A Collaborative Framework for Collecting Thai Unknown Words from the Web

Choochart Haruechaiyasak, Chatchawal Sangkeettrakarn, Pornpimon Palingoon
Sarawoot Kongyoung and **Chaianun Damrongrat**
Information Research and Development Division (RDI)
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand
rdi5@nnet.nectec.or.th

Abstract

We propose a collaborative framework for collecting Thai unknown words found on Web pages over the Internet. Our main goal is to design and construct a Web-based system which allows a group of interested users to participate in constructing a Thai unknown-word open dictionary. The proposed framework provides supporting algorithms and tools for automatically identifying and extracting unknown words from Web pages of given URLs. The system yields the result of unknown-word candidates which are presented to the users for verification. The approved unknown words could be combined with the set of existing words in the lexicon to improve the performance of many NLP tasks such as word segmentation, information retrieval and machine translation. Our framework includes word segmentation and morphological analysis modules for handling the non-segmenting characteristic of Thai written language. To take advantage of large available text resource on the Web, our unknown-word boundary identification approach is based on the statistical string pattern-matching algorithm.

Keywords: Unknown words, open dictionary, word segmentation, morphological analysis, word-boundary detection.

1 Introduction

The advent of the Internet and the increasing popularity of the Web have altered many aspects of natural language usage. As more people turn to the Internet as a new communicating channel, the textual information has increased tremendously and is also widely accessible. More importantly, the available information is varied largely in terms of topic difference and multi-language characteristic. It is not uncommon to find a Web page written in Thai lies adjacent to a Web page written in English via a hyperlink, or a Web page containing both Thai and English languages. In order to perform well in this versatile environment, an NLP system must be adaptive enough to handle the variation in language usage. One of the problems which requires special attention is unknown words.

As with most other languages, unknown words also play an extremely important role in Thai-language NLP. Unknown words are viewed as one of the problematic sources of degrading the performance of traditional NLP applications such as MT (Machine Translation), IR (Information Retrieval) and TTS (Text-To-Speech). Reduction in the amount of unknown words or being able to correctly identify unknown words in these systems would help increase the overall system performance.

The problem of unknown words in Thai language is perhaps more severe than in English or other latin-based languages. As a result of the information technology revolution, Thai people have become more familiar with other foreign languages especially English. It is not uncommon to hear a few English words over a course of conversation between two Thai people. The foreign words along with other Thai named entities are among the new words which are continuously created and widely circulated. To write a foreign word, the transliterated form of Thai alphabets is often used. The Royal Institute of Thailand is the official organization in Thailand who has respon-

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 345–352,
Sydney, July 2006.

sibility and authority in defining and approving the use of new words. The process of defining a new word is manual and time-consuming as each word must be approved by a working group of linguists. Therefore, this traditional approach of constructing the lexicon is not a suitable solution, especially for systems running on the Web environment.

Due to the inefficiency of using linguists in defining new lexicon, there must be a way to automatically or at least semi-automatically collect new unknown words. In this paper, we propose a collaborative framework for collecting unknown words from Web pages over the Internet. Our main purpose is to design and construct a system which automatically identifies and extracts unknown words found on Web pages of given URLs. The compiled list of unknown-word candidates is to be verified by a group of participants. The approved unknown words are then added to the existing lexicon along with the other related information such as meaning and POS (part of speech).

This paper focuses on the underlying algorithms for supporting the process of identifying and extracting unknown words. The overall process is composed of two steps: unknown-word detection and unknown-word boundary identification. The first step is to detect the locations of unknown-word occurrences from a given text. Since Thai language belongs to the class of non-segmenting language group in which words are written continuously without using any explicit delimiting character, detection of unknown words could be accomplished mainly by using a word-segmentation algorithm with a morphological analysis. By using a dictionary-based word-segmentation algorithm, locations of words which are not previously included in the dictionary will be easily detected. These unknown words belong to the class of *explicit* unknown words and often represent the transliteration of foreign words.

