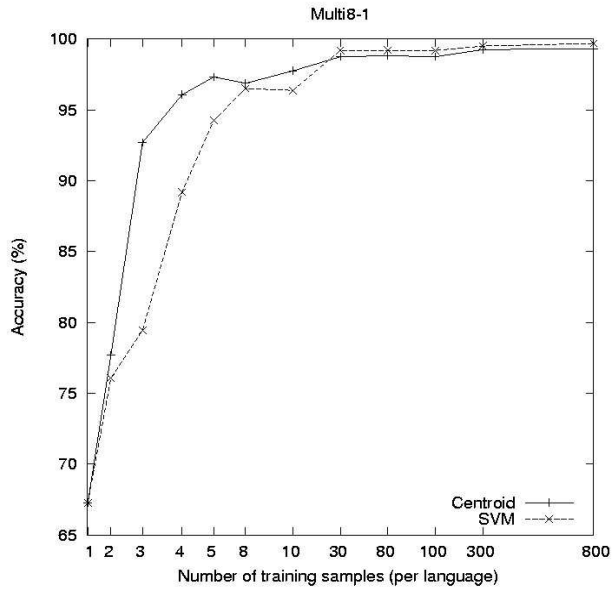

Fig. 3 Classification accuracy on Multi8-1.

In total, comparing to centroid-based method, SVM classifier yields slightly better performance when given more training samples, while the centroid-based classifier performs better for small numbers of training samples.

We hypothesize that, for the SVM classifier, only the critical support vectors are retained after the training process, whereas the centroid-based classifier can exploit the feature combination of the representative centroid that is the mean vector of all training samples. However, both had shown significant results in identifying languages of whatever groups of the languages.

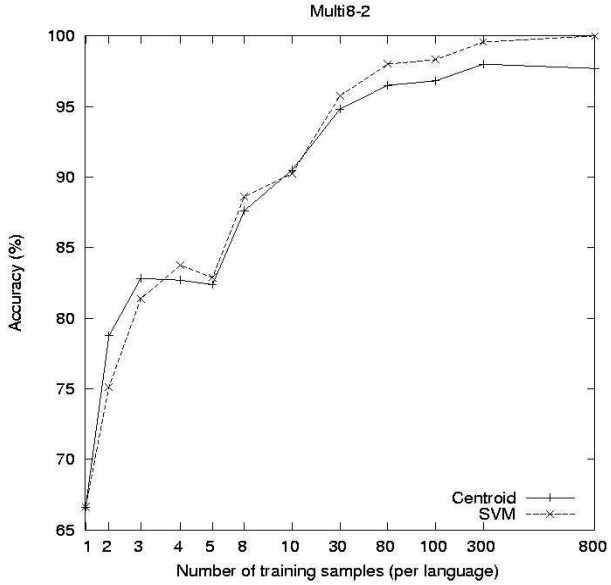

Fig. 4 Classification accuracy on Multi8-2.

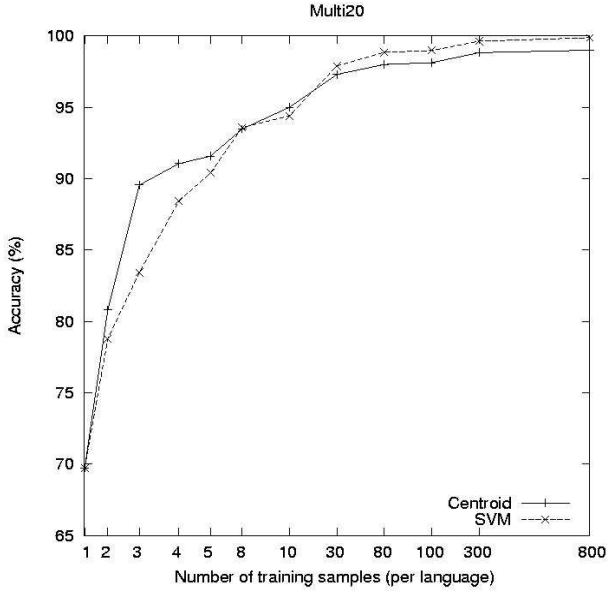

Fig. 5 Classification accuracy on Multi20.

