JIS

Word segmentation for the Myanmar language

Tun Thura Thet and Jin-Cheon Na

Division of Information Studies, School of Communication and Information, Nanyang Technological University, Singapore 637718

Wunna Ko Ko

Myanmar NLP Research Center, Hlaing, Yangon, Myanmar

Abstract.

This study reports the development of a Myanmar word segmentation method using Unicode standard encoding. Word segmentation is an essential step prior to natural language processing in the Myanmar language, because a Myanmar text is a string of characters without explicit word boundary delimiters. The proposed method has two phases: syllable segmentation and syllable merging. A rule-based heuristic approach was adopted for syllable segmentation, and a dictionary-based statistical approach for syllable merging. Evaluation of test results showed that the method is very effective for the Myanmar language.

Keywords: Myanmar language; word segmentation; natural language processing; syllable segmentation; syllable merging; collocation strength; mutual information

1. Introduction

Word segmentation is the process of determining word boundaries in a piece of text. In English language texts, word boundaries are easily determined because of the presence of white spaces or punctuation between words. However, it is not straightforward in languages like Myanmar, which do not have inter-word spacing or any other delimiter in their written texts.

In linguistics, a word is a basic unit of language that carries meaning and can be spoken or written [1]. It can consist of one or more morphemes that are linked more or less tightly together. Typically, a word will consist of a root or stem and zero or more affixes. Words can be combined to form phrases, clauses and sentences. A word consisting of two or more stems joined together is known as a compound word.

To process text computationally, words have to be determined first. For instance, search engines require documents to be indexed by words [2]. When a query is submitted to a search engine, key words of the query are compared against the indexed words of the documents to return search

Correspondence to: Jin-Cheon Na, Division of Information Studies, School of Communication and Information, Nanyang Technological University, Singapore 637718. Email: tjcna@ntu.edu.sg

results. Word segmentation is therefore an essential pre-requirement in applications where data and information are to be computationally processed in their natural language. The manipulation of text for automatic indexing has been recognized as an important area of research in Myanmar Natural Language Processing (NLP) [3].

The purpose of this study is to develop an effective and practical word segmentation algorithm for the Myanmar language, which is used by more than 50 million people. No published work or research has been found for Myanmar word segmentation using the Unicode standard. Without a word segmentation solution, no NLP application (such as Part-of-Speech (POS) tagging and translation) can be developed. Currently, Myanmar NLP is at an initial stage, but word segmentation is very important for future progress. The Myanmar NLP committee [3] has been working with the International Organization for Standardization (ISO) on the standardization of a workable character set for the Myanmar script, and, very recently, the Myanmar character set was successfully proposed and approved for Unicode standard version 5.1. The Myanmar character set had been quite unstable in the past and its recent enhancement in Unicode version 5.1 in 2006 has resolved many issues. As far as word segmentation is concerned, the changes from the previous versions of the Unicode standard are significant. The Myanmar NLP committee is also working on a Myanmar lexicon database.

The Myanmar language is the official language of Myanmar and is more than one thousand years old. Texts in the Myanmar language use the Myanmar script, which is descended from the Brahmi script of ancient South India [4]. Other Southeast Asian descendants of this script, known as Brahmic or Indic scripts, include Thai, Khmer and Lao.

Myanmar writing does not use white spaces between words or between syllables. Thus, the computer has to determine syllable and word boundaries by means of an algorithm. Moreover, a Myanmar syllable can be composed of multiple characters. Vietnamese syllables are also multiple characters, but with inter-syllable white spaces; hence, only word boundaries need to be determined. The problems faced by Myanmar NLP are similar to those for Thai, Khmer and Lao, but with different characteristics. Studies conducted on two other Southeast Asian languages were reviewed in this paper, to help us develop an effective and workable solution for the Myanmar language.

This paper is organized into six sections. Section 2 examines the word segmentation methods used for two Southeast Asian languages. Section 3 describes our proposed Myanmar word segmentation method. Sections 4 and 5 report on the evaluation and implementation of the proposed method. The last section summarizes the proposed development of Myanmar word segmentation and discusses issues for future work.

2. Review of word segmentation methods for the Thai and Vietnamese languages

The following subsections will look at the word segmentation methods used for the Thai and Vietnamese languages. The former is a Brahmic script, but the latter is not.

2.1. Thai language

2.1.1. Introduction to the Thai language

The Thai script is a member of the Brahmic or Indic family of scripts, descended from Brahmi, although it has modified the original Brahmic letter shapes. It has also increased the number of letters in order to accommodate features of the Thai language, and does not use the conjunct consonant mechanism and independent vowel letters found in most other Brahmic scripts [4].

Each Thai letter is a consonant possessing an inherent vowel sound as well as inherent tones. Both the inherent vowel and inherent tone can be modified by means of vowel signs and tone marks attached to the base consonant letter. All of the tone marks and some of the vowel signs are rendered in the script as diacritics attached above or below the base consonant. In the Unicode memory representation, these combining signs and marks are encoded after the modified consonant [4].

In the Thai language, a 'word' is difficult to define, as it does not exhibit explicit word boundaries. Traditional methods for determining word boundaries depend on human judgement and have several limitations. One main cause of the problems with Thai word segmentation is this lack of a clear definition of a Thai 'word' [5]. If there is no agreement on word segmentation, a training corpus will not be useful, because results will be different for different researchers. Therefore, the first thing is to decide on the definition of a Thai 'word'. In linguistics, a word is defined as a linguistic unit made up of one or more morphemes. Thai grammar books, however, usually view a word as a composition of syllables, and distinguish it as either a simple or compound word. Simple words are those that can have one or more syllables, but in the case of a multi-syllable word, the meaning of the word is not related to the meaning of any syllable. Compound words, on the other hand, are composed of two or more simple words. The meaning of a compound word may not be the sum of the meanings of its parts, though it can be related to the meanings of its parts. For example, the word 'river' in Thai is composed the two Thai words 'water' and 'mother'. Though its meaning is not 'mother of water', it is related. Another example is 'rice cooker', which is a compound of three words: 'pot', 'cook' and 'rice' [5].

2.1.2. Research on Thai word segmentation

Like many other Asian languages, the Thai language does not use white spaces for word boundaries. Most word segmentation approaches use a dictionary for segmenting running texts. If the coverage of the dictionary is not good enough, this will lead to a great number of unknown or unrecognized words, adversely affecting the results [6]. Traditional methods of Thai word segmentation are based on unclear criteria and procedures, and have several limitations. Most methods use "longest matching", where the outcome is very much dependent on the coverage of the dictionary used [7]. Since 1981, various methods have been proposed, implemented and tested, but no satisfying solution has been found.