The other class of unknown words is *hidden* unknown words. This class includes new words which are created through the combination of some existing words in the lexicon. The *hidden* unknown words are usually named entities such as a person's name and an organization's name. The *hidden* unknown words could be identified using the approaches such as n-gram generation and phrase chunking. The scope of this paper focuses only on the extraction of the *explicit* unknown words. However, the design of our framework also includes the extraction of *hidden* unknown words. We will continue to explore this issue in our future works.

Once the location of an unknown word is detected, the second step involves the identification of its boundary. Since we use the Web as our main resource, we could take advantage of its large availability of textual contents. We are interested in collecting unknown words which occur more than once throughout the corpus. Unknown words which occur only once in the large corpus are not considered as being significant. These words may be unusual words which are not widely accepted, or could be misspelling words. Using this assumption, our approach for identifying the unknown-word boundary is based on a statistical pattern-matching algorithm. The basic idea is that the same unknown word which occurs more than once would likely to appear in different surrounding contexts. Therefore, a group of characters which form the unknown word could be extracted by analyzing the string matching patterns.

To evaluate the effectiveness of our proposed framework, experiments using a real data set collected from the Web are performed. The experiments are designed to test each of the two main steps of the framework. Variation of morphological analysis are tested for the unknown-word detection. The detection rate of unknown words were found to be as high as approximately *96%*. Three variations of string pattern-matching techniques were tested for unknown-word boundary identification. The identification accuracy was found to be as high as approximately *36%*. The relatively low accuracy is not the major concern since the unknown-word candidates are to be verified and corrected by users before they are actually added to the dictionary. The system is implemented via the Web-browser environment which provides user-friendly interface for verification process.

The rest of this paper is organized as follows. The next section presents and discusses related works previously done in the unknown-word problem. Section 3 provides an overview of unknown-word problem in the relation to the word-segmentation process. Section 4 presents the proposed framework with underlying algorithms in details. Experiments are performed in Section 5 with results and discussion. The conclusion is given in Section 6.

2 Previous Works

The research and study in unknown-word problem have been extensively done over the past decades. Unknown words are viewed as problematic source in the NLP systems. Techniques in identifying and extracting unknown words are somewhat language-dependent. However, these techniques could be classified into two major categories, one for segmenting languages and another for non-segmenting languages. Segmenting languages, such as latin-based languages, use delimiting characters to separate written words. Therefore, once the unknown words are detected, their boundaries could be identified relatively easily when compared to those for non-segmenting languages.

Some examples of techniques involving segmenting languages are listed as follows. Toole (2000) used multiple decision trees to identify names and misspellings in English texts. Features used in constructing the decision trees are, for example, POS (Part-Of-Speech), word length, edit distance and character sequence frequency. Similarly, a decision-tree approach was used to solve the POS disambiguation and unknown word guessing in (Orphanos and Christodoulakis, 1999). The research in the unknown-word problem for segmenting languages is also closely related to the extraction of named entities. The difference of these techniques to those in non-segmenting languages is that the approach needs to parse the written text in word-level as opposed to character-level.

The research in unknown-word problem for non-segmenting languages is highly active for Chinese and Japanese. Many approaches have been proposed and experimented with. Asahara and Matsumoto (2004) proposed a technique of SVM-based chunking to identify unknown words from Japanese texts. Their approach used a statistical morphological analyzer to segment texts into segments. The SVM was trained by using POS tags to identify the unknown-word boundary. Chen and Ma (2002) proposed a practical unknown word extraction system by considering both morphological and statistical rule sets for word segmentation. Chang and Su (1997) proposed an unsupervised iterative method for extracting unknown lexicons from Chinese text corpus. Their idea is to include the potential unknown words to the augmented dictionary in order to improve the word segmentation process. Their proposed approach also includes both contextual constraints and the joint character association metric to filter the unlikely unknown words. Other approaches to identify unknown words include statistical or corpus-based (Chen and Bai, 1998), and the use of heuristic knowledge (Nie et al. , 1995) and contextual information (Khoo and Loh, 2002). Some extensions to unknown-word identification have been done. An example include the determination of POS for unknown words (Nakagawa et al. , 2001).

The research in unknown words for Thai language has not been widely done as in other languages. Kawtrakul et al. (1997) used the combination of a statistical model and a set of context sensitive rules to detect unknown words. Our framework has a different goal from previous works. We consider unknown-word problem as collaborative task among a group of interested users. As more textual content is provided to the system, new unknown words could be extracted with more accuracy. Thus, our framework can be viewed as collaborative and statistical or corpus-based.