For the purpose of comparison, we used TextCat implemented based on [11] as the baseline. The text collection of Universal Declaration of Human Rights (UDHR) was used to evaluate the performance of both of our approaches i.e. string kernel SVM method (SKSVM),

and kernelized centroid based method (KCB) comparing to TextCat. Ten languages (India10) in UDHR were selected. The difficulty in identifying the languages (Bengali, Gujarati, Hindi, Kannada, Magahi, Marathi, Punjabi, Sansakrit, Saraiki, and Tamil) in this group is the scripts (Devanagari and Arabic) that they share in writing.

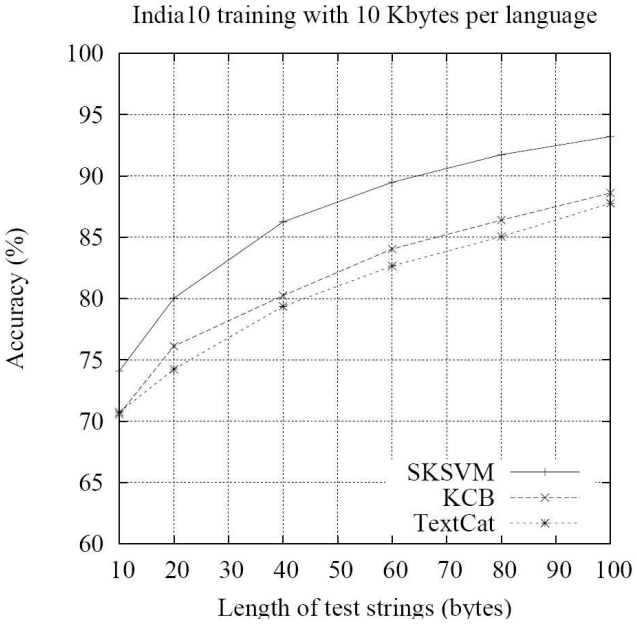

Fig. 6 Accuracy on India10 with 10 Kbytes training data per language.

Fig. 6 shows the accuracy of identification on India10 testing text by varying the length of the test strings. Both of our approaches outperform the TextCat on every length of the test strings. SKSVM shows a significantly higher accuracy than the other two. The results conclude that the string kernel model can very well represent the nature of the language comparing to the simple n-gram model.

3. Word Extraction

It is always arguable about what should be an appropriate word list for a non-segmented language. There is no any explicit rule for explaining what a word looks like. Semantically defining that word is a unit of language that native speakers can identify is still vague. It mostly depends on individual's perception about word i.e. 'bookstore' vs. 'book store', 'open source' vs. 'opensource', 'newspaper' vs. 'news paper', etc..

A preliminary study of co-occurrence of substring has shown a promising result in extracting open compounds from text corpora [7]. The significant change in occurring frequency of a substring when expanded has invoked the possible observation of word boundary. We proposed another method for automatic word extraction from raw texts based on the algorithm that reflected the understanding of word being a string frequently used in most part of the texts [8]. We employed the C4.5 decision tree induction program [4] as the learning algorithm for word extraction. The induction algorithm proceeds by evaluating content of a series of attributes and iteratively building a tree from the attribute values with the leaves of the decision tree being the value of the goal attribute. At each step of learning procedure, the evolving tree is branched on the attribute that partitions the data items with the highest information gain. Branches will be added until all items in the training set are classified. To reduce the effect of overfitting, C4.5 prunes the constructed entire decision tree. It recursively examines each subtree to determine whether replacing it with a leaf or branch would reduce expected error rate. This pruning makes the decision tree better in dealing with the data different from the training data.

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

(i) Left Mutual Information and Right Mutual Information

The left mutual information (*Lm*), and right mutual information (*Rm*) of string xyz are defined as:

$$Lm(x\mathbf{y}z) = \frac{p(x\mathbf{y}z)}{p(x)p(\mathbf{y}z)} \tag{8}$$

$$Rm(x\mathbf{y}z) = \frac{p(x\mathbf{y}z)}{p(x\mathbf{y})p(z)} \tag{9}$$

where

x is the leftmost character of $x\mathbf{y}z$
$\mathbf{y}$ is the middle substring of $x\mathbf{y}z$
z is the rightmost character of $x\mathbf{y}z$
p( ) is the probability function.