A study conducted by Sornlertlamvanich et al. [8] used automatic corpus-based word extraction. It employed the C4.5 decision tree induction program [9] as a learning algorithm for word extraction. The induction algorithm evaluates the content of a series of attributes and interactively builds a tree. The leaves of the decision tree represent the values of the goal attributes. The method used C4.5 to prune the entire decision tree, in order to reduce the effect of over-fitting. It recursively travelled to each sub-tree to determine if the leaf or branch could reduce the expected error rate. The attributes of the learning algorithm are mutual information, entropy, frequency and string length. Evaluation of the method was carried out with a 1 MB corpus, consisting of 75 articles from various fields. A total of 30,000 strings were manually tagged and compared with the results produced by the method, which recorded 84.1% accuracy for the test dataset.

Another study, conducted by Wirot [5], used a two-part approach: a syllable-based trigram model for syllable segmentation, and maximum collocation for syllable merging. Syllable segmentation was done on the basis of trigram statistics, whereas syllable merging was done on the basis of collocation between words. Many word segmentation ambiguities were resolved during the syllable segmentation process. Since a syllable is a more well-defined unit than a word, and also more consistent in analysis, syllable segmentation is more effective and reliable than methods which use only word segmentation. A Thai syllable is composed of three parts: vowel forms, initial consonants and final consonants. In some syllables, however, vowel forms can be omitted, while some syllables use more than one character for vowel forms. Further, some syllables can have more than one initial or final consonant. A total of about 200 syllable patterns are defined. Using a training corpus of 553,372 syllables, a newspaper was manually syllable segmented. Witten-Bell discounting [10] was used for smoothing, and Viterbi algorithms were used for determining the best syllable segmentation. When tested on another corpus of 30,498 syllables, the results were 99.8% correct, with only 52 segmentation errors. Of these 52 errors, 22 were proper names and foreign words written in Thai. After syllable segmentation, the strategy was to use collocation strength between syllables to merge syllables. Collocation refers to co-occurrence of syllables observed from the training corpus. If a word contains two or more syllables, those syllables will always co-occur. Thus, the probability of co-occurrence will be much greater than by chance [11]. From experiments conducted, the maximum collocation approach

did not obviously out-perform the 'longest matching' approach in terms of precision, recall, or *F*-measure. The 'longest matching' approach relies heavily on words listed in the dictionary, and always prefers compound words over simple words. On the other hand, the maximum collocation approach does not exhibit such a preference. Therefore, further study was deemed necessary to find out which method was more appropriate for syllable merging.

2.2. Vietnamese language

2.2.1. Introduction to the Vietnamese language

The Vietnamese language has a special linguistic unit called a 'tieng' (equivalent to hanzi in Chinese). One 'tieng' is one sound unit and has one syllable. Unlike the hanzi of Chinese, one 'tieng' has only one possible pronunciation. One or more sound units can be combined to form a word in Vietnamese. The word boundary may vary from person to person, but this is not a problem for native speakers. In Vietnamese, the combination of sound units is the only way to form new lexical units to describe new concepts. There is no prefix or suffix in Vietnamese, only syllables or sound units. A large part of the vocabulary of Vietnamese comes from ancient Chinese. For example, in Vietnamese, 'cong nhan' means 'worker' and 'thuong nhan' means 'businessman', where 'nhan' means 'person' in Chinese. In this example, the modifier ('cong' or 'thuong') is followed by the head 'nhan'. In pure Vietnamese, the head comes before the modifier [12]. Such issues make Vietnamese a difficult language to work with. For a native speaker, it is not a problem to segment the words and understand the meaning. For a computer, however, it is confusing and ambiguous. It is clear that what constitutes a word is not resolved. Another major problem is that there are few effective lexical resources available. The only option is to use pure statistical methods. One of the most difficult tasks in machine translation to or from Vietnamese is the elimination of ambiguity in human languages [13]. A word in Vietnamese often has different meanings depending on its syntactical position in the sentence and the context.

Vietnamese, traditionally a spoken language, is a monosyllabic language that belongs to the Southeast Asian language family. Even though its alphabet (called quoc ngu) is based on the Latin alphabet, its differences from Indo-European languages make it difficult for the Vietnamese not only to learn European languages but also to develop techniques for natural language processing. The available options are either to fit Vietnamese into a well-established European language framework or to come up with a new framework. The former had been tried but did not achieve good results, while the latter requires substantial human and material resources [12].

Each syllable or sound unit in Vietnamese is composed of a (sequence of) character(s) separated by spaces, and there is also some separation for words. From the beginning to the middle of the twentieth century, Vietnamese scholars proposed a new approach for better word segmentation: to use spaces for word segmentation while eliminating spaces between syllables or sound units. However, no general agreement was reached [12].

2.2.2. Research on Vietnamese word segmentation

Three methods for Vietnamese word segmentation are reviewed here. The first, by Dinh et al. [14], utilized weighted finite state transducer (WFST) and neural network techniques; the second, by Ha [12], adopted a pure statistical model, using the trigram maximum probability approach; and the third, by Nguyen et al. [15], also made use of a statistical model, using the N-gram mutual information (MI) approach.

The study conducted by Dinh et al. [14] considered Vietnamese word segmentation as a stochastic transduction problem, and applied the WFST model. The first step was a preprocessing stage, where errors in sentence presentation were eliminated. The normalization of accenting was also done at this stage. Then, the sentences were introduced to the WFST model, where reduplicatives, proper nouns, dates-times and numbers were further defined. A dictionary was used in this approach, and was arranged as a multiway tree in which each node represented a Vietnamese letter. Attributes such as POS, word frequency and syntactic features were stored for each word. The selection of the best word segmentation was done using the Like-Viterbi algorithm. A Neural Network model was

applied as the last step, when the WFST model proved insufficient to determine word boundaries. Machine learning for ambiguous sentences using a neural network model was used instead of a rule-based model. The corpus used for calculating probability was about 2 million words from five different sources. The size of the dictionary used was about 34,000 entries. When the approach was evaluated by comparing it with a manually annotated corpus, it achieved 97% accuracy on a corpus of Vietnamese electronic textbooks. Out of a total of 700 sentences, 683 sentences were correct. The problems highlighted by this study are the absence of an exhaustive dictionary and the ambiguity of Vietnamese words.

