Myanmar OOV Words Extraction with Maximal Substrings and its Application to Document Clustering

Yuzana Win[1] and Tomonari Masada[2]

[1] Yangon Technological University, Insein, Yangon, Myanmar
yuzanawinn@gmail.com
[2] Graduate School of Engineering, Nagasaki University, Nagasaki, Japan
masada@nagasaki-u.ac.jp

Abstract. This paper proposes a method for out-of-vocabulary (OOV) words extraction from Myanmar text with maximal substrings. Our method aims to extract OOV words that can be added into the Myanmar dictionary. The outcome of our method are new compound words which are not described in the Myanmar dictionary. Our method, *firstly*, extracts maximal substrings from Myanmar news articles. Maximal substrings are defined as the substrings whose numbers of occurrences are reduced by any of its extensions. *Secondly*, we make a post-processing of maximal substrings, because the resulting maximal substrings contain noisy characters. In our post-processing, we reduce the number of maximal substrings and remove maximal substrings whose prefixes and suffixes are meaningless characters. We keep only the substrings that consist of words from the existing dictionary. As a result, we obtain the substrings as candidates of new compound words that can be added into the existing Myanmar dictionary. We perform the evaluation both from the subjective and quantitative perspectives. From the subjective perspective, we compare the new compound words extracted by our method with those extracted by word bigrams method. It is found that our method is better than the word bigrams method based on the evaluation using a pooling procedure. From the quantitative perspective, we use the extracted compound words as additional features in K-means clustering. The experimental results show that the document clusters given by our method are better than those given by word bigrams method in precision, recall and F-score.

Keywords: OOV Words, Maximal Substrings, Compound Words, K-means Clustering, Document Clustering

1 Introduction

Recently, researchers have focused on the mining of a large amount of data in natural language processing tasks such as machine translation, word segmentation, part of speech (POS) tagging and so on. Many of those tasks rely on unsupervised methods because it is difficult to prepare a large amount of annotated training data for supervised methods. However, in most of such unsupervised methods, we should first split the text of the documents into sequences of words. In English, French or German, we can easily extract such sequences of words, because each word is delimited by white spaces. In contrast, for the languages such as Japanese, Chinese and Myanmar, it is not a trivial task to extract such sequences of words, because there are no white spaces between words or phrases. In Myanmar script, although spaces are sometimes used for separating words or phrases to read more easily, there are no clear rules for using spaces between words in a sentence. Therefore, word extraction is a challenging task. Moreover, Myanmar words can be combined to make a new word. For instance, the two words "ခရီးသွား" (travel) and "လုပ်ငန်း" (occupation; business; job) are isolated words. However, by combining them, a new word "ခရီးသွားလုပ်ငန်း" (tourism) is formed. A new word "ခရီးသွားလုပ်ငန်း" (tourism) is a combination of two words but the meaning of this combination of two words cannot be derived from those of its constituents. The word is thus regarded as a new word in a non-trivial manner. In this way, when the new word is not presented in the data sets of known words (e.g. in the existing Myanmar dictionaries), out-of-vocabulary (OOV) word problem occurs. That is, we need to make distinction between the cases where a sequence of words should be split into isolated words and the cases where a sequence of words should be regarded as a single *compound word*.

From the above analysis, this paper provides a method for out-of-vocabulary (OOV) words extraction from Myanmar text with maximal substrings [8]. Our method extracts maximal substrings from a given Myanmar document set. The extracted maximal substrings are expected to be useful as new compound words in Myanmar text. Some of the compound words like "ရာသီဥတုပြောင်းလဲ" (climate change), "စီးပွားရေးပိတ်ဆို့" (economic sanctions), "စီးပွားရေးနယ်ပယ်" (economic field), examples of the outcome of our method, are not described in the existing Myanmar dictionary. Compound words are a concatenation of the words that can be found in dictionaries. However, not all concatenations of dictionary words are a compound word. This is the reason why we extract maximal substrings, which are a good candidate of unknown compound words. Precisely speaking, the outcome of our method is a set of candidates of compound words. However, when no ambiguity exists, we call them compound words, not candidates of compound words, for avoiding redundancy.

We give a brief explanation of our method. Our method is twofold. *Firstly*, we extract maximal substrings from a large set of Myanmar news articles. After the extraction, the resulting maximal substrings contain some noisy characters. Therefore, *secondly*, we make a post-processing of maximal substrings. Our post-processing consists of three steps. In the first step, we reduce the number of maximal substrings because the number of resulting maximal substrings is too large. In the second step, we remove maximal substrings that begin or end with meaningless characters (e.g. '့', 'း', '။'). After the removal, as the third step, we extract OOV words that are the substrings consisting of two words from the existing dictionary. Details of each step will be explained later.

Our first paper [15] describes a method for exploring out-of-vocabulary (OOV) words from Myanmar text by using maximal substrings. However, not all the substrings given by our method can be regarded as an OOV word. Therefore, in this paper, we perform evaluations to clarify how useful our method is in finding OOV words. We evaluate the performance of the extracted compound words as additional features in an unsupervised manner. This is because our method does not require any domain knowledge. We do not need to prepare the training data sets where each document is given a category label that may be used as a hint for extracting compound words. For evaluating the effectiveness of our method, we compare the compound words given by our method with those given by word bigrams method, which is also an unsupervised method. We compare these two methods both from the subjective and quantitative perspectives. This evaluation is not included in [15].