If *xyz* is a word, both *Lm(xyz)* and *Rm(xyz)* should be high. On the contrary, if *xyz* is a non-word string but consists of words and characters, either of its left or right mutual information or both must be low. For example, 'กปรากฏ' ('ก' (a Thai alphabet) + 'ปรากฏ' (The word means 'to appear' in Thai.)) must have low left mutual information.

(ii) Left Entropy and Right Entropy

Entropy [5] is the information measuring disorder of variables. The left and right entropy is exploited as another two attributes in our word extraction. Left entropy (*Le*), and right entropy (*Re*) of string y are defined as:

$$Le(\boldsymbol{y}) = -\sum_{\forall x \in \boldsymbol{A}} p(x\boldsymbol{y} \mid \boldsymbol{y}) \cdot log_2 p(x\boldsymbol{y} \mid \boldsymbol{y}) \quad (10)$$

$$Re(\boldsymbol{y}) = -\sum_{\forall z \in \boldsymbol{A}} p(\boldsymbol{y}z \mid \boldsymbol{y}) \cdot log_2 p(\boldsymbol{y}z \mid \boldsymbol{y}) \quad (11)$$

where

y is the considered string,
A is the set of all alphabets
x, *z* is any alphabets in ***A***.

If *y* is a word, the alphabets that come before and after *y* should have varieties or high entropy. If *y* is not a complete word, either of its left or right entropy, or both must be low. For example, 'ปราก' is not a word but a substring of word 'ปรากฏ' (appear). Thus the choices of the right adjacent alphabets to 'ปราก' must be few and the right entropy of 'ปราก', when the right adjacent alphabet is 'ฏ', must be low.

(iii) Frequency

It is obvious that the iterative occurrences of words must be higher than those of non-word strings. String frequency is also useful information for our task. Because the string frequency depends on the size of corpus, we normalize the count of occurrences by dividing by the size of corpus and multiplying by the average value of Thai word length:

$$F(\boldsymbol{s}) = \frac{N(\boldsymbol{s})}{Sc}.Avl \quad (12)$$

where

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

We employed the frequency value as another attribute for the C4.5 learning algorithm.

(iv) Length

Short strings are more likely to happen by chance than long strings. Then, short and long strings should be treated differently in the disambiguation process. Therefore, string length is also used as an attribute for this task.

(v) Functional Words

Functional words such as 'จะ' (will) and 'ก็' (then) are frequently used in Thai texts. These functional words are used often enough to mislead the occurrences of string patterns. To filter out these noisy patterns from word extraction process, discrete attribute Func(s):

*Func(**s**) = 1 if string **s** contains functional words,*
= 0 if otherwise,

is applied.

(vi) First Two and Last Two Characters

A very useful process for our disambiguation is to check whether the considered string complies with Thai spelling rules or not. We employ the words in the Thai Royal Institute dictionary as spelling examples for the first and last two characters. Then we define attributes and for this task as follows.

$$Fc(\boldsymbol{s}) = \frac{N(s_1 s_2 *)}{ND} \quad (13)$$

$$Lc(\boldsymbol{s}) = \frac{N(* s_{n-1} s_n)}{ND} \quad (14)$$

where

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

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

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

ND is the number of words in the dictionary.

We apply [10]'s algorithm to extract all strings from a plain and unlabelled 1-MB corpus which consists of 75 articles from various fields. For practical and reasonable purpose, we select only the 2 to 30 character strings that occur more than 2 times, have positive right and left entropy, and conform to simple Thai spelling rules. To this step, we get about 30,000 strings. These strings are manually tagged as words or non-word strings by native speakers. The strings' statistics explained above are calculated for each string. Then the strings' attributes and tags are used as the training example for the learning algorithm. The decision tree is then constructed from the training data.

In order to test the decision tree, another plain 1-MB corpus (the test corpus), which consists of 72 articles from various fields, is employed. All strings in the test corpus are extracted and filtered out by the same process as used in the training set. After the filtering process, we get about 30,000 strings to be tested. These 30,000 strings are manually tagged in order that the precision and recall of the decision tree can be evaluated.

To measure the accuracy of the algorithm, we consider two statistical values: precision and recall. As shown in Table 1 and 2, the precision of our algorithm is 87.3% for the training set and 84.1% for the test set. The recall of extraction is 56% in both training and test sets.

