N-gram of Orthographic Syllable Indexing for Khmer Information Retrieval System

Channa VAN and Wataru KAMEYAMA

Abstract In the conventional approach of text retrieval, words are usually used as the unit of indexing. However, in many Asian languages such as Chinese, Japanese and Thai, word segmentation is still an open issue and still actively addressed by many researchers to achieve high segmentation accuracy especially when dealing with the out-of-vocabulary words. This issue also occurs in Khmer, and it affects Khmer information retrieval (IR). To solve the issue, we propose an indexing approach based on n-grams of Khmer orthographic syllable instead of words. The experiments show that the proposed approach outperforms the conventional word based indexing approach. Therefore, n-grams of orthographic syllables can be an alternative to words in case of Khmer IR.

1 Introduction

In text retrieval, words are widely used as the unit of indexing for many languages because semantics of words can be used to retrieve the most relevant documents corresponding to the search query. In general, texts in documents are split into words, and the words are indexed for searching afterward. Therefore, the performances of the text retrieval system depend on the performance of the word segmentation. In many languages such as Chinese, Japanese and Khmer, achieving very high performance in word segmentation is difficult due to the complexity of the writing system of these languages in which words are written continuously without any delimiters.

Instead of using a conventional word-segmentation based approach, the n-gram of characters or syllables approach has been investigated by many researchers. Theeramunkong et al. have studied character cluster indexing based on various proposed sets of rules to segment Thai characters [1]. In their study, their results show that the proposed approach outperforms the traditional methods in all measures. Nie has studied the use of word and n-grams for Chinese IR (Information Retrieval) systems, and the result of the n-gram is comparable to that of word-based indexing [2]. For Japanese, Ogawa and Matsuda have proposed an efficient approach using n-gram indexing for Japanese IR systems, where character level indexing has shown the significant result compared to conventional word-based indexing [3]. Character level indexing based on n-gram is easy to achieve in the case of languages like Chinese. In the Chinese writing system, each character represents a phonologic syllable and might be a word. However, it is difficult to make character level indexing for language such as Khmer because of the complexity of the Khmer writing system. The Khmer writing system consists of many types of characters including consonants, two types of vowels, many diacritics and symbols. The complexity of the writing makes it difficult to segment the Khmer phonologic syllables. The other option is to use orthographic syllables of which the segmentation is easy to achieve.

In this research, we explore the utilization of the Khmer orthographic syllables in the field of Khmer IR. N-gram of orthographic syllables is used instead of n-gram of phonologic syllables as in the case of Chinese. The results are evaluated by comparing to the conventional approach that is based on words.

This paper is organized as follows: in the following section, we briefly introduce the Khmer writing system and the issues of Khmer word segmentation as well as the issue of Khmer phonologic syllable segmentation. In Section 3, details of the proposed approach of Khmer orthographic syllable segmentation are described. Then, the building Khmer test collection for evaluating the proposed experimental IR system is presented in Section 4. Section 5 shows the experimental procedure as well as the results of our proposed approach. The comparison between our proposed approach and the conventional approach is also shown in this section. Finally, we conclude our work in Section 6.

2 Khmer Writing System

2.1 Khmer Writing System

Unicode is the only existing encoding that can be
used to encode the Khmer text. In the Unicode chart, the Khmer script consists of 35 consonants, 17 independent vowels, 16 dependent vowels, 13 diacritics, 7 punctuations, a special subscript character and several other signs. Words can be formed by only a consonant or the combination of consonants, vowels, subscripts and diacritics together. These symbols are arranged in 5 layers as shown in Fig. 1.

Fig. 1 The 5 Layers of Khmer Writing System

In Khmer text, words are written continuously without any delimiter. Fig. 2 shows an example of Khmer phrase, which consists of 3 words: "Cambodia", "kingdom" and "wonder". The first line in the figure is the original text written in Khmer script, and the second line shows the boundaries of the words in the sentence by the vertical lines.

Fig. 2 Example of Words in a Khmer Phrase

2.2 Issues of Khmer Word Segmentation

There are two obvious issues in the Khmer word segmentation: over-segmentation and word-segmentation ambiguity. The over-segmentation issue is the most common one, which is mostly caused by the OOV (Out of Vocabulary) words found in text. Fig. 3 shows two examples of the over-segmentation. The first one is the case of a name "Obama", which is segmented into three different clusters of characters using a dictionary-based approach because there is no such word in it. The second example shows the OOV word "supermarket", which is incorrectly over-segmented into two different words. The issue dramatically decreases the precision and the recall of the word segmentation as it increases the number of incorrectly segmented terms as well as decreases the correct segmented words.