The rest of this paper is organized as follows. Section 2 describes the related works. Section 3 presents the nature of Myanmar language. Section 4 explains the procedure of proposed method. Section 5 includes the results of evaluation experiment. Finally, Section 6 concludes the paper with discussion on future work.

2 Related Works

The extraction of OOV words, e.g. proper names, locations, foreign words and new words, is relevant to our problem. The OOV words are important content words which are of great interest in text mining. For example, in a search system for news articles, where many OOV words appear as compound words, without identifying such OOV words, the system cannot find the specific information that a user needs. Therefore, many researchers have proposed various methods to address the extraction of OOV words. OOV words detection methods are effectively used in machine translation [4, 5, 11] and cross language information retrieval [12]. Language models [9, 17] and lexicon-induction-based methods [14] have also been used to address specific kind of OOV words. A recent research [13] focused on OOV words in the speech recognition task.

Zhang et al. [10] proposed a method for Chinese OOV term detection and POS guessing based on the fusion of multiple features and supervised learning. After performing their post-processing of word segmentation, they extracted candidate strings as features by using local, statistical and global feature representation. Then, they used the constraints and heuristic rules to get OOV term candidates. According to their results, the proposed method successfully detected important OOV words that do not exist in Chinese Basic Dictionary.

Shen et al. [6] proposed a method for Chinese unknown word extraction by using maximized substrings [7]. The authors extracted maximized substrings, i.e., the substrings whose number of occurrences are decreased after adding any character before or after them, as unknown word candidates. They extracted maximized substrings by using a two-level hash structure. In their post-processing, they removed meaningless characters by applying short-term store and lexicon-based techniques. According to their results, the proposed method successfully explored important OOV words such as names of persons, locations and technical terms but was not successful in finding compounds, noun and verb phrases, and partial words.

Our approach. We propose a different method for OOV words extraction. Our method consists of two steps. We, *first*, extract maximal substrings as sequences of characters by applying the tool in Becher et al. [1]. *Second*, we make a post-processing of the results given by the extraction of maximal substrings. Our post-processing is threefold. We reduce the number of maximal substrings based on their frequencies and remove the maximal substrings that begin or end with meaningless characters. We then find OOV words that are the substrings consisting of two words from the existing dictionary. Consequently, our method obtains new compound words that can be added into the existing dictionary. Shen et al. [6] also extract Chinese OOV words with maximal substrings [7], but their post-processing is totally different from ours. This is because most Chinese words are composed of single-characters that have their own meanings. In contrast, only a few single-characters in Myanmar language have a meaning (e.g. ‘ဤ’ (this), ‘၌’ (at), ‘၍’ (therefore), ‘၏’ (ending word “to be”), ‘၎င်း’ (that; above-mentioned)). Our method also differs from Zhang et al. [10] by using maximal substrings that are good candidates of compound words.

3 The Nature of Myanmar Language

The Myanmar script is an *abugida* or syllabic writing system [16] in which most of the syllables are composed of more than one character. Most of Myanmar characters are rounded in shape and the script is written from left to right. Myanmar alphabet are composed of 33 main consonants (e.g. က,…, အ), 4 medials (e.g. -ျ, ြ, ွ, ှ), 12 basic vowels (e.g. ါ, -ာ, -ိ, -ီ, -ု, -ူ, ေ-, -ဲ, -်, -ံ , -့, -း), 11 independent vowels (e.g. ဤ, ဧ, ဪ, ၌, ၍, ၏, ဣ, ဉ, ဦ, ဩ, ၎င်း), and 10 digits (e.g. ၀, ၁, ၂, ၃, ၄, ၅, ၆, ၇, ၈, ၉), respectively.

There are nine parts of speech (POS) in Myanmar grammar, namely: noun (N), pronoun (Pron), adjective (Adj), verb (V), adverb (Adv), particle (Part), conjunction (Conj), post-positional marker (P) and interjection (Inter) [3]. Myanmar words are composed of single or multiple syllables, and there are no spaces between words or syllables. Words in Myanmar language can be divided into simple words, compound words and complex words. In this paper, we mainly focus on compound words because we can use them as a single unit of meaning. A compound word is a combination of two or more simple words. Compound words are widely seen in every language including Myanmar language. Myanmar Language is also very rich in compound words. In our proposed method, we can recognize compound words like “မော်တော်ယာဉ်ထုတ်လုပ်မှု” (motor vehicle production), “ဆိပ်ကမ်းတည်ဆောက်” (port construction), “ရွှေ့ပြောင်းဒုက္ခသည်” (displaced person) and “ပြောင်းရွှေ့နေထိုင်” (immigrate), which are not described in the existing Myanmar dictionary.

4 The Proposed Method

In this section, we explain the details of the two steps in our proposed method.

4.1 Maximal Substrings

Maximal substrings are defined as the substrings each giving a smaller number of occurrences by any of its extensions and occurring at least twice. We give an explanation of maximal substring as follows. For example, the string S = abeacadabea. In the string S, ‘abea’ is a maximal substring and it occurs twice in S. Each of the extensions of this substring occurs less than twice. While ‘abe’ occurs twice in S, but it is not a maximal substring, because one of its extensions ‘abea’ also appears the same number of times. In this manner, when there is a way to extend a substring without decreasing the number of its occurrences, we cannot call such substring maximal substring.