We compare the recall of our word extraction with the recall from using the Thai Royal Institute dictionary (RID). The recall from our approach and from using RID are comparable and our approach should outperform the existing dictionary for larger corpora. Both precision and recall from training and test sets are quite close. This indicates that the created decision tree is robust for unseen data. Table 3 also shows that more than 30% of the extracted words are not found in RID. These would be the new entries for the dictionary.

Table 1: The precision of word extraction

	Total extracted strings by our approach	Words	Non-words
Training Set	1882 (100%)	1643 (87.3%)	239 (12.7%)
Test Set	1815 (100%)	1526 (84.1%)	289 (15.9%)

Table 2: The recall of word extraction

	Words found in the extracted strings	Words extracted by our approach	Words found in RID
Training Set	2933 (100%)	1643 (56.0%)	1833 (62.5%)
Test Set	2720 (100%)	1526 (56.1%)	1580 (58.1%)

Table 3: Words extracted by our approach and RID

	Words extracted by our approach	Words found in RID	Words not found in RID
Training Set	1643 (100%)	1082 (65.9%)	561 (34.1%)
Test Set	1526 (100%)	1046 (68.5%)	480 (31.5%)

The attributes of such the character-based mutual information and entropy provide significant information to C4.5 algorithm for selecting appropriate candidates for words. The approach greatly supports the process of nominating word candidates for developing a dictionary, and later is extended to fulfill a dictionary-less search engine [6]. The search engine has introduced a word score as a heuristic value to determine the word likelihood of a string. The word score is a normalized value of a mutual information value. The minimum score of the left and right hand side of a string in question is assigned as the word score of the string. Based on the proposed approach, we successfully implemented a multi-lingual search engine with minimum modification.

4. Dictionary-less Search Engine

The performance of dictionary-based search engines is directly affected by the accuracy of word segmentation algorithms. Our previous work [6] discussed about two possible errors affected by the accuracy of dictionary-based word segmentation modules. Assuming that a dictionary contains 6 words: **a**, **b**, **c**, **ac**, **bc** and **cb**.

Case 1: Incorrect word segmentation
The content of the document ***A*** is **abcbcb**. By using a word segmentation module, the content is separated into **a|bc|bc|b**. Assuming that the correct segmentation is **a|b|cb|cb**. If the query is **cb**, it cannot be found in the document ***A*** or if the query is **bc**, the document ***A*** will be incorrectly returned.

Case 2: Unregistered word problem
The content of document ***A*** is **abcdac**. By using the word segmentation module, the content is separated into **a|bc|d|ac**. Assuming that the correct segmentation is **a|b|cd|ac** and **cd** is an unregistered word to the word segmentation. If the query is **bc**, the result from this document will be incorrect or if the query is **cd**, it cannot be found in the document ***A***.

The architecture of our dictionary-less search engine is illustrated in Fig. 7 comparing to the typical dictionary-based search engine in Fig. 8. It is composed of three major modules: (1) data indexing, (2) searching and (3) document ranking.

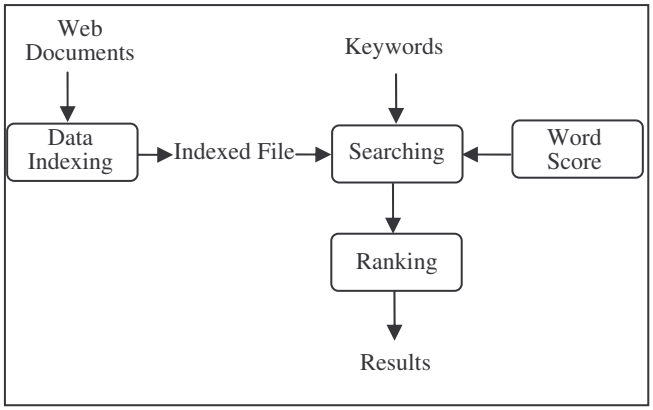

Fig. 7 Dictionary-less Search Engine.