The word-segmentation ambiguity issue is rather rare in Khmer word segmentation. But occasionally, two words may be incorrectly segmented into a single word. Fig. 4 shows an example of the word-segmentation ambiguity.

As the majority of the issues in the Khmer word segmentation are caused by the over-segmentation mainly by OOV words, solving the problem of OOV words is the most crucial work in order to increase the performance of the word segmentation.

Fig. 4 Example of Word Segmentation Ambiguity

2.3 Issues of Khmer Phonologic Syllable Segmentation

There is no study yet on Khmer phonologic syllable segmentation as well as it is not well studied yet from the linguistic point of view. Due to the complexity of the Khmer Unicode writing system as well as the pronunciation, it is very difficult to define a set of general rules to segment the Khmer phonologic syllables.

As in the example shown in Fig. 4, ambiguity also occurs in phonologic syllable segmentation. In this example, it is difficult to define the exact rule of phonologic syllable segmentation without using the context to disambiguate the word. Another example in Fig. 5 shows two possible pronunciations of a word. Generally, the first pronunciation is the correct way to pronounce the word to obtain the correct meaning which is "ice". However, the second pronunciation is also possible. In the second case, the word becomes an OOV word, and its meaning is changed to "a type of water" or "Kaka water".

Fig. 5 Ambiguity of Khmer Phonologic Syllables

In Khmer, each consonant can be a phonologic syllable. When it is combined with another consonant, it produces a new phonologic syllable. For example, ណ[kaː] and ណ[kák] in Fig. 5 are two different phonologic syllables. This shows how difficult to read Khmer text even for native speakers without any context understanding.

Based on the presented examples, the issues of phonologic syllable segmentation are comparable to the issues of word segmentation in terms of complexity. To solve
these issues, further study of Khmer phonological syllable segmentation is required to improve its segmentation accuracy. However, when comparing it to the orthographic syllable segmentation which is presented in the next section, the orthographic syllable segmentation becomes easy and robust when using our proposed rule of Khmer orthographic syllables. Since our goal is not to study Khmer phonologic syllables segmentation but n-gram indexing for Khmer information retrieval, we have decided to use only orthographic syllables for the n-gram indexing.

3 Orthographic Syllable Segmentation and Indexing

3.1 Khmer Unicode Orthographic Syllable

Since Unicode 4.0 was available, a standard Khmer Unicode encoding has been defined. According to Solá [4], a Khmer orthographic syllable in Unicode can be expressed in the following form:

\[ B \{ R \mid \{ Z \} F \} \{ S \{ R \} \} \{ \{ Z \} F \} \{ O \} \{ S \} \] (1)

Where \( B \) is the base character which can be a consonant or an independent vowel, \( R \) is “Ꭰ” character, \( F \) is a consonant shifter, \( S \) is a subscript consonant or an independent vowel, \( V \) is a dependent vowel, \( Z \) is the zero width non-joiners, and \( O \) is any other sign. This form is quite complex due to the many types of characters, and it has been proposed for analyzing the issues of writing and displaying Khmer Unicode characters. Therefore, specifically for orthographic syllable segmentation, we propose a simplified form of Khmer orthographic syllable as shown in Eq. 2 which is a simplified form based on the Eq. 1. Table 1 shows the comparison of our proposed form with the original form.

The proposed Khmer orthographic syllable form consists of only 3 groups of characters as follows:

\[ I \{ JC \} \{ C \{ A \} \} ^+ \{ CJ \{ I \} \{ C \{ JC \} \} \{ A \} \} ^* \] (2)

Where \( I \) is an independent vowel, \( C \) is a consonant, \( J \) is a subscript sign, and \( A \) is any other character. The simplification is done by dividing and grouping the character groups in Eq. 1. The characters in \( B \) are divided into two groups of characters \( I \) and \( C \) while the characters in \( S \) are divided into three groups of characters \( J \), \( C \), and \( A \). On the other hand, the remaining groups are grouped into \( A \). This new form shows that any Khmer orthographic syllables can be in one of three basic forms ranging from simple to complex: \( I \{ JC \} \), \( C \{ A \} \) or \( CJ \{ I \} \{ C \{ JC \} \} \{ A \} \).

In Khmer, a word can be composed of a single orthographic syllable or many orthographic syllables together, and the orthographic syllable itself is composed of a single character or more complex characters. Therefore, an orthographic syllable is a unit that varies from a character to a word. This makes the orthographic syllable a good candidate as the unit of segmentation instead of using words.