On the left panel of Figure 1, *firstly*, we extract maximal substrings from a large set of Myanmar news articles. Before extracting maximal substrings, we collect the document sets in Myanmar news articles as the input text. We remove the punctuation mark (e.g. ‘။’ (pote ma) which is placed at the end of a sentence, comma, question mark, etc.) and other functional characters (e.g. parentheses, hyphen, center dot, etc.) in each sentence. We make a single long string of characters by concatenating all paragraphs for each news articles as an input file to the extraction of maximal substrings. We then apply the tool developed by Becher et al. [1, 2] [1].

In our experiment, we set the minimum length of maximal substrings to be 50. There are two reasons why we use the minimum length of maximal substrings is 50. The one reason is that the minimum length larger than 50 is too long to obtain a sufficient amount of compound words. The other reason is that the minimum length smaller than 50 is too small to make a compound word, because each Myanmar character is encoded with three bytes (e.g. “ဂျပန်ဝန်ကြီးချုပ်” (Prime Minister of Japan) consisting of 17 characters and thus of 51 bytes). This is the reason why we can extract new compound words most effectively when the minimum length of maximal substrings is 50 [15]. After the extraction, the results given by maximal substrings contain noisy characters as illustrated in Table 1. Therefore, we make a post-processing of the obtained maximal substrings to get clear and precise results.

4.2 Post-processing of Maximal Substrings

Secondly, our method applies a post-processing to the results given by the extraction of maximal substrings. Table 1 gives examples of the maximal substrings obtained in our experiment. All Myanmar syllables start with a consonant (e.g. ‘က’,…,‘အ’), the vowel ‘ေ’ or the medial ‘ြ’. But, for example, the substrings like “့ရွေးကောက်ပွဲတွေ”, “းလျှပ်စစ်ဓာတ်အား”, “။းပေါင်းဆောင်ရွက်” and “ါနဲ့ပတ်သတ်လို့” have the first character that is not a consonant (i.e., ‘့’, ‘း’, ‘။’, ‘ါ’). On the other hand, the substrings like “ရင်းနှီးမြှုပ်နှံသ”, “မြန်မာအစိုးရနဲ့ြ” and “တွင်တိုင်းရင်းသားေ” have the last character that is a consonant, a medial or a vowel (i.e., ‘သ’, ‘ြ’, ‘ေ’). If the character is a consonant, a medial or a vowel, we cannot say that this character is placed at the end of a word. The above observation is used in our post-processing, which consists of three steps as shown in the right panel of Figure 1.

[1] Our usage is: ./findpat [options] *input1 input2 output_dir min_length* Let *input1* be an input file. And we set *input2* to /dev/null. We use *–p* and *+f* options to find any strings of characters. We choose the value of the minimum length of maximal substrings (we called *min_length*) to control the output size.

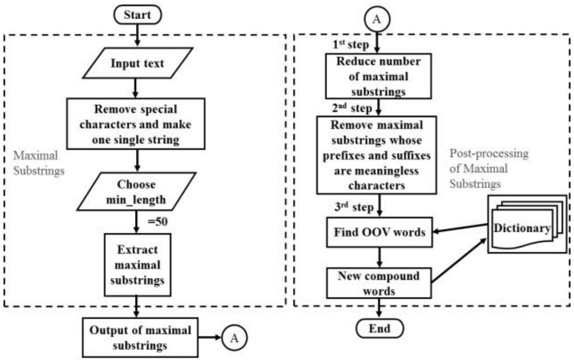
Start
Input text
Remove special characters and make one single string
Maximal Substrings
Choose min_length
=50
Extract maximal substrings
Output of maximal substrings
A
A
1st step
Reduce number of maximal substrings
2nd step
Remove maximal substrings whose prefixes and suffixes are meaningless characters
Post-processing of Maximal Substrings
3rd step
Find OOV words
Dictionary
New compound words
End

Fig. 1. The flowchart of our proposed method

Table 1. Examples of maximal substrings in VOA, MMTimes and Mizzima data sets

မှာရင်းနှီးမြှုပ်နှံ	ပ်ငန်းရှင်များ	င်္မာနေရှင်နယ်လိဂ်
တွေတိုးမြှင့်သွား	းလျှပ်စစ်ဓာတ်အား	အသင်းကိုကိုင်တွယ်
ရင်နိုင်ငံရေးပါတီ	ရင်းနှီးမြှုပ်နှံသ	င်ငံခြားရေးဝန်ကြီး
့ရွေးကောက်ပွဲတွေ	တွင်စားသောက်ဆိုင်	စ္စတာအိုဘားမားက
မြန်မာအစိုးရနဲ့ြ	ူးပေါင်းဆောင်ရွက်	တ္တရားနှောင့်ယှက်မှ
ါနဲ့ပတ်သတ်လို့	သူမြန်မာနိုင်ငံသား	တွင်တိုင်းရင်းသားေ

(1) Reduce the number of maximal substrings

In the first step, we reduce the number of maximal substrings by setting the frequency threshold for each data set, because the number of resulting maximal substrings is too large. Our experiment uses six data sets, i.e., VOA, MMTimes, Mizzima, Mizzima_MMTimes, VOA_MMTimes and Mizzima_MMTimes1, whose details are given later. For each of the six data sets, we remove the maximal substrings whose frequencies are less than or equal to 7, 7, 5, 4, 8 and 7, respectively. These threshold values are automatically calculated and chosen so that we could obtain the best results in evaluation.