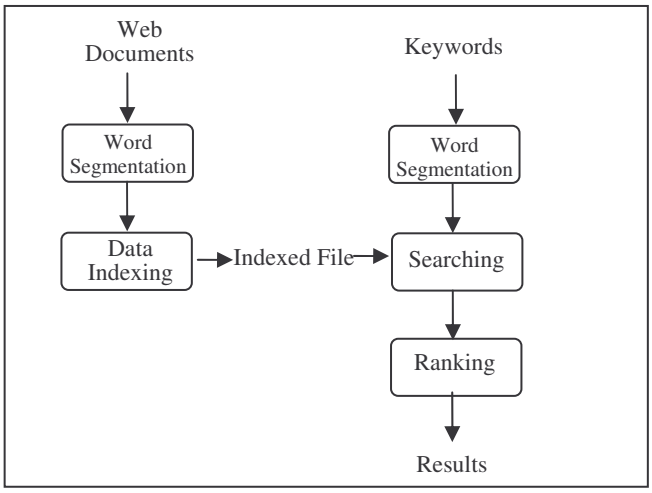

Fig. 8 Dictionary-based Search Engine.

(1) Data Indexing

In typical search engines, web documents are separated into words to provide a word list for generating the indexes. In our approach, the data is considered to be the sequence of characters and indexed character by character. We adopt the enhanced suffix array [10] for indexing the data. All suffixes of the data string are indexed. Thus, the number of indexes is equal to the data size. The advantage of this indexing method is that it guarantees all search strings to be found, whereas the word indexing method depends on the word segmentation. This indexing method can also be applied to other languages since it does not require any dictionary and language-specific knowledge.

(2) Searching

Based on the enhance suffix array, it requires *O(mlogN)* to access the string in the data, where m is a length of the search string and N is the number of indexes. The use of the suffix array guarantees that all search strings will be found. However, only the meaningful strings are preferred. If the found pattern is a part of other word, that pattern is inseparable. As a result, it is not valid as a meaningful word.

For example, assuming that the search query is short and likely to be a part of other strings such as 'ยา' (drug), two strings are found i.e. (1) 'กินยา' (take a drug) and (2) "พัทยา" (Pattaya, name of a district in Thailand). The first string can be separated into two words: (1) "กิน" (take, eat) and (2) "ยา" (drug). Thus, the word "ยา" in the first string is a meaningful word. For the second string, the first part "พัท" is a meaningless string and is strongly connected to the second part "ยา". Thus, the word "ยา" in the second string is meaningless since this string is inseparable.

From the example, the validity of a word can be decided from its surrounding context. If the word is strongly connected to other word and inseparable, it is likely to be a meaningless string. In contrast, the word is likely to be a meaningful string if it is loosely connected to other word and separable.

(3) Word Score

We use mutual information (MI) [1] to measure the degree of the co-occurrence of the query and its context. Let *xy* be a query, *ab* is the left context and *cd* is the right context of the string *xy*, the mutual information can be determined by the Equations 15-18.

$$MI_L(abxy) = \frac{p(abxy)}{p(ab)p(xy)} \tag{15}$$

$$MI_L(abxy) \approx \frac{Count(abxy)}{Count(ab)Count(xy)} \tag{16}$$

$$MI_R(xycd) = \frac{p(xycd)}{p(xy)p(cd)} \tag{17}$$

$$MI_R(xycd) \approx \frac{Count(xycd)}{Count(xy)Count(cd)} \tag{18}$$

If the MI value is high, *xy* is likely to be a part of the context. On the other hand, *xy* should be independent from the context if the MI value is low. We define the inverse of MI as the word score. The word score is calculated by the Equations 19-20.

$$wscoreL(xy|ab) = 1 - norm(MI_L(abxy)) \tag{19}$$

$$wscoreR(xy|cd) = 1 - norm(MI_R(xycd)) \tag{20}$$

The norm(·) is the normalizing function which normalizes the argument from 0 to 1. At this point, the word score determines the probability of being a word of the string. In practice, we use the minimum value of word score for both sides comparing with a threshold to determine the word candidate for a string.

(4) Ranking

The word score from the previous step is not only used to determine the word boundary, it is also used to rank the document. That is, the document with higher word score will attain high rank.

We conduct an experiment in order to compare the dictionary-less search engine with the dictionary-based search engine. The experiment is based on the article [3]. We assign 20 queries to 5 volunteers. For each query, the volunteer is shown the top 10 results from each system. Then, each volunteer will choose the documents that are relevant to the query. Each result is considered to be relevant if at least 3 of the 5 volunteers assigned it as relevant for the query. Finally, the satisfaction score of both systems will be calculated for each query. We define

the satisfaction score as the ratio of the relevant results to all available results.

