Query Expansion for Khmer Information Retrieval

Channa Van and Wataru Kameyama
GITS, Waseda University
Honjo, Saitama, Japan
channa@fuji.waseda.jp, wataru@waseda.jp

Abstract

This paper presents the proposed Query Expansion (QE) techniques based on Khmer specific characteristics to improve the retrieval performance of Khmer Information Retrieval (IR) system. Four types of Khmer specific characteristics: spelling variants, synonyms, derivative words and reduplicative words have been investigated in this research. In order to evaluate the effectiveness and the efficiency of the proposed QE techniques, a prototype of Khmer IR system has been implemented. The system is built on top of the popular open source information retrieval software library Lucene[1]. The Khmer word segmentation tool (Chea et al., 2007) is also implemented into the system to improve the accuracy of indexing as well as searching. Furthermore, the Google web search engine is also used in the evaluation process. The results show the proposed QE techniques improve the retrieval performance both of the proposed system and the Google web search engine. With the reduplicative word QE technique, an improvement of 17.93% of recall can be achieved to the proposed system.

1 Introduction

Similar to the other major languages in the world, the number of Khmer digital content has been rapidly growing over the world-wide web, and it is becoming very difficult to obtain the relevant information as needed from the Internet. Although some major web search engine providers such as Google has a localized version of Khmer in its web search engine[2], it is not specifically designed for Khmer due to the lack of integration of Khmer specific characteristics. Consequently, it misses a lot of information of Khmer websites found in the Internet. For this reason, we propose the QE techniques using the specific characteristic of Khmer to improve the effectiveness of Khmer IR system. Four types of QE technique are proposed based on the four types of Khmer specific characteristic that are spelling variants, synonyms, derivative words (Khin, 2007) and reduplicative words (Long, 2007). Moreover, a prototype of Khmer IR system is implemented in order to evaluate the effectiveness of these proposed QE techniques. The proposed system is built on top of the popular open source information retrieval software library Lucene in order to take the advantage from its powerful indexing and searching algorithms. Furthermore, to improve the accuracy of indexing and searching, we have also implemented the specific Khmer word segmentation into the system. Due to the lack of Khmer text collection which is required in the evaluation process, we have also created our own Khmer text corpus. The corpus has been built to be useful and beneficial to all types of the research in Khmer Language Processing.

2 Khmer Specific Characteristics

Khmer is the official language of Cambodia. It is the second most widely spoken Austroasiatic language family[3]. Due to the long historical contact

[1] Apache Lucene: http://lucene.apache.org

[2] Google Khmer: http://www.google.com.kh

[3] Khmer language: http://en.wikipedia.org/wiki/Khmer_language

Proceedings of the 8th Workshop on Asian Language Resources, pages 80–87,
Beijing, China, 21-22 August 2010.

with India, Khmer has been heavily influenced by the Sanskrit and Pali. However, Khmer still possesses its own specific characteristics so far such as the word derivation rules and word reduplication techniques. Furthermore, the specific written rule is also found in Khmer, for instance the case of multiple spelling words. These characteristics are very useful especially in the Khmer IR due to the lexical-semantic relation between the words.

2.1 Spelling Variants

In Khmer, multiple spelling words exist. They have same meaning and pronunciation but only different in spelling. Most of the spelling variants are loan words, and others are the result of the substitutability between characters. For example:

- The word សមុទ្រ [săműt] "sea" ,which is originated from Sanskrit, can also be spelled សមុទ្ធ [săműt] "sea" (Khmer Dictionary, 1967) which is originated from Pali.
- And the word ឫទ្ធិ [rŏtthĭ] "power" has another spelling រឹទ្ធិ [rŏtthĭ] "power" because "ឫ" can be substituted by "រឹ".

2.2 Synonyms

Synonyms exist in all the natural languages. Thus, there is no exception for Khmer as well. Khmer has rich and variety of synonym vocabularies. Most of these synonyms are found in the loan words (influenced by Sanskrit and Pali) and the social status's words. For instance the word ញុាំ "to eat" has many synonyms for each social status: ស៊ី (impolite word), ពិសា (polite word), ឆាន់ (religious word), សោយ (royal word) and etc.

2.3 Derivative Words

Derivative words in Khmer are the words which are derived from the root words by affixation, including prefixation, infixation and suffixation (Jenner, 1969). Interestingly, derivative word's meaning is semantically related to its root word. For example:

ចាស់ "old" + កញ (prefix) →កញ្ជាស់ "very old"

ដើរ "to walk" + មណ (infix) →ដំណើរ "the walk"

ឯករាជ្យ "independence" + ភាព (suffix) →
ឯករាជ្យភាព "the state of independence"

2.4 Reduplicative Words

Reduplicative words are very common in Khmer. They are used for several purposes including emphasis, pluralization and complex thought expressions. Three kinds of duplicating techniques are found in Khmer: word duplication, synonym duplication and symmetrical duplication (Noeurng and Haiman, 2000).