3 Unknown-Word Problem in Word Segmentation Algorithms

Similar to Chinese, Japanese and Korea, Thai language belongs to the class of non-segmenting languages in which words are written continuously without using any explicit delimiting character. To handle non-segmenting languages, the first required step is to perform word segmentation. Most word segmentation algorithms use a lexicon or dictionary to parse texts at the character-level. A typical word segmentation algorithm yields three types of results: known words, ambiguous segments, and unknown segments. Known words are existing words in the lexicon. Ambiguous segments are caused by the overlapping of two known words. Unknown segments are the combination of characters which are not defined in the lexicon.

In this paper, we are interested in extracting the unknown words with high precision and recall results. Three types of unknown words are *hidden*, *explicit* and *mixed* (Kawtrakul et al. , 1997). Hidden unknown words are composed by different words existing in the lexicon. To illustrate the idea, let us consider an unknown word *ABCD* where *A*, *B*, *C*, and *D* represents individual characters. Suppose that *AB* and *CD* both ex-

ist in a dictionary, then *ABCD* is considered as a hidden unknown word. The explicit unknown words are newly created words by using different characters. Let us again consider an unknown word ABCD. Suppose that there is no substring of *ABCD* (i.e., *AB*, *BC*, *CD*, *ABC*, *BCD*) exists in the dictionary, then *ABCD* is considered as explicit unknown words. The mixed unknown words are composed of both existing words in a dictionary and non-existing substrings. From the example of unknown string *ABCD*, if there is at least one substring of *ABCD* (i.e., *AB*, *BC*, *CD*, *ABC*, *BCD*) exists in the dictionary, then *ABCD* is considered as a mixed unknown word.

It can be immediately seen that the detection of the *hidden* unknown words are not trivial since the parser would mistakenly assume that all the fragments of the words are valid, i.e., previously defined in the dictionary. In this paper, we limit ourself to the extraction of the *explicit* and *mixed* unknown words. This type of unknown words usually represent the transliteration of foreign words. Detection of these unknown words could be accomplished mainly by using a word-segmentation algorithm with a morphological analysis. By using a dictionary-based word-segmentation algorithm, locations of words which are not previously defined in the lexicon could be easily detected.

4 The Proposed Framework

The overall framework is shown in Figure 1. Two major components are information agent and unknown-word analyzer. The details of each component are given as follows.

- **Information agent**: This module is composed of a Web crawler and an HTML parser. It is responsible for collecting HTML sources from the given URLs and extracting the textual data from the pages. Our framework is designed to support multi-user and collaborative environment. The advantage of this design approach is that unknown words could be collected and verified more efficiently. More importantly, it allows users to select the Web pages which suit their interests.

- **Unknown-word analyzer**: This module is composed of many components for analyzing and extracting unknown words. Word segmentation module receives text strings from the information agent and segments them into a list of words. N-gram generation module is responsible for generating hidden unknown-word candidates. Morphological analysis module is used to form initial explicit unknown-word segments. String pattern matching unit performs unknown-word boundary identification task. It takes the intermediate unknown segments and identifies their boundaries by analyzing string matching patterns The results are processed unknown-word candidates which are presented to linguists for final post-processing and verification. New unknown words are combined with the dictionary to iteratively improve the performance of the word segmentation module. Details of each component are given in the following subsections.

4.1 Unknown-Word Detection

As previously mentioned in Section 3, applying a word-segmentation algorithm on a text string yields three different segmented outputs: known, ambiguous, and unknown segments. Since our goal is to simply detect the unknown segments without solving or analyzing other related issues in word segmentation, using the *longest-matching* word segmentation algorithm previously proposed by Poowarawan (1986) is sufficient. An example to illustrate the word-segmentation process is given as follows.