The web documents used in the experiment can be divided into two groups. The first group is obtained from websites of newspapers, consisting of 5,853 documents (approximately 65 Mb). The second group contains general articles, not related to news. The second group consists of 7,710 documents (approximately 35 Mb).

Table 4: List of queries

	Unsegmented queries	*Segmented queries*
1	บริจาค, สืนามิ	บริจาค, สืนามิ
2	เส้นตาย, ซัดดม	เส้นตาย, ซัดดม
3	มนัส, โชว์ตัว, จีน	มนัส, โชว์ใต้ว, จีน
4	ผลกระทบ, ราคาน้ำมันแพง	ผลกlระlทบ, ราคาlน้ำมันlแพง
5	ไข้หวัดนก	ไข้หวัดlนก
6	ทุจริต, การเลือกตั้ง	ทุจริต, การlเลือกตั้ง
7	จับกุม, ผู้ก่อการร้าย, ภาคใต้	จับกุม, ผู้ก่อการร้าย, ภาคlใต้
8	นโยบาย, แก้ไข, ปัญหายาเสพติด	นโยบาย, แก้ไข, ปัญหาlยาlเสพติด
9	ทดลอง, ลดค่าทางด่วน	ทดลอง, ลดlค่าlทางด่วน
10	ซื้อคืน, สัมปทาน, รถไฟฟ้า	ซื้อlคืน, สัมปทาน, รถไฟฟ้า
11	ลงทุน, ในพม่า	ลงทุน, ในlพม่า
12	ส่งเสริม, การท่องเที่ยว, ไทย	ส่งเสริม, การlท่องเที่ยว, ไทย
13	เลือกตั้ง, ประธานาธิบดี, ปาเลสไตน์	เลือกตั้ง, ประธานาธิบดี, ปาlเลสไตน์
14	เลือกตั้ง, ผู้ว่า, กทม.	เลือกตั้ง, ผู้lว่า, กทมl.
15	สินค้าไทย, การส่งออก	สินค้าlไทย, การlส่งlออก
16	แปรรูปรัฐวิสาหกิจ	แปรรูปlรัฐวิสาหกิจ
17	อุ้ม, นายสมชาย	อุ้ม, นายlสมlชาย
18	ฉลองปีใหม่	ฉลองlปีlใหม่
19	พรทิพย์, ลาออก	พรlทิพย์, ลาlออก
20	สวนสนุก	สวนlสนุก

The test queries are listed in Table 4. The queries are related to news. Thus, only results from the first group are preferred. The satisfaction score of two systems for each query is presented in Table 5. The average of the satisfaction score of the dictionary-less search engine is higher than that of the dictionary-based approach. Furthermore, there are 10 queries that the dictionary-less approach is better than the dictionary-based approach, while the results of 5 queries are equal and the dictionary-based approach achieves higher results on other 5 queries. The differences of the satisfaction score are high when the results of dictionary-less approach are superior comparing to when the results of the dictionary-based approach are superior. The difference of the satisfaction score also insists that some inferior results of the dictionary-less approach are acceptable.

We observe that the dictionary-based search engine faces difficult situation when the query is excessively segmented, and the segmented words are likely to be general terms. For example, one of the queries is "การส่งออก" (export). The word segmentation module separates this word as "การ | ส่ง | ออก". All three terms still have some meaning in Thai, but not directly relevant to the compound word. Moreover, these terms are general words and often parts of several words. Thus, the dictionary-based approach tends to return irrelevant documents, but have several locations of these general terms.