(2) Remove maximal substrings whose prefixes and suffixes are meaningless characters

In the second step, we further reduce the number of maximal substrings based on the nature of Myanmar syllables. We set five rules to remove maximal substrings that begin or end with meaningless characters.

The first rule is that the first character must be a consonant, the vowel 'ေ' or the medial 'ြ'. We thus remove maximal substrings like "့ရွေးကောက်ပွဲတွေ", "းလျှပ်စစ်ဓာတ်အား", "ူးပေါင်းဆောင်ရွက်", "င်္မာနေရှင်နယ်လိဂ်" and "ါနဲ့ပတ်သတ်လို့" because their first characters (i.e., '့', 'း', 'ူ', 'င်္' and 'ါ') are vowel symbols. The second one is that the first character must not start with the consonant character with Asat '်' symbol (e.g. 'ပ်', 'င်', etc.). That is, we discard the substrings like "ပ်ငန်းရှင်များ", "င်ငံခြားရေးဝန်ကြီး", etc. This is because Asat '်' is used in conjunction with double consonants (i.e., 'လုပ်' (do; work, carry out, etc.), 'နိုင်' (win)). The third one is that the first character must not be a preposition, a particle, a conjunction or a pronoun. Therefore, we remove the substrings like "မှာရင်းနှီးမြှုပ်နှံ", "တွေတိုးမြှင့်သွား", "ရင်နိုင်ငံရေးပါတီ" and "သူမြန်မာနိုင်ငံသား" because their prefixes (i.e., 'မှာ' (at), 'တွေ' (plural noun), 'ရင်' (if) and 'သူ' (one who does something; he)) are a preposition, a particle, a conjunction or a pronoun. The fourth one is that the first character must not start with double-stacked consonants. So, we remove the substrings like "တ္တရားနှောင့်ယှက်မှု" and "စ္စတာအိုဘားမားက", because they should be stored at the middle position of the syllables e.g. "ဝတ္တရား" (responsibility; duty) and "မစ္စတာ" (to write the title Mr.), not as a prefix. The fifth one is that the last character must not be a consonant, the vowel 'ေ' or the medial 'ြ'. We remove the substrings like "ရင်းနှီးမြှုပ်နှံသ", "တွင်တိုင်းရင်းသားေ" and "မြန်မာအစိုးရနဲ့ြ" because their last characters (i.e., 'သ', 'ေ' and 'ြ') need to be combined with some characters to compose meaningful character streams.

(3) Find OOV words by using the dictionary

As the third step, we find OOV words, i.e., the words which are not contained in the dictionary of known words, by using the dictionary. We collect the words from Myanmar-English Dictionary[2], which includes 20,778 words in total.

To perform the OOV word extraction, there are two methods for using the dictionary. The first method is a simple approach, i.e., the extraction of the substrings consisting of two words. The second method is the extraction of the substrings consisting of more than two words by implementing the recursive search.

In the first method, we first find a prefix of the substrings that can be recognized as a word in the dictionary. Then, we also find a suffix of the substrings that can be matched with a word in the dictionary. If the substrings consist of the prefix and the suffix that are found in the dictionary, it can be said that those substrings are candidates of compound words that may be inserted into the existing dictionary. In the second method, if a prefix of the substrings is present in the dictionary, then we recursively check for the remaining part. We may find that the remaining part can be segmented into a sequence of dictionary words by applying the recursive search.

According to the preliminary experiment, we found that the substrings obtained by the second method were not useful. This is because many of the substrings consisting of more than two dictionary words cannot be regarded as compound words. Therefore, we only use the first method.

5 Experimental Results

5.1 Data Sets

We used six data sets in our experiment. Each data set contains already categorized articles downloaded from the Web. All Myanmar news articles are written in Zawgyi-One Myanmar font, because this font is the most commonly used in Myanmar Web pages. We chose this experimental data set for evaluation the quality of compound words extracted by our method to predict the categories by clustering documents. We

[2] http://myanmar-dictionary.blogspot.sp/2009/08/myanmar-english- dictionary-version-10.html

discuss how we collected each data set as below.

The first data set is a set of news articles from the VOA (Voice of America)[3] Burmese website. We denote this data set as VOA. This data set originally consists of six categories. We collected news articles from the two categories: International and Myanmar domestic. The numbers of the documents contained in each category is almost the same as given in Table 2. This table also includes the specifications of other data sets. The second one is a set of news articles from the Myanmar Times[4] Burmese website. We denote this data set as MMTimes. This data set also originally consists of six categories. We collected news articles from the two categories: Business and National-news. The third one is a set of news articles from the Mizzima[5] Burmese website. We denote this data set as Mizzima. This data set also originally consists of six categories. We collected news articles from the two categories: Sports and World-news.

The fourth one is a mixture of three news categories, i.e., Sports, World-news, and Business categories, from Mizzima and MMTimes data sets. We denote this data set as Mizzima_MMTimes. The fifth one is a mixture of three news categories, i.e., International, Myanmar domestic, and National-news categories, from VOA and MMTimes data sets. We denote this data set as VOA_MMTimes. The sixth one is a mixture of all news categories, i.e., Sports, World-news, Business and National-news categories, from Mizzima and MMTimes data sets. We denote this data set as Mizzima_MMTimes1.