Let the following string denotes a text string written in Thai language: $\{a_1a_2...a_ib_1b_2...b_jc_1c_2...c_k\}$. Suppose that $\{a_1a_2...a_i\}$ and $\{c_1c_2...c_k\}$ are known words from the dictionary, and $\{b_1b_2...b_j\}$ be an unknown word. For the explicit unknown-word case, applying the word-segmentation algorithm would yield the following segments: $\{a_1a_2...a_i\}\{b_1\}\{b_2\}...\{b_j\}\{c_1c_2...c_k\}$. It can be observed that the detected unknown positions for a single unknown word are individual characters in the unknown word itself. Based on the initial statistical analysis of a Thai lexicon, it was found that the averaged number of characters in a word is equal to *7*. This characteristic is quite different from other non-segmenting languages such as Chinese and Japanese in which a word could be a character or a combination of only a few characters. Therefore, to reduce the complexity in unknown-word boundary identification task, the unknown segments could be merged to form multiple-character segments. For exam-

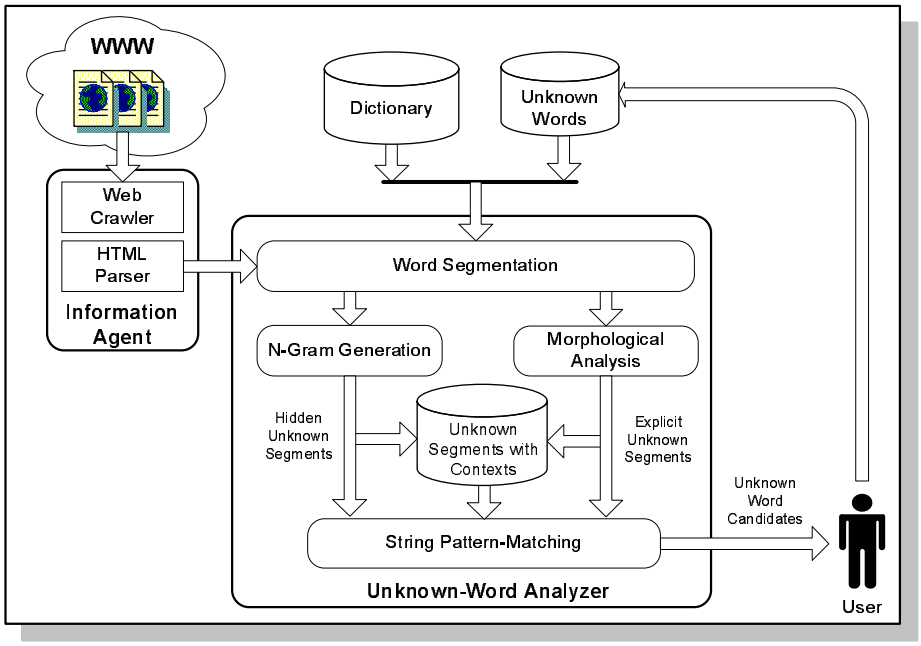

Figure 1: The proposed framework for collecting Thai unknown words.

ple, a merging of two characters per segment would give the following unknown segments: $\{b_1b_2\}\{b_3b_4\}...\{b_{j-1}b_j\}$. In the following experiment section, the merging of two to five characters per segment including the merging of all unknown segments without limitation will be compared.

Morphological analysis is applied to guarantee grammatically correct word boundaries. Simple morphological rules are used in the framework. The rule set is based on two types of characters, front-dependent characters and rear-dependent characters. Front-dependent characters are characters which must be merged to the segment leading them. Rear-dependent characters are characters which must be merged to the segment following them. In Thai written language, these dependent characters are some vowels and tonal characters which have specific grammatical constraints. Applying morphological analysis will help making the unknown segments more reliable.

4.2 Unknown-Word Boundary Identification

Once the unknown segments are detected, they are stored into a hashtable along with their contextual information. Our unknown-word boundary identification approach is based on a string pattern-matching algorithm previously proposed by Boyer and Moore (1977). Consider the unknown-word boundary identification as a string pattern-matching problem, there are two possible strategies: considering the longest matching pattern and considering the most frequent matching pattern as the unknown-word candidates. Both strategies could be explained more formally as follows.