<table>
<thead>
<tr>
<th>Objective</th>
<th>Proposed for analyzing the issue of written and displaying Khmer Unicode</th>
<th>Proposed for segmenting Khmer orthographic syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>Simple and easy:</td>
<td>Easy:</td>
</tr>
<tr>
<td></td>
<td>• Less representations of characters</td>
<td>• Orthographic syllables are clearly separated.</td>
</tr>
<tr>
<td></td>
<td>• Easy to understand</td>
<td>• 100% of segmentation accuracy is achieved in our experiment as shown in subsection 3.1.</td>
</tr>
<tr>
<td>Implementation for segmentation</td>
<td>Difficult:</td>
<td>Easy:</td>
</tr>
<tr>
<td></td>
<td>• Because the objective is different from segmentation.</td>
<td>• Easy to understand</td>
</tr>
<tr>
<td></td>
<td>• Confusing information makes it impossible to be implemented.</td>
<td>• No confusing information.</td>
</tr>
</tbody>
</table>

3.1 Proposed Orthographic Syllable Segmentation Approach

Based on Eq. 2, we can define a general grammar of a Khmer Unicode text \( T \). In Khmer text, many punctuation symbols are used, such as question mark, exclamation mark, Khmer period symbol and so on. Therefore, a Khmer Unicode text \( T \) is a sequence of orthographic syllables and the punctuation symbols \( P \). Eq. 3 shows the general form of the Khmer Unicode text:

\[ T = \{ P \} \{ I \{ JC \} \} \{ C \{ A \} \} ^+ \{ CJ \{ I \} \{ C \{ JC \} \} \{ A \} \} ^* \] (3)

Eq. 3 obviously shows that any Khmer Unicode texts can be segmented easily into the sequence of orthographic syllables and punctuation symbols. For example:

- 柬柬 (Cambodia) = 柬 柬 | 柬 = 柬 CJC4 | C4
- 例例 (example) = 例 例 | 例 例 例 = 例 C4 | C | C | C4
- 乐乐 (music) = 乐 乐 | 乐 = C4 | CJC | C | CJC4

By employing this general form, we can easily and
effectively segment the orthographic syllables of the Khmer Unicode texts. We have also evaluated the orthographic syllables segmentation results with a text consisting of 2456 orthographic syllables manually segmented. The accuracy of the segmentation achieves 100%.

3.2 Orthographic Syllable and Writing Constraints

Compared with word segmentation, the disadvantage of orthographic segmentation is that it generates more segmented tokens. Therefore, in order to minimize the number of generated tokens, we apply the Khmer writing constraints, which is basically a set of agglomeration rules that use the local context of a character. These rules are used to determine whether two tokens shall be combined together or not. Unfortunately, there are no comprehensive studies about them, thus we propose some of the rules based on our observation and knowledge as a native speaker of Khmer.

We define the two types of the rules as below:

- A token to be connected to its next token: any token whose terminal character is “？”.
- A token to be connected to its previous token:
  - Terminal character is one of {“ CATEGORY_3”},
  - In form of $C[\{I\} \{C\} \{J\} \{C\}] \{A\}$, and the first C consonant and the second C consonant are the same.
  - In form of $C[\{I\} \{C\} \{J\} \{C\}] \{A\}$, and the first C consonant is one of {“ CATEGORY_1”}, “ CATEGORY_5"}, “ CATEGORY_6”.
  - In form of $C[\{I\} \{C\} \{J\} \{C\}] \{A\}$, and the pair of the first C consonant and the second C consonant are in the Table 2.

3.3 Orthographic Syllable N-gram Indexing

We have already mentioned in Section 1 that the character-level n-gram indexing has shown promising results compared to the word-based indexing in some languages without explicit delimiters. However, as the character unit is too small in Khmer and almost semantically meaningless, we propose an orthographic syllable n-gram indexing scheme, instead. Since words are composed of one or many orthographic syllables, n-grams of orthographic syllables can be expected to represent the general sense of words; this is especially important when dealing with OOV words, where no information on the word as a whole is available.

Fig. 6 shows the process of the proposed n-gram indexing of orthographic syllables. Text is segmented into orthographic syllables using our proposed segmentation method, then n-gram generator generates different kinds of n-gram of orthographic syllables, and finally the n-grams are indexed.