In the last three data sets, we mixed documents from different sets to make our evaluation more reliable. This is because it makes the experiment more reliable to evaluate the methods under a larger number of different settings.

Table 2. Numbers of documents belonging to each category in the six data sets

VOA		
International	*Myanmar Domestic*	*Total*
7,320	7,176	14,496

MMTimes		
Business	*National-news*	*Total*
1,130	4,188	5,318

Mizzima		
Sports	*World-news*	*Total*
1,463	2,589	4,052

Mizzima_MMTimes			
Sports	*World-news*	*Business*	*Total*
1,463	2,589	1,130	5,182

VOA_MMTimes			
International	*Myanmar Domestic*	*National-news*	*Total*
7,320	7,176	4,188	18,684

Mizzima_MMTimes1				
Sports	*World-news*	*Business*	*National-news*	*Total*
1,463	2,589	1,130	4,188	9,370

[3] http://burmese.voanews.com/

[4] http://myanmar.mmtimes.com/

[5] http://mizzimaburmese.com/

5.2 Evaluation in Document Clustering

Document clustering is the process of grouping documents into topics without any knowledge of the category structure that exists in the entire document set. All semantic information is derived from the documents themselves and it is often referred to as unsupervised clustering. Therefore, we evaluated the candidate compound words extracted by our method as additional document features in an unsupervised manner. This is because our method itself does not require any domain knowledge. We do not need to prepare the training data sets where each document is given a category label that may be used as a hint for extracting compound words. We applied K-means clustering for obtaining document clusters, because K-means is widely used method for document clustering. And then we checked whether the compound words extracted by our method improved the quality of clusters in the prediction of document categories.

We evaluated our method in document clustering of news articles obtained from the six data sets described in Section 5.1. We removed all function words by regarding them as stop words, e.g. preposition, pronouns, conjunctions, particles, etc. To obtain a bag of words representation for each document, we applied a Burmese word segmenter[6]. This segmenter parses Myanmar syllabic structure by using a dictionary. However, the segmenter could not give any good evaluation results. Therefore, we tokenized the sentences into words simply by using white spaces according to the original texts. This led to better evaluation results and this is the *baseline* method in our evaluation. We used TF-IDF term weighting to obtain a feature vector for each document based on the formula: $tf\_idf(t,d) = tf(t,d) \times log(N/df(t))$, where $tf(t,d)$ is the frequency of the term t in document d, and $df(t)$ is the document frequency of t, i.e., the number of documents where t appears. N is the total number of documents.

In our experiment, we prepared another compared method. We found dictionary words in each document and used them as additional document features other than the strings separated by white spaces. The dictionary words are expected to represent particular topics better than the strings simply separated by white spaces. However, this method can only find the compound words already presented in the dictionary. In contrast, our method can provide candidate compound words by using maximal substrings. Therefore, we used the candidates given by our method as additional document features and compared the resulting cluster quality with those achieved by the two methods given above. If the candidate compound words given by our method are important in the sense that each of them is closely related to a particular topic and thus helps discriminating among different topics, clustering results may be improved.

We measured the results of document clustering in precision, recall and F-score, each defined by Eqs. (1), (2) and (3), respectively. We calculated precision, recall and F-score for each category.

$$Precision = \frac{TP_{correct}}{TP_{correct} + FP_{incorrect}} \tag{1}$$

$$Recall = \frac{TP_{correct}}{TP_{correct} + FN_{missing}} \tag{2}$$

[6] https://github.com/lwinmoe/segment

$$F\text{-}score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \,, \tag{3}$$

where $TP_{correct}$ is the total number of correctly identified documents in the same category, $FP_{incorrect}$ is the total number of incorrectly identified documents in the same category and $FN_{missing}$ is the total number of correctly identified documents in different categories. *F-score* is the harmonic mean of precision and recall. Our evaluation measure is the macro averaged F-score over all categories. For each data set and each compared method, we used K-means clustering by setting the number of clusters K equal to or larger than the true number of clusters. To choose the correct number of clusters K is important so that we adopted the *silhouette analysis*. We used scikit-learn machine learning library for performing K-means clustering and computing silhouette coefficients[7]. To make our evaluation more reliable, we compute the *F-score* averaged over the ten results obtained from ten different random seeds used in K-means. Then the mean and standard deviation of *F-score* were recorded.

Table 3 summarizes the *p*-values obtained by comparing the *Baseline* (i.e., using no dictionary words and no compound words), *Find_dict* (i.e., only using dictionary words as additional features) and *Ours* (i.e., using both dictionary words and the compound words given by our method). For each data set and each compared methods, we removed words of lower frequency to obtain better clustering results. In Table 3, the *p*-values are obtained in a paired two-sided *t*-test. When the *p*-value is less than 0.05, we can say that the improvement is statistically significant and thus give the *p*-value in bold. Table 3 shows the number of clusters K given by the silhouette analysis. Only for two data sets, i.e., MMTimes and Mizzima, the *Baseline* method could get a significantly better F-score than the *Find_dict* and our method. In contrast, for three data sets, i.e., VOA, Mizzima_MMTimes and VOA_MMTimes, we could get a significantly better F-score than the *Baseline* and *Find_dict*.

Based on these results, it can be said that the document clusters given by our method are better than those given by the *Baseline* and *Find_dict* method. So we claim that our method can extract the features that are useful in discriminating among different topics.