Ha [12] employed a pure statistical model, using the maximum probability of a tri-gram in a given chunk of syllables. An unannotated corpus of 10 million words was used in this study. In Vietnamese, names can contain meanings, and the initial characters are in uppercase. Thus, not everything could simply be converted into lower-case (as with English) before counting the n-grams. The study counted n-grams in both cases. The aim was to maximize the probability of the chunk, using different segmentations. The probability of a chunk is the product of its n-gram probabilities, and the chunk with the maximum probability was selected as the final result. Dynamic programming was used to address the problem of combinatorial explosion, where the number of possible segmentations is an exponential function of the length of the chunk. Evaluation was conducted on 100 randomly selected chunks, containing 614 words identified by the model. A native-speaking evaluator was in 'agreement' with 315 (51%) of the words, and considered 402 (65%) of them 'reasonable'. (See Table 8.) The problems encountered in this study were vague definitions of what constitutes a Vietnamese word and the lack of concrete evaluation schemes for word segmentation approaches.

The study by Nguyen et al. [15] also adopted a statistical model, using Mutual Information (MI) formulae for n-grams. The interesting aspect in this approach is that the statistical information was retrieved directly from a commercial search engine by using a genetic algorithm to find the most reasonable segmentation. A genetic algorithm was applied to evolve a population in which each individual was a particular way of segmenting. If enough statistical information was cached, the processing time for a document was about one minute. A new way to calculate the MI of a Vietnamese n-gram was considered in this study for better performance, but there was no significant difference in word segmentation results among the various MI formulas. Using the principles of evolution and heredity, genetic algorithms have long been known for their ability to traverse very large search spaces effectively, and to find approximate global optimal solutions instead of local optimal solutions [16]. About 80% of the segmented words were considered 'acceptable' by two native speakers. (See Table 8.)

3. Myanmar word segmentation

3.1. Introduction to the Myanmar language

A Myanmar text is a string of characters without explicit word boundary markup, written in sequence from left to right without regular inter-word spacing, although inter-phrase spacing may sometimes be used.

Myanmar characters can be classified into three groups: consonants, medials and vowels. The basic consonants in Myanmar can be multiplied by medials. Syllables or words are formed by consonants combining with vowels. However, some syllables can be formed using just consonants, without any vowel. Other characters in the Myanmar script include special characters, numerals, punctuation marks and signs.

There are 34 basic consonants in the Myanmar script, as shown in Table 1. They are known as 'Byee' in the Myanmar language [17]. Unicode encodes the consonants between U+1000 and U+1021. Note that the consonants 'ဩ' and 'ဥ' are stored as different codes, although they can be considered the same consonant. Consonants serve as the base characters of Myanmar words, and are similar in pronunciation to those of other Southeast Asian scripts, such as Thai, Lao and Khmer.

Medials are known as 'Byee Twe' in Myanmar [17]. There are four basic medials and six combined medials in the Myanmar script. The 10 medials can modify the 34 basic consonants to form 340

Table 1
Myanmar characters

Basic Consonants (Byee)				
က	ခ	ဂ	ဃ	င
စ	ဆ	ဇ	ဈ	ဉ / ည
ဋ	ဌ	ဍ	ဎ	ဏ
တ	ထ	ဒ	ဓ	န
ပ	ဖ	ဗ	ဘ	မ
ယ	ရ	လ	ဝ	သ
ဟ	ဠ	အ		

Basic Medials (Byee Twe)				
ျ				
ြ				
ွ				
ှ				
Combined Medials (Byee Twe)				
ျ	+	ွ		
ြ	+	ွ		
ျ	+	ှ		
ြ	+	ှ		
ျ	+	ှ	+	ွ
ြ	+	ှ	+	ွ

Vowels (Thara)				
ေ-	◌	◌		
◌	◌	◌	◌	

Others				
anusvara	*Atha*	*dot*	*visarga*	*kinzi*
◌	◌	◌	◌	◌

Special Characters				
ဥ	၍	၏	ဤ	ဩ

Numerals				
၀	၁	၂	၃	၄
၅	၆	၇	၈	၉

Punctuation marks				
၊	။			

additional multi-clustered consonants. Therefore, a total of 374 consonants exist in the Myanmar script, although some consonants have the same pronunciation.

Vowels are known as 'Thara' [17]. Vowels are the basic building blocks of syllable formation in the Myanmar language, although a syllable or a word can be formed without a vowel. As in other languages, multiple vowel characters can exist in a single syllable.

Special characters, represented by separate codes, are usually complete words or syllables by themselves. They form independent words with complete meanings and do not require any extra consonant, medial or vowel to become a word. Two noteworthy special characters are Kinzi (U+1039), also known as stacking or 'Htutsint', and Atha (U+103A), commonly called 'Killer'. Unlike Atha, Kinzi is invisible when it is typed, but changes the rendering of characters by stacking one consonant above the following one, example '္က' in the word 'ကမ္ဘာ'.

Myanmar numerals are decimal-based, and Table 1 shows zero to nine in sequence. No thousand separators are used; instead, spaces are sometimes used between digits for easy reading.

The two punctuation marks function in a similar manner to the comma and the period in English, respectively. In addition to punctuation, white spaces also separate phrases which can be part of a sentence. Additional signs may be used, such as 'kyats', which is the currency in Myanmar.

A Myanmar syllable has a base character, and may also have (or not) a pre-base character, a post-base character, an above-base character and a below-base character [18]. Figures 1 and 2 illustrate how a syllable is formed. Regardless of the appearance of the characters on the screen, the characters are to be stored consistently in a sequence specified by the Unicode standard. Further, the order in which the characters are stored may not be the same as their keyboarding sequence.

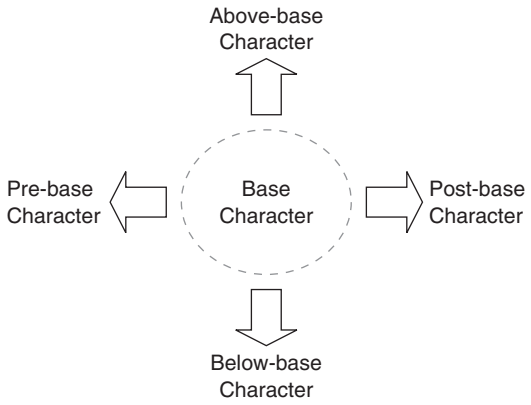

Fig. 1. Positioning of characters in a Myanmar syllable.