Suppose a text $T = O_1, O_2, O_3, ..., O_n$ where $T$ is a sequence of tokens drawn from a set of orthographic symbol types ($O_1, O_2, O_3, ..., O_n$). The n-gram generator generates a set of sets that are subsets of all the possible n-grams occurring in the text $T$ (see Table 3). Each set may contain all n-grams for a single value of n (for example, only bigrams as in the second row of the table), or it may contain all the n-grams for pairs of consecutive values of n (for example, the set may contain all unigrams and bigrams, as in the fourth row of the table).

<table>
<thead>
<tr>
<th>N-gram Subset Name</th>
<th>N-gram Subset Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>${O_1, O_2, ..., O_n}$</td>
</tr>
<tr>
<td>Bigram</td>
<td>${O_1, O_1O_2, O_2, O_2O_3, ..., O_n}$</td>
</tr>
<tr>
<td>Trigram</td>
<td>${O_1, O_1O_2, O_2O_3, ..., O_n-1O_n}$</td>
</tr>
<tr>
<td>Unigram &amp; Bigram</td>
<td>${O_1, O_1O_2, O_2O_3, ..., O_n-1O_n}$</td>
</tr>
<tr>
<td>Bigram &amp; Trigram</td>
<td>${O_1, O_1O_2, O_2O_3, ..., O_n-1O_n}$</td>
</tr>
</tbody>
</table>

4 Test Collection

A test collection is very important for evaluating the accuracy as well as the performance of IR system. However, no test collection for Khmer has been ever built. Therefore, we have also built a small test collection to evaluate our proposal when used in a Khmer IR system. We have followed the TREC standard test collection (file format, format, guidelines, creation procedure) [6]. The task starts with collecting Khmer text using a web-crawler from
various websites daily publishing contents in Khmer. Then, the test collection is made using the following pooling technique to create the relevant judgments by combining the results from various retrieval systems.

In order to create the relevant judgments, three popular open source IR systems, Apache Lucene, Sphinx and Xapian, are used. The selection process for relevant documents is as follows:

1. A topic is selected and used in the three IR systems.
2. The top 100 retrieved documents by each IR system are selected to create a pool.
3. Then, a pool of retrieved documents is created by applying the union of the top 100 retrieved documents.
4. Finally, a manual judgment process of each document in the created pool is carried out. The document, which consists of the information relevant to the corresponding topic, is included in the test collection otherwise it is excluded.

The task of building the test collection requires a lot of human involvement and is time consuming. Due to our limited resources, only the top 100 documents are used to create this test collection. Nonetheless, we have created a test collection consisting of 20 topics or queries including 10 OOV words and 10 normal queries. The queries from 1 to 10 are the normal queries without OOV, and the queries from 11 to 20 are the OOV queries. The OOV queries consist of words that are not found in the dictionary, such as names of islands, cities, countries, people, brands, events and so on. All the 10 proposed OOV queries consist of segmentation mistakes when applying word segmentation. Table 4 shows the number of words, the number of orthographic syllable segmentation segmented tokens, and the number of word segmented tokens of each query. The word segmentation tool created by the Cambodia PAN Localization project has been used in the word segmentation process. The average number of the orthographic syllables is 5, and the average number of word segmented tokens is 2.96. Thus, the number of orthographic syllables is almost double of that by word segmented tokens.

5 Experiments and Results

5.1 Experimental Setup and Procedure

In order to evaluate our proposed orthographic syllable n-gram indexing, a small IR system based on Apache Lucene is set up for the experiment. Apache Lucene is a Java-based full-text search engine that can be used for index generating and searching based on the generated index. We have used the default indexing algorithm of Apache Lucene which is based on the inverted index. Furthermore, the default searching algorithm has been also used including scoring and relevant order that are calculated based on the combination of vector space model (VSM) and Boolean model.

Fig. 7 shows the proposed IR system. The input query is segmented into orthographic syllables, then n-gram generator creates the n-gram of the orthographic syllables of the query which is used by Lucene’s searcher to search for the relevant documents, and finally a search result is returned. The index is composed of the orthographic syllable n-grams described in Subsection 3.3.

Furthermore, a text collection consisting of 80K documents is used for indexing and searching. The test collection described in Section 4 was used to evaluate the result afterward. The evaluation is made based on the precision of the top 100 retrieved documents of each query. Moreover, the comparison is also done in order to compare the performance of the proposed n-gram indexing of orthographic syllable and the word-based indexing.