5.3 Comparing with Word Bigrams Method

To discuss the special nature of compound words extracted by our method, we compared our method with word *bigrams* method, i.e., a method extracting all *bigrams* of dictionary words from documents. The preliminary experiment revealed that many of the concatenations of more than two dictionary words cannot be regarded as compound words. Therefore, we only consider word bigrams.

Word bigrams method extracts all concatenations of two dictionary words as candidate compound words. In contrast, our method first extracts maximal substrings and then refines them by finding bigrams of dictionary words in the 3rd step of the post-processing as described in Section 4.2. Therefore, the compound words given by our method are different from those given by word bigrams method. However, to achieve a fair comparison, we made the number of candidate compound words given by word

[7] http://scikit-learn.org/

bigrams method as close as possible to that of candidate compound words given by our method. Therefore, we removed bigrams of low frequency by adjusting the threshold. For each of the six data sets, i.e., VOA, MMTimes, Mizzima, Mizzima_MMTimes, VOA_MMTimes and Mizzima_MMTimes1, we removed the word bigrams whose frequencies are less than or equal to 9, 3, 1, 1, 11 and 3, respectively. These threshold values turned out to lead to a better result in evaluation than other threshold values. We compared our method and word bigrams method both from the subjective and quantitative perspectives.

In the evaluation from the subjective perspective, we asked five Myanmar natives to examine which candidates could be regarded as new compound words for both methods. For example, the compound words like "ငြိမ်းချမ်းရေးဆု", "ဆိပ်ကမ်းတည်ဆောက်", "အခွင့်ရေးကောင်း", "ပြောင်းရွှေ့နေထိုင်", "ရွှေ့ပြောင်းဒုက္ခသည်", "နှစ်ပတ်လည်အစည်းအဝေး", "အလွန်အမင်းစိုးရိမ်", examples of the outcome of our method, are not described in the existing Myanmar dictionary. They can be found using two nouns (N+N), noun and verb (N+V), noun and adjective (N+Adj), two verbs (V+V), verb and noun (V+N), adjective and noun (Adj+N), and adverb and verb (Adv+V), respectively. We regarded them as compound words because these words are no ambiguity exists. However, not all concatenation of two words is a compound word. For example, the two words like "စီးပွားရေး" (economy) and "ကြိုတင်" (advance) are isolated words. By combining them, we can obtain the word like "စီးပွားရေးကြိုတင်". The compound word "စီးပွားရေးကြိုတင်" is not a meaningful word, because the first word "စီးပွားရေး" is a noun, and it is not followed by an adverb "ကြိုတင်". In this manner, when the two words create a misleading word, we cannot regard them as a compound word.

From the above analysis, we compared the methods in subjective evaluation in terms of precision, recall and F-score. The precision is defined as W_{new} / $W_{candidate}$, where W_{new} is the number of new compound words, i.e., the word bigrams that are regarded as compound words by Myanmar natives and are not listed in the dictionary, and $W_{candidate}$ is the number of candidate compound words given by the two compared methods. The recall is defined as W_{new} / $W_{correct}$, where $W_{correct}$ is the number of all correct compound words found by our method and word bigrams method. That is, we adopted a pooling procedure for calculating recall. The F-score is defined as the harmonic mean of precision and recall.

Table 4 summarizes the results of evaluation from the subjective perspective. For all data sets, i.e., VOA, MMTimes, Mizzima, Mizzima_MMTimes, VOA_MMTimes and Mizzima_MMTimes1, our method provided a larger number of new compound words than word bigrams method. We could add 214 new compound words for VOA, 150 for MMTimes, 88 for Mizzima, 220 for Mizzima_MMTimes, 169 for VOA_MMTimes and 202 for Mizzima_MMTimes1 to the existing Myanmar dictionary as new words, respectively. Further, the precision, recall and F-score of our method were better than those of word bigrams method for all data sets. It is indicated that our method provided new compound words than word bigrams method even though the number of candidate compound words is the same. This is because word bigrams method provided many misleading compound words. Based on these results, it can be concluded that our method is better than word bigrams method from the subjective perspective.

From the quantitative perspective, we compared our method with word bigrams method also in document clustering described in Section 5.2. We used K-means clustering for obtaining document clusters and checked if the compound words extracted by our method were more useful as additional document features. The number of clusters was chosen based on the silhouette analysis. The F-score was averaged over the ten results obtained from ten different random seeds. Then the mean and standard deviation of F-score were recorded.

Table 5 summarizes the *p*-values obtained by comparing word bigrams method and our method in terms of F-score. Both for word bigrams method and our method, we removed words of lower frequency to obtain better clustering results. The *p*-values are obtained in a paired two-sided *t*-test. When the *p*-value is less than 0.05, we can say that the difference is statistically significant and thus give the *p*-value in bold. Table 5 also shows the number of clusters K given by the silhouette analysis. Only for two data sets, i.e., Mizzima and Mizzma_MMTimes1, the word bigrams method could get a significantly better F-score than our method. In contrast, for the three data sets, i.e., VOA, Mizzima_MMTimes and VOA_MMTimes, we could get a significantly better F-score than word bigrams method.

The experimental results show that the document clusters given by our method are better than those given by word bigrams method. This is because word bigrams method would degrade the quality of clustering since it would provide many misleading compound words. Therefore, it can be said that our method extracts better compound words than word bigrams method also from the quantitative perspective.