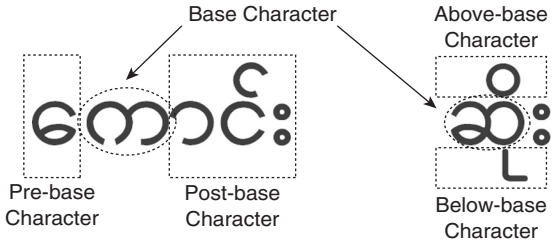

Fig. 2. Two Myanmar syllables.

The proposed word segmentation method in this study is in two phases: syllable segmentation and syllable merging. The method uses a rule-based heuristic approach for syllable segmentation and a dictionary-based statistical approach for syllable merging.

3.2. Syllable segmentation

A syllable is a basic sound unit or a sound. A word can be made up of one or more syllables. Every syllable boundary can be a potential word boundary. In some cases, a word can include other words, in which case it is called a compound word.

In Myanmar, a syllable is formed based on rules that are quite definite and unambiguous. A syllable can contain multiple consonants, multiple medials and multiple vowels. These constituents can appear in different sequences; e.g. consonants followed by medials followed by vowels, or consonant, then medials, then vowels, followed by more consonants and vowels.

The following syllable segmentation rules are proposed in this study:

1. Single character rule (R1)
2. Special ending characters rule (R2)
3. Second consonant rule (R3)
4. Last character rule (R4)
5. Next starter rule (R5)
6. Miscellaneous rules (R6)
   - Non-Myanmar characters
   - Numeric characters
   - Punctuation marks, spaces and similar characters.

3.2.1. Single character rule (R1)

Characters such as 'ဤ', 'ဥ' , '၍' and '၏' are single characters and do not need any medials or vowels to become a syllable or a word. Once these characters are encountered, they can be immediately segmented as a syllable or a word.

3.2.2. Special ending characters rule (R2)

Some characters, such as ':' and '့', represent the end of a syllable. When any of these characters is found, it is safe to assume that it is the end of a syllable. The rules can be refined to handle exceptional cases, for example, the spelling of words from other languages, such as English. For example, if the name 'George' is written as 'ဂျော့ခ်ျ' in Myanmar, it will violate the special ending characters rule.

3.2.3. Second consonant rule (R3)

When a syllable has two consonants, the second consonant should come with either the Atha (Killer) or the Kinzi (Htutsint). This rule will segment a Killer/Kinzi pair of consonants as a syllable.

3.2.4. Last character rule (R4)

The last character in a sentence, a phrase, or the input file can be regarded as the end of a syllable. This is because in Myanmar texts, spaces appear between phrases and punctuation marks between sentences. When the algorithm hits the end of a phrase or a sentence, that is also the end of a syllable or a word. This rule can prevent improper merging of words in the syllable merging stage by preserving white spaces and punctuation marks.

3.2.5. Next starter rule (R5)

This rule works as a complement to the second consonant rule, R3, but without Killer or Kinzi. It provides proper segmentation of an invalid sequence of entries involving the vowel 'ေ'. According to the Unicode standard's encoding rule, the vowel 'ေ' should follow the consonant, with or without medial. The problem arises because in the written form, the vowel 'ေ' precedes the consonant. This gives rise to keyboarding errors and invalid entries. This rule tolerates such invalid entries, and breaks up the syllable when it sees 'ေ' appearing after a complete syllable. This rule is necessary because the test documents may contain such invalid entries.

3.2.6. Miscellaneous rules (R6)

These rules cover numbers, special characters and non-Myanmar characters. Whenever the language changes from Myanmar to English or from English to Myanmar, it signals the syllable breaker. This can be accomplished by checking the characters' values. For example, Myanmar characters have hexadecimal values between (U+1000) and (U+104F), and any character out of this range will be a non-Myanmar character. Similarly, numeric characters can be segmented by checking the range (U+1040 to U+1049). Special characters and punctuation marks can also be segmented by checking their range.

3.3. Syllable merging

The next step is to merge the segmented syllables into words. Our proposed method uses a dictionary-based statistical approach to perform syllable merging.

First, the input text of segmented syllables is broken down into sentences and phrases by looking at punctuation marks and spaces. For each sentence or phrase, all possible combinations of merged words are generated by matching segmented syllables in the sentence or phrase with word entries in the dictionary. From the resulting combinations, the one with the minimum number of merged words is selected, and taken as the correctly merged words of the sentence or phrase. This approach is biased to prefer longer word matching in the dictionary, as the dictionary-based approach with longest matching worked well in our internal tests for syllable merging. When there are two or more combinations with the same minimum number of merged words, the following statistical approach is used to resolve the problem.

In the statistical approach, the mutual information [19] of two syllables (i.e. bi-grams) is pre-calculated with the corpus, and then used to calculate the collocation strength of a sentence or phrase.

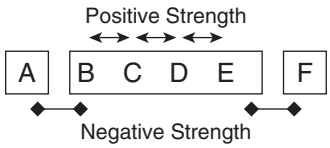

Fig. 3. Collocation strength of the word 'BCDE'.

The collocation strength of a sentence or phrase is the sum of the collocation strengths of all the merged words in the sentence or phrase, whereas the collocation strength of an individual word is the sum of the positive strength minus the sum of the negative strength. The equations for calculating the collocation strength of a sentence or phrase are as follows:

$$CS_s = \sum_{w=1}^{nw} CS_W \tag{1}$$

$$CS_W = \sum_{i=1}^{ns-1} MI\,(i,\ i+1) - (MI\,(Syllable_{LeftNeighbor},\ Syllable_{FirstCurrent}) + MI\,(Syllable_{LastCurrent},\ Syllable_{RightNeighbor})) \tag{2}$$

$$MI(Sx, Sy) = \log_2 \frac{P(Sx, Sy)}{P(Sx)P(Sy)} \tag{3}$$

where CS_s = collocation strength of a sentence or phrase; w = index for words; nw = total number of words in a sentence or phrase; CS_w = collocation strength of an individual word ($CS_w = 0$ if a word is a single syllable); $MI()$ = mutual information between two syllables; $P(S_x,S_y)$ is the probability of observing S_x and S_y together, and $P(S_x)$ and $P(S_y)$ are the probabilities of observing S_x and S_y independently; i = index for current word's syllables; and ns = total number of syllables in a word.