The experiments are carried out based on different types of n-grams of orthographic syllables: unigram, bigram, trigram, combinations of unigram and bigram, and combinations of bigram and trigram. The 20 queries of the test collection are input into the proposed IR system. The same type of n-gram of the index is applied for the search query in order to optimize the retrieved results. For example, if the index is bigram of orthographic syllables, the query must be also bigram of orthographic syllables. Similarly, the same 20 queries are also used in the word-based IR system for evaluation and comparison. In the word-based IR experiments, a word segmentation tool is used to segment the test queries.
5.2 Experimental Results and Discussion

Tables 5 and 6 show the number of the relevant documents of the top 100 results in case of normal queries and OOV queries, respectively. Furthermore, Fig. 8 shows the average of the top 100 precisions of the normal queries, these results were unexpected. Despite using just the n-gram of orthographic syllable indexing even in case of normal queries, the bigram, and the combination of unigram and bigram of orthographic syllables significantly outperform word-based indexing, while the result of the combination of bigram and trigram of orthographic syllables is slightly better. Only the result of the trigram is lower than word-based indexing approach.

On the other hand, a quite similar result can be observed also in case of the OOV queries as shown in Fig. 9. The bigram, the combination of unigram and bigram, and combination of the bigram and trigram of orthographic syllables still get better results than word-based approach.
while the trigram still gets the lowest result. Furthermore, the average precision of word-based indexing is quite low. Just 0.5310 in the case of OOV queries can be achieved compared to 0.7520 in the case of normal queries. This clearly shows that the OOV words strongly affect the results of the IR system when a word is used as indexing unit. However when using n-gram of orthographic syllables, the variation from the normal queries to the OOV queries is not as much as in the case of words. For example, in case of the bigram, 0.6530 is obtained in case of OOV queries while 0.7930 is obtained in case of normal queries.

Taking the case of the query number 18, the word-based and trigram indexing perform the worst, i.e. only 3 relevant documents were retrieved from the top 100. The query was segmented into 3 orthographic syllables and 3 word segmented tokens as follows:

- Orthographic syllables: ʦʧ tai ʦʧ
- Word segmented tokens: ʦʧ tai ʦʧ

Exploring the results based on word segmentation, most of the retrieved documents match at least one of the word segmentation tokens. However, most of these matching tokens are fragments from other OOV words, for example: ʦʧ tai (Taiwan) and ʦʧ tai (Tai chi). These OOV words share the common token ʦʧ as in the query. As a result, it reduces the precision of word based indexing. On other hand, in the case of trigram indexing, it was expected to work better than other approaches since the trigram of the orthographic syllables forms the original words. However, it performs poorly. Surprisingly, most of the retrieved documents in the top 10 did not contain any text matching the query. The relevant documents did occur in the list beyond the 100th entry. As our evaluation is based on only the top 100 retrieved documents, the document relevance ranking using VSM and Boolean models [14] simply does not function well in trigram indexing. Moreover, when combining bigram and trigram, up to 47 relevant documents are retrieved due to the contribution of the bigram.

In summary, the bigram of orthographic syllables performs the best. It achieved the highest performance in our experiments for both types of queries compared to other indexing units. Despite the fact that the average number of orthographic syllables is ~1.7 times larger than that of the words, the bigram is still the best while the trigram is the worst. In addition, the combination of the unigram and bigram as well as the combination of the bigram and trigram did not improve the precision over the bigram. This may be because the combinations generate more constituents than that of the bigram, and the most of constituents do not have any related meaning which decreases the precision. On the other hand, the bigram generates fewer constituents which have related meaning. Furthermore, this result has shown that the bigram of orthographic syllables can replace the conventional word based indexing in Khmer IR systems.

### 6 Conclusion

The proposed orthographic syllable indexing approach shows very promising results and a significant improvement over words-based indexing. We have demonstrated the effective adaptation of an existing n-gram approach to Khmer IR using Khmer orthographic syllables for indexing. Our experimental results show that n-grams of orthographic syllables outperform word-based indexing.

The results will allow us to implement a wide variety of Khmer IR systems based on bigrams of orthographic syllables although Khmer IR is not well studied yet due to the many fundamental issues such as the lack of resources including text corpora and test collections as well as the fundamental research on Khmer natural language processing. The use of n-grams, particularly bigrams, of Khmer orthographic syllables for indexing in Khmer IR is a strong approach that we have shown to be capable of overcoming the primary obstacle encountered in Khmer IR that is word segmentation.

Nevertheless, we believe this research is a topic to be revisited in the future when more progress has been made in fundamental Khmer NLP technology, especially in Khmer word segmentation. Moreover, as phonologic syllables have not been well studied yet, this study and a comparison between phonologic syllable indexing and orthographic syllable indexing shall be investigated in the future.

### References