2.4.1 Word duplication

Word duplication is the process of duplicating a word to express pluralisation meaning. The duplication symbol "ៗ" is put after the word to indicate the duplication. For example:

ឆ្កែធំ [chhkê thum] "a big dog" →
ឆ្កែធំៗ [chhkê thum thum] "many big dogs"

2.4.2 Synonym duplication

Also called synonym compounding words. This kind of words are created by combining two different words which are either synonyms or related meaning. For example:

បំណង "goal" + ប្រាថ្នា "intention" →
បំណងប្រាថ្នា "goal and intention"

2.4.3 Symmetrical duplication

Symmetrical duplication is the process of creating a word by combining a word and its similar phonetic syllables. It is similar to create words which sound like "*zigzag*" or "*hip hop*" in English. This duplication technique is usually used to emphasize the meaning of the original word. There are quite remarkable amount of this kind of words found in Khmer. For example:

ធំ [thum] "big" + ធេង [théng] (similar phonetic syllable) →ធំធេង [thum théng] "very big"

ក្រី [krei] (similar phonetic syllable) + ក្រ [krar] "poor" →ក្រីក្រ [krei krar] "very poor"

3 Khmer Text Corpus

A large collection of Khmer text is required in the evaluation process. Due to the lack of Khmer language resource especially the Khmer text corpus, we have built our own Khmer text corpus. The corpus was designed to be useful and beneficial to all

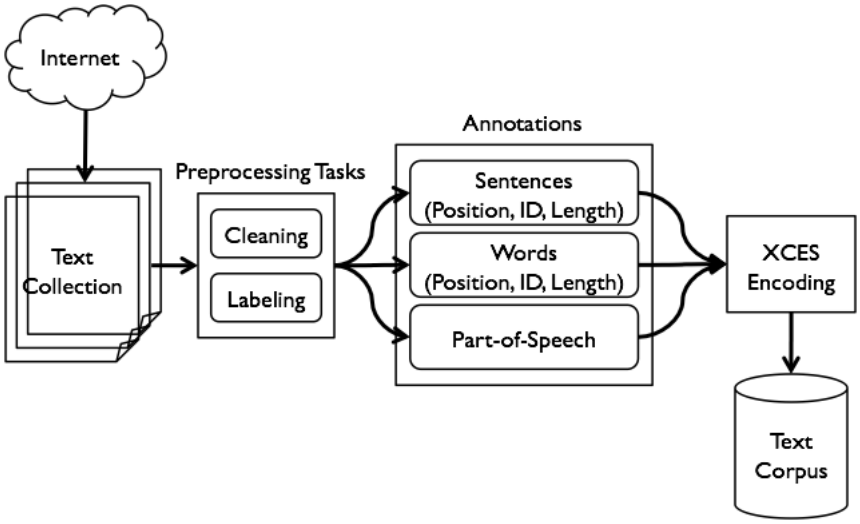

Figure 1: System Design of Building a Khmer Text Corpus

kinds of the research in Khmer language processing. Building such text corpus for Khmer is a challenging task since there is no implementation yet on Khmer optical character recognition, and a few research in Khmer have been done . All texts have to be collected manually from various Khmer website in the Internet. The corpus includes some basic corpus annotations such as word annotation, sentence annotation and part-of-speech annotation. It was encoded in eXtensible Corpus Encoding Standard (XCES) (Nancy et al., 2000) to assure the future extensibility. Figure 1 illustrates the whole process of building a Khmer text corpus in which four main steps were carried out: text collection, preprocessing tasks, corpus annotations and corpus encoding. The detail of each step is described in the following subsections:

3.1 Text Collection

Collecting Khmer digital text is the most difficult and time consuming task in the process of building this corpus. As there is no implementation on Khmer optical character recognition, it is not possible to scan or extract Khmer texts neither from books nor from other paper sources. However, thanks to the websites as well as Khmer blog community that provide the valuable Khmer digital texts for this research. All text were manually collected from all the available sources.

3.2 Preprocessing Tasks

The texts collected from the Internet are usually not clean and unstructured. It may contain unwanted elements such as images, links and some HTML elements. Therefore, cleaning process is carried out to remove the unwanted elements and to restructure the texts before proceeding to the next step.

After cleaning, each text is categorized by its domain according to its content by the labeling process. There are twelve domains in this corpus: newspaper, magazine, medical, technology, culture, history, law, agriculture, essay, novel, story and other. The text descriptions such as author's name, publisher's name, publishing date and publisher's digital address, are kept along with each text.