As an example, for the word 'BCDE' in the sentence 'ABCDEF', its positive strength is the sum of the mutual information of B–C, C–D and D–E, whereas its negative strength is the sum of the mutual information of A–B and E–F (see Figure 3.) In most cases, if 'BCDE' is a correctly merged word, its positive collocation strength will be higher than its negative collocation strength.

For each sentence or phrase, we merge the segmented syllables into words with the help of a dictionary, and then select the combination with the minimum number of merged words. When there are two or more such combinations, their collocation strengths are calculated, and the combination with the highest collocation strength is selected.

During the syllable merging process, three kinds of problems can occur. First, missing entries in the dictionary can cause the system to be unable to merge all the words it should. In this study, we made use of a base dictionary, with about 30,000 entries, provided by the Myanmar NLP team, and its coverage has been found to be relatively effective compared to dictionaries in other languages. We also used context-based extensions to the base dictionary. For instance, when processing Myanmar literature documents, we used the base dictionary together with an appropriate extension. The second problem arises when errors (e.g. spelling errors) occur in either the test documents or the dictionary itself. This problem can be reduced by correcting the error entries in the dictionary. The third problem, an infrequent occurrence, can happen when segmented syllables are matched with various words in the dictionary. For instance, the segmented syllables 'ABCD' can be merged in various ways with the following dictionary entries: A, AB, CD, and BCD. For the known cases, this problem can be solved by our proposed approach using the statistical information of the corpus (see example 2 in Section 3.4).

3.4. Two examples of syllable segmentation and syllable merging

Example 1.

Input text:	ဤနေရာတွင်ကျောင်းသားများစာဖတ်နေသည်။
Segmented syllables:	ဤ + နေ + ရာ + တွင် + ကျောင်း + သား + များ + စာ + ဖတ် +နေ +သည် + ။
Merged syllables:	ဤ + နေရာ + တွင် + ကျောင်းသားများ + စာ ဖတ် + နေ +သည် + ။ This + place + at + students + read + doing + is +
In English	Students are reading at this place

The above input text was segmented by applying the rules described in Section 3.2 as follows: the syllable 'ဤ' was segmented by rule R1, the single character rule. Syllables 'နေ', 'ရာ' and 'တွင်' were segmented by rule R3, the second consonant rule. The syllables 'ကျောင်း', 'သား' and 'များ' were segmented by rule R2, the special ending character rule. Syllables 'စာ' and 'ဖတ်' were segmented by rule R3. The syllable 'နေ' can be segmented by rule R3, but it can also be segmented by R5 because of the vowel 'ေ'. Lastly, the syllable 'သည်' was segmented by rule R6, the miscellaneous rule.

Applying our proposed approach for syllable merging, there is only one combination with the minimum number of words, and the output is 'ဤ+နေရာ+တွင်+ကျောင်းသားများ+စာဖတ်+နေ+သည်+။'.

Example 2.

Input text:	သဘာဝဟာသဘာဝပါ။
Segmented syllables:	သ+ဘာ+ဝ+ဟာ+သ+ဘာ+ဝ+ပါ+။
Merged syllables:	သဘာဝ+ဟာ+သဘာဝ+ပါ+။ Nature + is + nature +
In English	Nature is nature

This example is a problematic case and is one of the most frequently quoted examples of problems in Myanmar word segmentation. The syllable segmentation is quite straightforward but the syllable merging is more challenging. The syllables 'သ', 'ဘာ', 'ဝ' and 'ဟာ' were segmented by rule R3. The last syllable, 'ပါ', was segmented by rule R4.

If the longest matching approach were used, the word segmentation output would be 'သဘာဝ+ဟာသ+ဘာဝ+ပါ+။' which is an invalid output. In our proposed approach for syllable merging, 15 combinations of words were generated for this example. Two of them had the same minimum number of words, that is, five, including the punctuation sign. Selecting the combination with the greater collocation strength yielded the correctly merged 'သဘာဝ+ဟာ+သဘာဝ+ပါ+။'

4. Evaluation and error analysis

The proposed method was evaluated using a set of 16 test documents, listed in Table 2. Resources such as lexicons and dictionaries were provided by the Myanmar NLP team, who also assisted in the evaluation. Feedback received was reviewed, and the method was further refined with some improvements.

No errors were detected in syllable segmentation, while we found 25 errors in syllable merging, most of which resulted from missing dictionary entries. Evaluation was carried out by calculating values for precision, recall, and *F*-measure [20].

$$\text{Recall} = \frac{\text{Number of correctly segmented words}}{\text{Number of all words}} \tag{4}$$

$$\text{Precision} = \frac{\text{Number of correctly segmented words}}{\text{Number of segmented words}} \tag{5}$$

$$\textit{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

4.1. Syllable segmentation

Sixteen documents, with a total of 32,567 syllables, were tested. No errors were found, thus achieving perfect precision, recall and *F*-measure. These results demonstrate that syllable segmentation using the heuristic approach can achieve 100% accuracy for the Myanmar language.

Table 2
Test data set

Document name	Characters	Syllables	Words	Base dictionary	Extended dictionary	Type of document
Wtdy1.txt	16490	6327	4555	Dic.txt	Dic_wtdy.txt	Myanmar literature (Waithanthayar Story)
Znk2.txt	5824	2242	1682	Dic.txt	Dic_znk.txt	Myanmar literature
Znk3.txt	6012	2287	1548	Dic.txt	Dic_znk.txt	(Zanaka Story)
Znk4.txt	5341	2057	1485	Dic.txt	Dic_znk.txt	
Znk5.txt	4947	1854	1371	Dic.txt	Dic_znk.txt	
Znk6.txt	6776	2339	1700	Dic.txt	Dic_znk.txt	
Znk7.txt	2416	775	574	Dic.txt	Dic_znk.txt	
Znk8.txt	9481	3396	2438	Dic.txt	Dic_znk.txt	
Znk9.txt	3304	1175	779	Dic.txt	Dic_znk.txt	
Znk10.txt	7956	2702	1968	Dic.txt	Dic_znk.txt	
Znk11.txt	7908	2668	1903	Dic.txt	Dic_znk.txt	
Bible31.txt	3147	1111	818	Dic.txt	Dic_bbl.txt	Myanmar Bible
Bible57.txt	2567	938	686	Dic.txt	Dic_bbl.txt	(Bible)
Bible63.txt	1690	620	454	Dic.txt	Dic_bbl.txt	
Bible64.txt	1852	672	513	Dic.txt	Dic_bbl.txt	
Bible65.txt	3980	1404	1011	Dic.txt	Dic_bbl.txt	
Total	89691	32567	23485			

It appears that rule-based syllable segmentation is the perfect solution for the Myanmar language, and no further investigation of other approaches for syllable segmentation, such as statistical models, is required.