Given a set of N text strings, $\{S_1S_2...S_N\}$, where S_i, is a series of len_i characters denoted by $\{c_{i,1}c_{i,2}...c_{i,len_i}\}$ and each is marked with an unknown-segment position, pos_i, where $1 \leq pos_i \leq len_i$. Given a new string, S_j, with an unknown-segment position, pos_j, the longest pattern-matching strategy iterates through each string, S_1 to S_N and records the *longest* string pattern which occur in both S_j and the other string in the set. On the other hand, the *most frequent* pattern-matching strategy iterates through each string, S_1 to S_N, but records the matching pattern which occur most frequently.

The results from the unknown-word boundary identification are unknown-word candidates. These candidates are presented to the users for verification. Our framework is implemented via a Web-browser interface which provides a user-friendly environment. Figure 2 shows a screen snapshot of our system. Each unknown word is listed within a *text field* box which allows a user to edit and correct its boundary. The contexts could be used as some editing guidelines and are also stored into the database.

Unknown Word	Context
อีเมล์ Add Discard	• 548 : ยังกบดานเงียบ แจ็กพอต 47 ล้าน **อีเมล์**ผู้ส่ง อีเมล์เพื่อน ข้อความถึงเพื่อน • เงียบ แจ็กพอต 47 ล้าน อีเมล์ผู้ส่ง **อีเมล์**เพื่อน ข้อความถึงเพื่อน ข่าวหน้าใน ข่
แจ็กพอต Add Discard	• เปิดเผยอีกว่า นับตั้งแต่มีการออกรางวัล**แจ็กพอต**มาฝ่ายจ่ายรางวัลยังไม่เคยพบว่ามีผู้ถู • บเงินล่าข้าขนาดนี้ที่ผ่านมาเคยมีผู้ถูก**แจ็กพอต**รายหนึ่ง มารับเงินล่าช้า แต่หลังออกรา • อยู่ 47 ล้านบาทนั้น จะนำมารวมกับรางวัล**แจ็กพอต**งวดวันที่1 ต.ค. นี้หรือไม่ ผอ.สำนักงา • ลากกล่าวว่า กรณีที่จะยกยอดเงินมารวมกับ**แจ็กพอต**งวดใหม่ก็ต่อเมืองวดดังกล่าวไม่มีผู้ถ
กสท Add Discard	• ล ที่สำนักสลากรูปแบบใหม่ ชั้น 22 อาคาร **กสท** โทรคมนาคม บางรักเจ้าหน้าที่จ่ายเงินรางว

Figure 2: Example of Web-Based Interface

5 Experiments and Results

In this section, we evaluate the performance of our proposed framework. The corpus used in the experiments is composed of *8,137* newspaper articles collected from a top-selling Thai newspaper's Web site (Thairath, 2003) during 2003. The corpus contains a total of *78,529* unknown words of which *14,943* are unique. This corpus was focused on unknown words which are transliterated from foreign languages, e.g., English, Spanish, Japanese and Chinese. We use the publicly available Thai dictionary *LEXiTRON*, which contains approximately *30,000* words, in our framework (Lexitron, 2006).

We first analyze the unknown-word set to observe its characteristics. Figure 3 shows the plot of unknown-word frequency distribution. Not surprisingly, the frequency of unknown-word usage follows a Zipf-like distribution. This means there are a group of unknown words which are used very often, while some unknown words are used only a few times over a time period. Based on the frequency statistics of unknown words, only about *3%* (*2,375* words out of *78,529*) occur only once in the corpus. Therefore, this finding supports the use of statistical pattern-matching algorithm described in previous section.

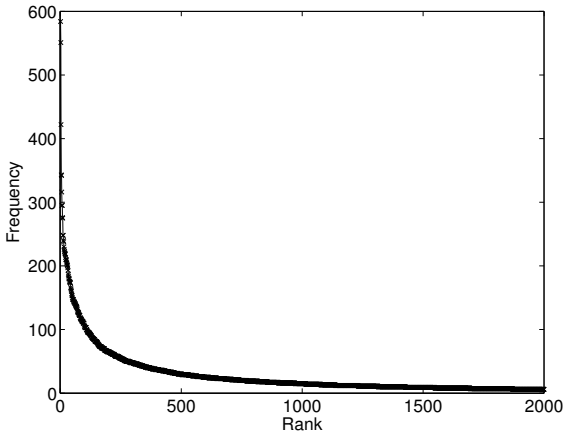

Figure 3: Unknown-word frequency distribution.