3.3 Corpus Annotations

Corpus annotation information is very important for the corpus-based research. Therefore, this corpus also includes the sentence annotation, word annotation and POS annotation.

3.3.1 Sentence Annotation

Each sentence is annotated with three kinds of information: position, identification and length.

1. Position: it is defined by the position of the first character and the last character of a sentence in the text.

2. Identification: the sequence number of a sentence within a text file.

3. Length: the number of characters of a sentence.

Like English, each sentence in Khmer can be separated by special symbols. In modern Khmer, there exists a series of characters that are used to mark the boundaries of different kind of sentences (Khin, 2007). Based on these characters, each sentence in a text can be separated easily.

- ។ and ។ល។ : end of declarative sentences.
- ? : end of interrogative sentences.
- ! : end of exclamative sentences.
- ៕ end of the last sentence in a text.

3.3.2 Word Annotation

The position, identification and length of each word are also annotated as in sentence annotation.

1. Position: it is defined by the position of the first character and the last character of a word in the text.

2. Identification: the sequence number of a word within a text file.

3. Length: the number of characters of a word.

Khmer is non-segmented script. There is no separator between words which is very difficult to segment. In order to do that, we have used the Khmer word segmentation tool (Chea et al., 2007) developed by PAN Localization Cambodia[4].

3.3.3 Part-of-Speech Annotation

To enhance the usefulness of the corpus, we also include the Part-of-Speech annotation. We have used a Khmer POS tagger which is based on the work of Nou et al. where a transformation-based approach with hybrid unknown word handling for Khmer POS tagger is proposed (Nou et al., 2007). There are 27 types of Khmer tagset which can be obtained by this Khmer POS tagger. Each obtained POS tag is assigned to each word in the corpus, and it is kept along with the word annotation.

[4]Cambodia PAN Localization: `http://www.pancambodia.info/`

3.4 Corpus Encoding

To assure the extensibility of corpus and the facility of development for the future works, this corpus has been encoded in eXtensible Corpus Encoding Standard (XCES) (Nancy et al., 2000). XCES is an XML-based standard to codify text corpus. It is highly based on the previous Corpus Encoding Standard (Nancy, 1998) but using XML as the markup language. Since the corpus encoding is based on XML, the corpus is suitable for many programming languages which support XML. Furthermore, it can fully take the advantage of the powerful XML framework including XQuery, XPath and so on. In addition, XCES supports many types of corpora especially the annotation corpora which our corpus is based on. The encoding of annotation files and text description files are conformed to XCES schema version 1.0.4.

3.5 Corpus Statistic

Table 1 shows the corpus statistic. We have achieved more than one million words within twelve different domains of text. The corpus size is relatively small at the moment, the expansion of the corpus size is continuously undergoing.

Table 1: Corpus Statistic

Domain	# of Article	# of Sentence	# of Word
Newspaper	571	13222	409103
Magazine	52	1335	42566
Medical	3	76	2047
Technical	15	607	16356
Culture	33	1178	43640
Law	43	5146	101739
History	9	276	7778
Agriculture	29	1484	30813
Essay	8	304	8318
Story	108	5642	196256
Novel	78	12012	236250
Other	5	134	5522
Total	**954**	**41416**	**1100388**

4 Retrieval Environment

This section provides the necessary background to understand the context in which the experiment

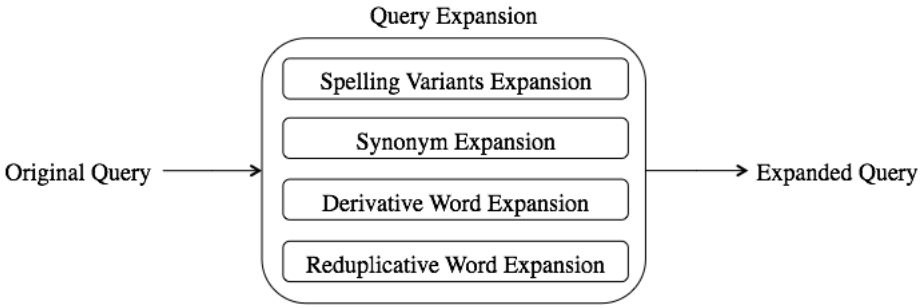

Figure 2: Query Expansion Procedures

was carried out.

4.1 Query Expansion Procedures

Query Expansion is a technique commonly used in IR (Manning et al., 2009) to improve retrieval performance by reformulating the original query. We have proposed four types of QE technique based on the four types of Khmer specific characteristics that we have presented in the section 2. During the expansion process, the original search query is analyzed and expanded corresponding to the type of words (Figure 2). Four types of QE is carried out: spelling variants expansion, synonym expansion, derivative word expansion and reduplicative word expansion. The expanded query is obtained after adding the the expansion term to the original query.