4.2. Syllable merging

Table 3 shows the test results for the 16 test documents. Test results averaged 99.05% recall, 98.94% precision and 98.99% *F*-measure.

4.3. Error analysis

A total of 15 different errors were found. They can be categorized into four groups: missing common words in the dictionary, proper nouns such as names of people and places, adopted words in Pali (or other languages), and numerical words.

4.3.1. Type 1 error – missing common words in the dictionaries

This type of error was caused by missing entries in the dictionaries. The problem can be reduced by extending the coverage of both the base and extended dictionaries. Table 4 lists the Type 1 errors encountered. In the column 'Error words', the '+' sign is used to represent the missed syllable merging. For example, 'a+b' denotes that 'a' and 'b' should have been merged but were not.

4.3.2. Type 2 error – proper nouns such as names of people and places

This type of error occurs because the names of people and places were not listed in the dictionaries used. Names of people and places can be added to a context-based extended dictionary. Alternatively, the Weighted Finite State Transducer (WFST) and neural network approach can be used, as Dinh et al. did [14]. Table 5 lists the Type 2 errors found.

4.3.3. Type 3 error – adopted words in Pali (or other languages)

The Myanmar language has long been influenced by the Pali language due to its Buddhist origin. It is quite common to see Pali words in Myanmar writing. The Pali derived words can be added to an

Table 3
Test results for syllable merging

Document name	Total number of words	Number of segmented words	Number of correctly segmented words	Recall	Precision	*F*-measure
wtdy1.txt	4555	4561	4550	99.89%	99.76%	99.82%
znk2.txt	1682	1682	1633	97.09%	97.09%	97.09%
znk3.txt	1548	1549	1537	99.29%	99.23%	99.26%
znk4.txt	1485	1485	1468	98.86%	98.86%	98.86%
znk5.txt	1371	1373	1361	99.27%	99.13%	99.20%
znk6.txt	1700	1705	1697	99.82%	99.53%	99.68%
znk7.txt	574	574	574	100.00%	100.00%	100.00%
znk8.txt	2438	2440	2406	98.69%	98.61%	98.65%
znk9.txt	779	782	769	98.72%	98.34%	98.53%
znk10.txt	1968	1969	1967	99.95%	99.90%	99.92%
znk11.txt	1903	1907	1889	99.26%	99.06%	99.16%
bible31.txt	818	818	818	100.00%	100.00%	100.00%
bible57.txt	686	686	670	97.67%	97.67%	97.67%
bible63.txt	454	454	414	91.19%	91.19%	91.19%
bible64.txt	513	514	500	97.47%	97.28%	97.37%
bible65.txt	1011	1011	1008	99.70%	99.70%	99.70%
Total/average	23,485	23,510	23,261	**99.05%**	**98.94%**	**98.99%**

Table 4
Type 1 error list

No.	Document	Error words	Meaning in English	Missed merging
1	wtdy1.txt	အ+ဖြေ	Answer	1
2	Znk8.txt	ကြိုး+စား	Try Hard	1
3	Znk8.txt	လက်+ယာ	Left side	1
4	Znk10.txt	စွမ်း+အန်	Capable	1
5	Znk11.txt	ဆောင်+နှင်း	Winter	1
6	Znk11.txt	တီး+မှုတ်	Play	1
7	Znk11.txt	မင်း+မြော်က်+တန်ဆာ	King's gifts	2
8	bible64.txt	တွေ့+မြင်	Visible	1
Total		8 errors		9

Table 5
Type 2 error list

No.	Document	Error words	Meaning in English	Missed merging
1	znk5.txt	ဗြာဟ္မဏ+မဟာ+သာလ	Pali name	2
2	znk6.txt	မိ+ထိ+လာ+ပြည်	Country name	3
Total		2 errors		5

Table 6
Type 3 error list

No.	Document	Error words	Meaning in English	Missed merging
1	wtdy1.txt	အ+လက်+က+ရုံ	Not applicable	3
2	wtdy1.txt	သာ+ကီ+ဝင်	Not applicable	2
3	znk3.txt	နိသမ္မ+ဉာဏ်	Not applicable	1
4	znk6.txt	ဣ+တိ+တသ္မာ	Not applicable	2
Total		4 errors		8

Table 7
Type 4 error list

No.	Document	Error words	Meaning in English	Missed merging
1	Znk9.txt	ကောင်း+ခြင်း+ငါး+ဖြာ	Five goodness	3
Total		1 error		3

extended dictionary specializing in Pali words. A similar situation can arise with words adopted directly from English and other languages. Table 6 lists the Pali words encountered.

4.3.4. Type 4 error – numerical words

In the Myanmar language, numerical words are followed by measure words, which are different for humans, animals and objects. This problem can be addressed with a rule-based heuristic approach. Alternatively, WFST and the neural network approach can be used [14].

4.4. Error analysis chart for syllable merging

Figure 4 shows the distribution of error categories for the syllable merging review. It can be observed that 93% of the errors come from dictionary coverage problems, while the remaining 7% can be eliminated by making some improvements to the current algorithm for merging numerical words.

From the evaluation results for syllable segmentation and syllable merging, we can conclude that our combination of heuristic and dictionary-based statistical approaches has been effective and practical. The limitation comes mainly from the limited coverage of the dictionaries.

4.5. Comparison of word segmentation methods for Thai, Vietnamese and Myanmar

Table 8 compares six word segmentation studies for the Thai, Vietnamese and Myanmar languages. All studies used statistical approaches in combination with other techniques, since the available dictionaries do not provide adequate coverage. Note that Vietnamese word segmentation does not require prior syllable segmentation, as syllables in Vietnamese words are separated by white spaces. In contrast, Thai and Myanmar word segmentation does require prior syllable segmentation because syllables are not separated. Sornlertlamvanich et al. [8] tried to segment words directly, without taking care of syllable segmentation first, but produced unsatisfactory results. Of the six studies, three scored above 95% accuracy, including our proposed method. Compared to the other five methods, our proposed approach for Myanmar word segmentation, which took into account the characteristics of the Myanmar language and script, has done very well.