Table 5: Satisfaction score of the test queries

Query No.	Dictionary -less	Dictionary -based	Difference
1	1	0.8	0.2
2	1	1	0
3	1	1	0
4	1	0.8	0.2
5	1	1	0
6	0.8	0.8	0
7*	0.2	0.68	-0.48
8*	0.75	1	-0.25
9	1	0.8	0.2
10*	0.68	1	-0.32
11*	0.25	0.8	-0.65
12	0.6	0	0.6
13*	0.6	1	-0.4
14	1	0	1
15	0.8	0.2	0.6
16	1	1	0
17	1	0.4	0.6
18	1	0.4	0.6
19	0.6	0.4	0.2
20	0.8	0.2	0.6
Average	0.8	0.66	0.14

Another observation is that the incorrect segmentation does not always affect the search performance of the dictionary-based search engine. For example, a part of one query is "ผลกระทบ" (effect). It is incorrectly segmented into "ผลก | ระ | ทบ". All three terms are meaningless and not general terms. However, the dictionary-based search engine still effectively discovers these terms since some of this terms are quite unique. We also observe that the correctness of word segmentation is less important than the generality of segmented words. When the query is excessively segmented, the dictionary-based search engine still performs well if the segmented terms are not quite general. In contrast, the dictionary-based search engine tends to return irrelevant documents if the query is excessively segmented and the segmented words are general terms. This also explains why the dictionary-based search engine performs better on some queries. Although those queries are sometime incorrectly

segmented, the dictionary-based search engine still finds related documents. The reason is that the segmented words are not general. Thus, these words are easily found.

6. Summary and Future Work

The tasks of language identification, word extraction, and dictionary-less search engine had been selected to study by means of the proposed model for unified language processing. The tasks had been evaluated and resulted in a significant performance. Therefore, a non-segmented language can be efficiently processed without relying on the word boundary information at all. It shows that the footprint of byte sequence of a language is sufficient for processing an input text. It is a direct input with most reliable information. The advantages in these statistical based approaches are also a fundamental work for unifying multi-lingual tasks where language dependent parts can be lessened. In terms of implementation efficiency, we still need to study more especially in case of memory space and computational time consuming. These topics are left for future improvement.

References

[1] Church, K. W., Robert, L., and Mark, L., "A status report on ACL/DCL," In Proceedings of the seventh Annual Conference of the UW Centre New OED and Text Research: Using Corpora, pp. 84–91, 1991.
[2] Haussler, D., "Convolution kernels on discrete structures," Technical Report UCSC-CRL-99-10,Department of Computer Science, University of California at Santa Cruz, 1999.
[3] Haveliwala, T., "Topic-sensitive pagerank," In Proceedings of the Eleventh International Conference on World Wide Web, pp. 517–526, 2002.
[4] Quinlan, J. R., "C4.5 Programs for Machine Learning," Morgan Publishers San Mated, California, 302p., 1993.
[5] Shannon, C. E., "A Mathematical Theory of Communication," Bell System Technical Journal 27, pp. 379-423, 1948.
[6] Sornlertlamvanich, V., Tarsaku, P., Srichaivattana, P., Charoenporn, T., and Isahara, H., "Dictionary-less Search Engine for the Collaborative Database," Proceedings of the Third International Symposium on Communications and Information Technologies (ISCIT-2003), Songkhla, Thailand, 2003.
[7] Sornlertlamvanich, V. and Tanaka, H., "The Automatic Extraction of Open Compounds from Text Corpora," Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pp. 1143-1146, 1996.
[8] Sornlertlamvanich, V., Potipiti, T. and Charoenporn, T., "Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm," Proceedings of the 18th International Conference on Computational Linguistics (COLING2000), Saarbrucken, Germany, July-August 2000, pp. 802-807, 2000.
[9] Watkins, C., "Dynamic alignment kernels," Technical Report CSD-TR-98-11, Royal Holloway, University of London, 1999.
[10] Yamamoto, M. and Church, K. W., "Using Suffix Arrays to Compare Term Frequency and Document Frequency for All Substrings in Corpus," Proceedings of the Sixth Workshop on Very Large Corpora, pp. 27-37, 1998.
[11] Cavnar, W. and Trenkle, J., "N-gram-based text categorization," In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, pp. 161-169, 1994.
[12] Vapnik, V., "The Nature of Statistical Learning Theory," Springer-Verlag, Berlin, 1995.
[13] Kruengkrai, C., Srichaivattana, P., Sornlertlamvanich, V., and Isahara, H., "Language Identification Based on String Kernel," In Proceedings of the Fifth Symposium on Communications and Information Technologies (ISCIT-2005), Beijing, China, 2005.
[14] Herbrich, R., "Learning Kernel Classifiers: Theory and Algorithms," The MIT Press, 2002.