5.1 Evaluation of Unknown-Word Detection Approaches

As discussed in Section 4, multiple unknown segments could be merged to form a representative unknown segment. The merging will help reduce the complexity in the unknown-word boundary identification as fewer segments will be checked for the same set of unknown words.

The following variations of merging approach are compared.

- No merging (*none*): No merging process is applied.
- N-character Merging (*N-char*): Allow the maximum of *N* characters per segment.
- Merging all segments (*all*): No limit on number of characters per segment.

We measure the performance of unknown-word detection task by using two metrics. The first is the detection rate (or recall) which is equal to the number of detected unknown words divided by the total number of previously tagged unknown words in the corpus. The second is the averaged detected positions per word. The second metric directly represents the overhead or the complexity to the unknown-word boundary identification process. This is because all detected positions from a single unknown word must be checked by the process. The comparison results are shown in Figure 4. As expected, the approach *none* gives the maximum detection rate of *96.6%*, while the approach *all* yields the lowest detection rate. Another interesting observation is that the approach *2-char* yields comparable detection rate to the ap-

	Unknown-Segment Merging Approach					
	none	2-char	3-char	4-char	5-char	all
Detection Rate (%)	96.6	95.3	93.9	92.9	91.6	85.4
Averaged Detected Positions Per Word	5.9	1.93	1.63	1.46	1.32	0.9

Figure 4: Unknown-word detection results

proach *none*, however, its averaged detected positions per word is about three times lower. Therefore to reduce the complexity during the unknown-word boundary identification process, one might want to consider using the merging approach of *2-char*.

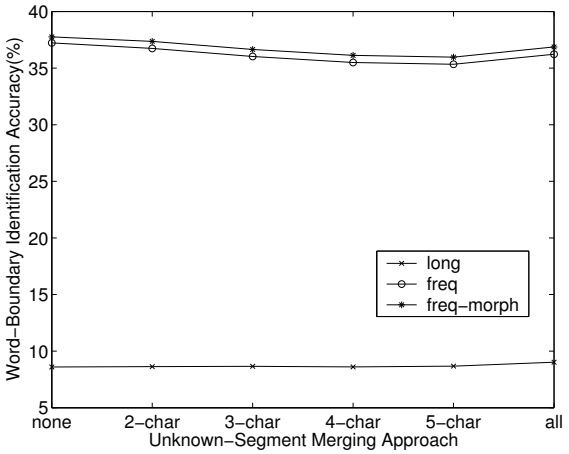

Figure 5: Comparison between different unknown-word boundary detection approaches.

5.2 Evaluation of Unknown-Word Boundary Identification

The unknown-word boundary identification is based on string pattern-matching algorithm. The following variations of string pattern-matching technique are compared.

- Longest matching pattern (*long*): Select the longest-matching unknown-word candidate
- Most-frequent matching pattern (*freq*): Select the most-frequent-matching unknown-word candidate
- Most-frequent matching pattern with morphological analysis (*freq-morph*): Similar the the approach *freq* but with additional morphological analysis to guarantee that the word boundaries are grammatically correct.

The comparison among all variations of string pattern-matching approaches are performed across all unknown-segment merging approach. The results are shown in Figure 5. The performance metric is the word-boundary identification accuracy which is equal to the number of unknown words correctly extracted divided by the total number of tested unknown segments. It can be observed that the selection of different merging approaches does not really effect the accuracy of the unknown-word boundary identification process. But since the approach *none* generates approximately *6* positions per unknown segment on average, it would be more efficient to perform a merging approach which could reduce the number of positions down by at least 3 times.

The plot also shows the comparison among three approaches of string pattern-matching. Figure 6 summarizes the accuracy results of each string pattern-matching approach by taking the average on all different merging approaches. The approach *long* performed poorly with the averaged accuracy of *8.68%*. This is not surprising because selection of the longest matching pattern does not mean that its boundary will be identified correctly. The approaches *freq* and *freq-morph* yield similar accuracy of about *36%*. The *freq-morph* improves the performance of the approach *freq* by less than *1%*. The little improvement is due to the fact that the matching strings are mostly grammatically correct. However, the error is caused by the matching collocations of the unknown-word context. If an unknown word occurs together adjacent to another word very frequently, they will likely be extracted by the algorithm. Our solution to this problem is by providing the users with a user-friendly interface so unknown-word candidates could be easily filtered and corrected.