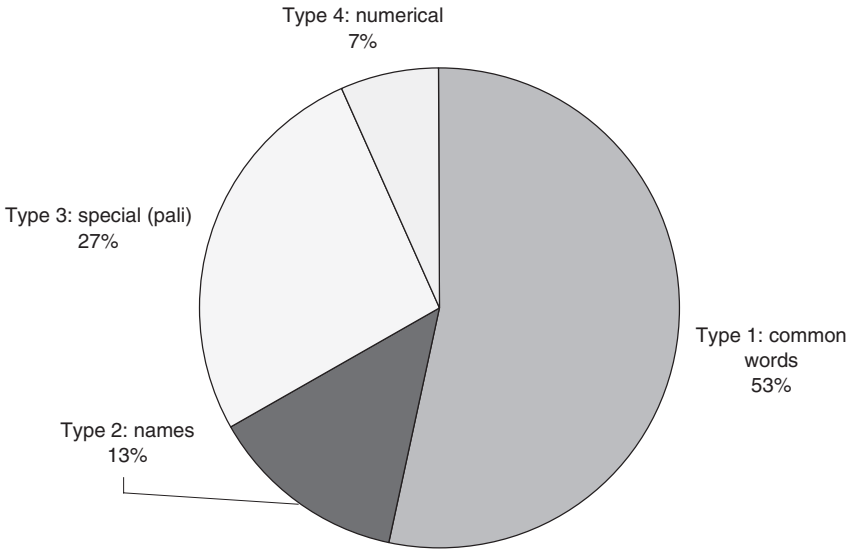

Fig. 4. Error analysis chart for syllable merging.

Table 8
Thai, Vietnamese and Myanmar word segmentation methods

Segmentation study	Method used	Dictionary	Corpus	Statistical-based	Other techniques	Results
Thai word segmentation (Sornlertlamvanich et al., 2000[8])	C4.5 learning algorithm using string length, frequency, mutual information and entropy		Y Y	Y Y	Y	Precision: 85% Recall: 56% *F*-measure: 67.5%
Thai word segmentation (Wirot, 2002 [5])	**Syllable segmentation:** statistics-based trigram model **Syllable merging:** maximum collocation strength	Y		Y		Precision: 96.36% Recall: 97.16% *F*-measure: 96.76%
Vietnamese word segmentation (Dinh et al., 2001 [14])	**Syllable segmentation:** white space **Syllable merging:** Weighted Finite State Transducer (WFST) model and neural network	Y		Y	Y	97% accuracy
Vietnamese word segmentation (Ha, 2003 [12])	**Syllable segmentation:** white space **Syllable merging:** statistical model and n-gram model using maximum probability		Y	Y		'Agreement': 51% 'Reasonable': 65%
Vietnamese word segmentation (Nguyen et al., 2006 [15])	**Syllable segmentation:** white space **Syllable merging:** statistical model using online corpus approach and genetic algorithm	Y	Y	Y		'Acceptable': 80%

(*continued*)

Table 8 (continued)

Segmentation study	Method used	Dictionary	Corpus	Statistical-based	Other techniques	Results
Myanmar word segmentation (our proposed method)	**Syllable segmentation:** rule-based heuristic approach **Syllable merging:** dictionary-based statistical approach	Y	Y	Y	Y	Precision: 98.94% Recall: 99.05% *F*-measure: 98.99%

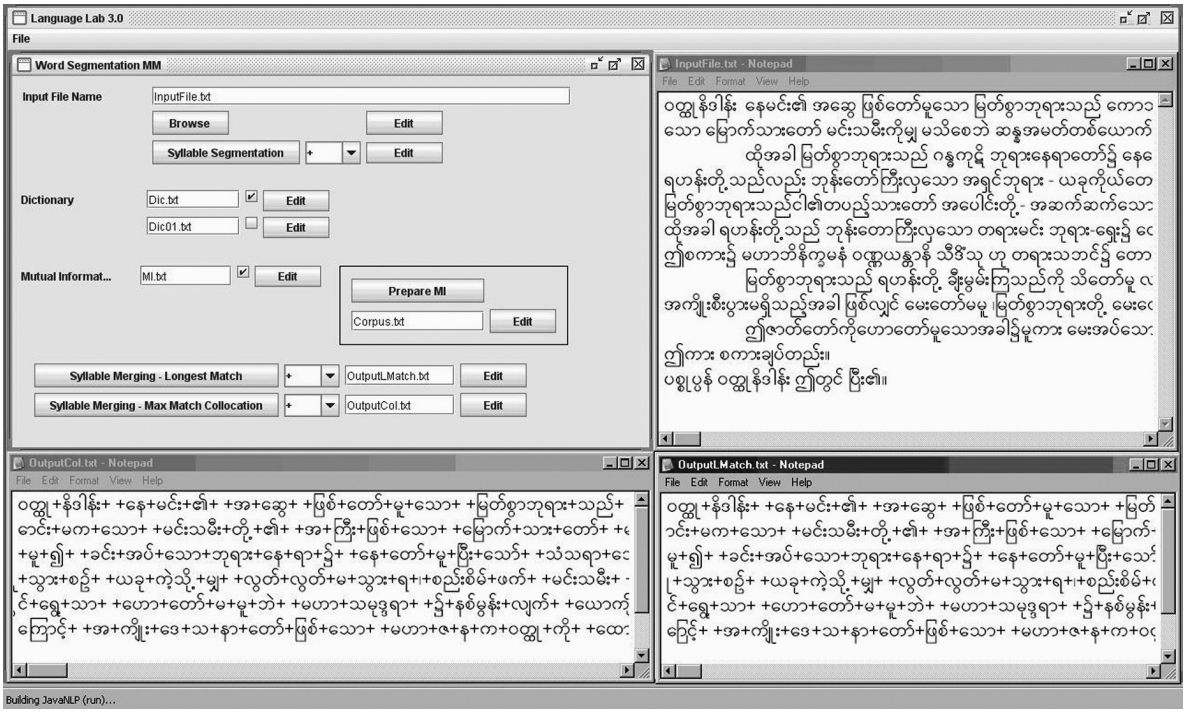

Fig. 5. Our proposed Myanmar word segmentation tool.

5. Implementation

A Myanmar word segmentation tool, based on our dictionary-based statistical approach, was designed using the Object-Oriented methodology and developed in the Java 1.5 language, using Swing components for the GUI interface.