5.4 Examples of New Compound Words

In order to examine whether maximal substrings extracted by our method can represent new compound words well, we here explain with examples how the outcome of our method can be regarded as new compound words.

We can observe that many multiword expressions that can be regarded as compound words given by our method. Table 6 summarizes the new compound words obtained by our method. For example, as presented in Table 6, the word "ငြိမ်းချမ်းရေး" (peace) and the other word "ဆု" (prize; award) are combined to form a meaningful compound word like "ငြိမ်းချမ်းရေးဆု" (peace prize). Further, the meaning of the compound word is sometimes derived from the meaning of its constituents in a non-trivial manner. For example, "အထိမ်းအမှတ်" + "ပြုလုပ်" = "အထိမ်းအမှတ်ပြုလုပ်" ("token" + "do" = "commemorate") is recall and show respect for someone or something. According to the above examples, the new compound words obtained by our method can utilize various specific meanings as a single unit of meaning.

6 Conclusion

In this paper, we proposed a method for out-of-vocabulary (OOV) words extraction from Myanmar text with maximal substrings. Our method extracted candidates of new compound words based on a procedure consisting of two steps. We evaluated the extracted compound words as additional document features in K-means clustering. The experimental results based on the comparison with the two other methods showed that the quality of clustering results was improved. We also compared the compound words given by our method with those given by word bigrams method. We evaluated the performance of our method both from the subjective and quantitative perspectives. From the subjective perspective, the precision, recall and F-score of new compound words given by our method were improved than those given by word bigrams method. From the quantitative perspective, the results of the experiment using K-means clustering in a similar manner showed that the document clustering results given by our method were better than those given by word bigrams method in precision, recall and F-score. Further, we described the examples of new compound words obtained by our method. It can be expected that our method will be also useful in other natural language processing tasks for Myanmar language data.

We also have a plan to evaluate maximal substrings extracted by our method in a multi-topic analysis e.g. by performing a probabilistic Latent Semantic Analysis (pLSA), where the extracted compound words are used as additional features.

Table 3. Results of the quantitative comparison of the three methods

Data Sets	***F-score***			***p-value***
	Baseline	***Find_dict***	***Ours***	
VOA	0.9135 ± 0.0005 (K=2)	0.8791 ± 0.0265 (K=3)	**0.9194 ± 0.0003** **(K=2)**	**0.001**
MMTimes	**0.7724 ± 0.0198** **(K=3)**	0.7611 ± 0.0265 (K=3)	0.6891 ± 0.0668 (K=3)	**0.005**
Mizzima	**0.9737 ± 0.0000** **(K=2)**	0.9735 ± 0.0001 (K=2)	0.9735 ± 0.0001 (K=2)	**8.5E-06**
Mizzima_MMTimes	0.8854 ± 0.0024 (K=5)	0.9595 ± 0.0005 (K=3)	**0.9599 ± 0.0003** **(K=3)**	**0.019**
VOA_MMTimes	0.8916 ± 0.0003 (K=4)	0.8947 ± 0.0006 (K=4)	**0.8950 ± 0.0007** **(K=4)**	**1.4E-07**
Mizzima_MMTimes1	0.7631 ± 0.0278 (K=6)	**0.7638 ± 0.0217** **(K=6)**	0.7604 ± 0.0255 (K=6)	0.877

Table 4. Results of the evaluation for comparing our method with word bigrams method from the subjective perspective

Data Sets	***Word Bigrams***						***Ours***					
	$W_{candidate}$	W_{new}	$W_{correct}$	***Precision***	***Recall***	***F-score***	$W_{candidate}$	W_{new}	$W_{correct}$	***Precision***	***Recall***	***F-score***
VOA	304	143	327	0.47	0.47	0.45	312	**214**	327	**0.69**	**0.65**	**0.67**
MMTimes	250	84	221	0.34	0.38	0.36	234	**150**	221	**0.64**	**0.68**	**0.66**
Mizzima	104	36	120	0.35	0.30	0.32	117	**88**	120	**0.75**	**0.73**	**0.74**
Mizzima_MMTimes	314	70	279	0.22	0.25	0.24	307	**220**	279	**0.72**	**0.79**	**0.75**
VOA_MMTimes	272	126	267	0.46	0.47	0.47	269	**169**	267	**0.63**	**0.63**	**0.63**
Mizzima_MMTimes1	296	95	282	0.32	0.34	0.33	301	**202**	282	**0.67**	**0.72**	**0.69**

Table 5. Results of the comparison from the quantitative perspective

Data Sets	***F-score***		***p-value***
	Word Bigrams	***Ours***	
VOA	0.7524 ± 0.0003 (K=3)	**0.9194 ± 0.0003** **(K=2)**	**2.4E-12**
MMTimes	0.6637 ± 0.0739 (K=4)	**0.6891 ± 0.0668** **(K=3)**	0.556
Mizzima	**0.9740 ± 0.0003** **(K=2)**	0.9735 ± 0.0001 (K=2)	**0.001**
Mizzima_MMTimes	0.9568 ± 0.0002 (K=3)	**0.9599 ± 0.0003** **(K=3)**	**3.4E-10**
VOA_MMTimes	0.8919 ± 0.0008 (K=4)	**0.8950 ± 0.0007** **(K=4)**	**2.1E-06**
Mizzima_MMTimes1	**0.8354 ± 0.0039** **(K=4)**	0.7604 ± 0.0255 (K=6)	**1.0E-05**