6 Conclusion

We proposed a framework for collecting Thai unknown words from the Web. Our framework

	Unknown-Word Boundary Identification Approach		
	long	freq	freq-morph
Averaged Accuracy (%)	8.68	36.18	36.79

Figure 6: Unknown-word boundary identification results

is composed of an information agent and an unknown-word analyzer. The task of the information agent is to collect and extract textual data from Web pages of given URLs. The unknown-word analyzer involves two processes: unknown-word detection and unknown-word boundary identification. Due to the non-segmenting characteristic of Thai written language, the unknown-word detection is based on a word-segmentation algorithm with a morphological analysis. To take advantage of large available text resource from the Web, the unknown-word boundary identification is based on the statistical pattern-matching algorithm.

We evaluate our proposed framework on a collection of Web Pages obtained from a Thai newspaper's Web site. The evaluation is divided to test each of the two processes underlying the framework. For the unknown-word detection, the detection rate is found to be as high as *96%*. In addition, by merging a few characters into a segment, the number of required unknown-word extraction is reduced by at least 3 times, while the detection rate is relatively maintained. For the unknown-word boundary identification, considering the highest frequent occurrence of string pattern is found to be the most effective approach. The identification accuracy was found to be as high as approximately *36%*. The relatively low accuracy is not the major concern since the unknown-word candidates are to be verified and corrected by users before they are actually added to the dictionary.

References

Masayuki Asahara and Yuji Matsumoto. 2004. Japanese unknown word identification by character-based chunking. *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, 459–465.

R. Boyer and S. Moore. 1977. A fast string searching algorithm. *Communications of the ACM*, 20:762–772.

Jing-Shin Chang and Keh-Yih Su. 1997. An Unsupervised Iterative Method for Chinese New Lexicon Extraction. *International Journal of Computational Linguistics & Chinese Language Processing*, 2(2).

Keh-Jianne Chen and Ming-Hong Bai. 1998. Unknown Word Detection for Chinese by a Corpus-based Learning Method. *Computational Linguistics and Chinese Language Processing*, 3(1):27–44.

Keh-Jianne Chen and Wei-Yun Ma. 2002. Unknown Word Extraction for Chinese Documents. *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, 169–175.

Asanee Kawtrakul, Chalatip Thumkanon, Yuen Poovorawan, Patcharee Varasrai, and Mukda Suktarachan. 1997. Automatic Thai Unknown Word Recognition. *Proceedings of the Natural Language Processing Pacific Rim Symposium*, 341–348.

Christopher S.G. Khoo and Teck Ee Loh. 2002. Using statistical and contextual information to identify two-and three-character words in Chinese text. *Journal of the American Society for Information Science and Technology*, 53(5):365–377.

Lexitron Version 2.1, Thai-English Dictionary. Source available: http://lexitron.nectec.or.th, February 2006.

Tetsuji Nakagawa, Taku Kudoh and Yuji Matsumoto. 2001. Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines. *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, 325–331.

Jian-Yun Nie, Marie-Louise Hannan and Wanying Jin. 1995. Unknown Word Detection and Segmentation of Chinese Using Statistical and Heuristic Knowledge. *Communications of COLIPS*, 5(1&2):47–57.

Giorgos S. Orphanos and Dimitris N. Christodoulakis. 1999. POS Disambiguation and Unknown Word Guessing with Decision Trees. *Proceedings of the EACL*, 134–141.

Yuen Poowarawan. 1986. Dictionary-based Thai Syllable Separation. *Proceedings of the Ninth Electronics Engineering Conference*.

Thairath Newspaper. Source available: http://www.thairath.com.

Janine Toole. 2000. Categorizing Unknown Words: Using Decision Trees to Identify Names and Misspellings. *Proceeding of the 6th Applied Natural Language Processing Conference (ANLP 2000)*, 173–179.