Figure 5 shows a screen shot of the various components of the tool. The user can use the 'Browse' button to browse and select an input text file in Unicode character codes from the file dialog box, and its contents will be loaded into the memory. The 'Edit' button can be used to edit the contents. The 'Syllable Segmentation' button segments the input text into syllables using the syllable segmentation rules in Section 3.2, and displays the results in a text file called syllables.txt. The drop down list next to the 'Syllable Segmentation' button allows the user to choose a syllable separator symbol, which by default is '+'. The button 'Prepare MI' is used to calculate mutual information of bigram syllables using an input corpus. The mutual information is stored in a text file called 'MI.txt'. The 'Syllable Merging – Longest Match' button merges the syllables into words using a dictionary-based 'longest matching' algorithm; this function is used for comparison with our dictionary-based statistical approach. The 'Syllable Merging – Max Match Collocation' button merges the syllables into words

using our dictionary-based statistical algorithm, and the merged words are displayed in an output text file named OutputCol.txt. The drop down lists next to the syllable merging buttons allow the user to choose a word separator symbol, which can be '+', '|' or a line break.

6. Conclusion

In this study, we developed a word segmentation method for the Myanmar language. Dictionaries and a lexicon database provided by the Myanmar NLP team were used during the evaluation of the proposed method.

The proposed strategy was in two parts: rule-based syllable segmentation and dictionary-based statistical syllable merging. First, input texts in Unicode character codes were scanned and segmented as syllables using a rule-based algorithm. Six syllable segmentation rules were applied in the algorithm. The rules proved to be very effective, with the algorithm achieving perfect precision, recall and *F*-measure for the 16 test documents. The next step adopted a dictionary-based statistical approach for syllable merging, using dictionaries and the collocation strength of a sentence or phrase. Scores for precision, recall and *F*-measure were above 98%. The reasons for such excellent results include the reliability of rule-based syllable segmentation, the compatibility of the proposed dictionary-based statistical approach with the Myanmar language, and the presence of many single-syllable Myanmar words. Analysis showed that most of the merging errors were due to inadequate dictionary coverage.

Although the study was a success, there are some obvious limitations, one being its dependence on dictionaries, which currently do not offer comprehensive coverage. Another is the lack of a large corpus of Myanmar Unicode documents.

Future work on our proposed method includes testing and evaluation on a larger data set, refining and adding to the syllable segmentation heuristic rules, and improving the dictionary-based statistical approach for better efficiency, effectiveness and functionality. Research is also needed to examine the application and effectiveness of pure statistical models without using any dictionaries. As more Myanmar documents become available in Unicode, the use of a large corpus and statistical models could lead to interesting results for Myanmar word segmentation and other NLP tasks.

References

[1] A.S. Hornby, *Oxford Advanced Learner's Dictionary* (Oxford University Press, Oxford, 2005).
[2] G.G. Chowdhury, *Introduction to Modern Information Retrieval* (Library Association, London, 1999).
[3] Myanmar NLP, *Myanmar Unicode Reference Documents and Research Papers* (2006). Available at: http://www.myanmars.net/unicode/doc/index.htm (accessed 1 January 2007).
[4] Unicode Consortium, *The Unicode Standard 4.0: Southeast Asian Scripts* (Addison Wesley, California, 2004).
[5] A. Wirot, Collocation and Thai Word Segmentation, *Proceedings of Joint International Conference of SNLP-Oriental COCOSDA 2002* (Thammasat University, Bangkok, Thailand, 2002).
[6] T. Theeramunkong and S. Usanavasin, Non-dictionary-based Thai word segmentation using decision trees, *Human Language Technology Conference: Proceedings of the first international conference on Human language technology research* (Association for Computer Linguistics, NJ, 2001).
[7] A. Kawtrakul, C. Thumkanon, Y. Poovorawan, P. Varasrai and M. Suktarachan, Automatic Thai unknown word recognition, *Proceedings of the nature language processing pacific rim symposium* (Phuket, Thailand, 1997).
[8] V. Sornlertlamvanich, T. Potipiti and T. Charoenporn, Automatic corpus-based Thai word extraction with the c4.5 learning algorithm, *Proceedings of the 18th Conference on Computational linguistics* (Association for Computer Linguistics, NJ, 2000).
[9] J.R. Quinlan, *C4.5 Programs for machine learning* (Morgan Kaufmann, California, 1993).
[10] S.F. Chen and J. Goodman, An empirical study of smoothing techniques for language modelling, *Proceedings of the 34th Annual Meeting on Association for Computer Linguistics* (Association for Computer Linguistics, NJ, 1996).

[11] C.D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing* (MIT Press, London, 2002).
[12] L.A. Ha, A method for word segmentation in Vietnamese, *Proceedings of Corpus Linguistic 2003,* (University of Lancaster, Lancaster, 2003).
[13] D. Dinh, Building a training corpus for word sense disambiguation in English-to-Vietnamese machine translation, *Proceedings of the 19th International Conference on Computational Linguistics* (Association for Computer Linguistics, NJ, 2002).
[14] D. Dinh, H. Kiem and N. V. Toan, Vietnamese word segmentation, *Proceedings of Neural Networks and Natural Language Processing* (Tokyo, Japan, 2001).
[15] T.V. Nguyen, H.K. Tran, T.T.T. Nguyen and H. Nguyen, Word segmentation for Vietnamese text categorization: an online corpus approach, *Proceedings of 4th IEEE International Conference on Computer Science Research, Innovation and Vision of the Future,* (HoChiMinh City, Vietnam, 2006).
[16] Z. Michalewicz, *Genetic Algorithms + Data Structure = Evolution Programs* (Springer-Verlag, London, 1996).
[17] Myanmar Sar A Phwae, *Myanmar Spelling Bible, Myanmar Thut Pon Kyan* (Myanmar Sar A Phwae, Yangon, 2003).
[18] Z. Htut, *Myanmar-Thai Co-workshop on Myanmar Language Implementation, Input Methods and Basic Encoding in Myanmar Language.* Available at: http://www.myanmars.net/unicode/doc (accessed 1 January 2007).
[19] K.W. Church and P. Hanks, Word association norms, mutual information, and lexicography, *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* (Association for Computer Linguistics, Vancouver, Canada, 1989).
[20] C.J. Van Rijsbergen, *The Geometry of Information Retrieval* (Cambridge University Press, London, 2004).