Table 6. Examples of new compound words

ငြိမ်းချမ်းရေး	+	ဆု	=	ငြိမ်းချမ်းရေးဆု
peace	+	prize; award	=	peace prize
(N)	+	(N)	=	(N)
အထိမ်းအမှတ်	+	ပြုလုပ်	=	အထိမ်းအမှတ်ပြုလုပ်
sign; symbol; indication; token	+	do; carry out	=	commemorate
(N)	+	(V)	=	(V)
ထွက်ပြေး	+	တိမ်းရှောင်	=	ထွက်ပြေးတိမ်းရှောင်
flee; run away	+	evade; avoid	=	abscond
(V)	+	(V)	=	(V)
ခေါင်းဆောင်	+	ဟောင်း	=	ခေါင်းဆောင်ဟောင်း
leader	+	old; ancient	=	former leader
(N)	+	(Adj)	=	(N)
စီးပွားရေး	+	အကျပ်အတည်း	=	စီးပွားရေးအကျပ်အတည်း
economy	+	difficulty; crisis	=	economic crisis
(N)	+	(N)	=	(N)

Acknowledgments

This work has been supported by the Grant-in-Aid for enhancement of engineering higher education of the Japan International Cooperation Agency (JICA) from 2017 to 2018. We are grateful for their support.

References

1. Becher. V., Deymonnaz, A., Heiber, P.: Efficient computation of all perfect repeats in genomic sequences of up to half a Gigabyte, with a case study on the Human genome.

Bioinformatics. vol. 25, no. 14, pp. 1746-1753 (2009). doi: 10.1093/bioinformatics/btp321.

2. Barenbaum, P., Becher, V., Deymonnaz, A., Halsband, M., Heiber, P.A.: Efficient repeat finding in sets of strings via suffix arrays. Discrete Math. Theor. Comput. Sci. vol. 15, no. 2, pp. 59-70 (2013).
3. 10th Grade Myanmar Grammar Book. Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar. vol. 2, no.5 (2015).
4. Razmara, M., Siahbani, M., Haffari, G., Sarkar, A.: Graph propagation for paraphrasing out-of-vocabulary in statistical machine translation. In: 51st Annual Meeting of the ACL 2013, pp. 1105-1115, Sofia, Bulgaria (2013).
5. Daumé III, H., Jagarlamudi, J.: Domain adaptation for machine translation by mining unseen words. In: 49th Annual Meeting of the ACL, pp. 407-412, Portland, Oregon (2011).
6. Shen, M., Kawahara, D., Kurohashi, S.: Chinese unknown word extraction by mining maximized substrings. In: 20th Annual Meeting of the Association for NLP, pp. 384-387, Sapporo, Japan (2014).
7. Shen, M., Kawahara, D., Kurohashi, S.: Chinese word segmentation by mining maximized substrings. In: 6th Inter. Joint Conf. on NLP (IJCNLP), pp. 171-179, Nagoya, Japan (2013).
8. Okanohara, D., Tsujii, J.: Text categorization with all substrings features. In: 9th SIAM Inter. Conf. on Data Mining (SDM), pp. 838-846 (2009).
9. Botha, J. A., Dyer, C., Blunsom, P.: Bayesian language modeling of German compounds. In: 24th Inter. Conf. on Computational Linguistics (COLING), pp. 341-356, Mumbai, December (2012).
10. Zhang, Y., Cen, L., Wu, W., Jin, C., Xue, X.: Fusion of multiple features and supervised learning for Chinese OOV term detection and POS guessing. In: 22nd Inter. Joint Conf. on Artificial Intelligence (IJCAI), vol. 22, no.3, pp. 1921-1926 (2011).
11. Durrani, N., Sajjad, H.: Integrating an unsupervised transliteration model into statistical machine translation. In: 14th Conf. of the European Chapter of the Association for Computational Linguistics, pp. 148-153, Gothenburg, Sweden (2014).
12. Huang, D., Zhao, L., Li, L., Yu, H.: Mining large-scale comparable corpora from Chinese-English news collections. In: COLING 2010: Poster Volume, pp. 472-480, Beijing (2010).
13. Sheikh, I., Illina, I., Fohr, D., Linarès, G.: Improved neural bag-of-words model to retrieve out-of-vocabulary words in speech recognition. In: INTERSPEECH 2016, pp. 675-679, San Francisco, United States (2016).
14. Kaewpitakkun, Y., Shirai, K., Mohd, M.: Sentiment lexicon interpolation and polarity estimation of objective and out-of-vocabulary words to improve sentiment classification on microblogging. In: 28th Pacific Asia Conf. on Language, Information and Computation (PACLIC 28), pp. 204-213 (2014).
15. Win, Y., Masada, T.: Exploring oov words from Myanmar text using maximal substrings. In: 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 657-663, Kumamoto, Japan (2016). doi: 10.1109/IIAI-AAI.2016.73.
16. Daniels, P.T., Bright, W.: The World's Writing Systems. New York Oxford: Oxford University Press (1996).
17. Xiong, D., Zhang, M., Li, H.: Enhancing language models in statistical machine translation with backward n-grams and mutual information triggers. In: 49th Annual Meeting of the ACL, pp. 1288-1297, Portland, Oregon (2011).