4.2 Khmer IR System Design and Implementation

A prototype of Khmer IR system, which is shown in the Figure 3, has been implemented to evaluate the efficiency of the proposed QE techniques. There are two main processes in the system implementation: indexing and searching. We have started to implement the system by constructing searching index from the text corpus. All texts in the corpus are tokenized into words. Then these words are indexed by the Lucene's indexer, and stored in an index database. On the other hand in the searching part, the search query is tokenized into words before being analyzed in the QE process. Finally, the search results are obtained by the Lucene's searcher which searches through the index database and returns results that correspond to the expanded query.

4.3 Indexing

Indexing is very important because it determines what search results are returned when a user submits query. Our proposed system's indexer is based on the Lucene's indexer with the modifications adapted to Khmer language. Lucene indexes a document by indexing the tokens, which are words, obtained by tokenizing text of the document (Hatcher et al., 2009). Since the default Lucene's tokenizer can only work with segmented western languages such as English or French where spaces are used between each word, it is impossible to tokenize document in Khmer which belongs to the class of non-segmenting language group, where words are written continuously without using any explicit delimiting character. Khmer word tokenizer, which is developed by the a research group of Cambodia PAN Localization, has been used to handle this task.

5 Experiments

The experiment to evaluate the proposed QE techniques was initially conducted only on the prototype of the Khmer IR system that we have implemented. As Google also has a localized Khmer version of its popular web search engine, and it can explicitly specify websites to be searched[5], we have extended our experiment to Google web search engine in order to obtain more precise result.

5.1 Experiment Setup

The Khmer text corpus, which consists 954 documents collected from various websites in the In-

[5] http://www.google.com/advanced_search

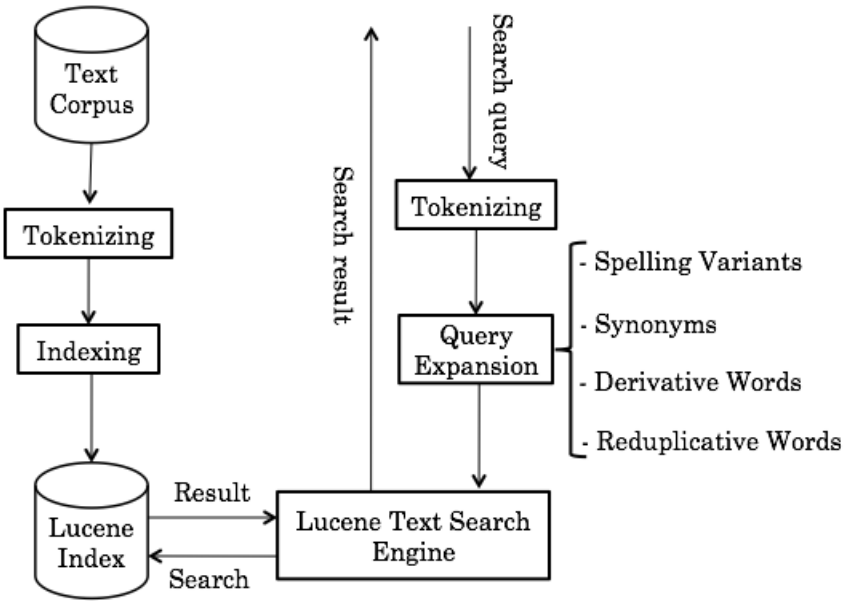

Figure 3: Proposed Khmer Information Retrieval System Design

ternet, was used for the experiments. A website, which contains all documents from the corpus, was hosted in our laboratory's web server in order that these documents can be indexed by the Google's indexer. Then we followed up the indexing progress by consulting the Google Webmaster Tools[6] service. In Khmer where words are written one after the other, word processing programs, Internet browsers and other programs that format text need to know where they can cut a sentence in order to start a new line. This kind of problem does not appear in the western languages where space are used between words. Thus, the zero-width space character was proposed to solve this problem by inputting the zero-width space at the end of each word while typing[7]. In Unicode, the zero-width space character is a type of space character but it is invisible. Using the zero-width space is very confusing because of the invisibility of the character, plus it is unnatural to Khmer writing system. As a result, most people only partly used it for the text display purpose. Therefore, we can find the zero-width space in almost all Khmer texts found in the Internet. Since all texts in the corpus are collected from the Internet, the zero-width space also can be found in almost all the texts in the corpus. Based on the zero-width space character, the Google can separate and index the Khmer texts in our text corpus hosted in our laboratory web server. After Google completely indexed all the documents, the experiment was proceeded.

5.2 Experiment Procedures

We conducted the experiment for each type of proposed QE technique in our implemented system and the Google web search engine. Four similar experiments of the four proposed QE techniques were carried out to the both systems. Due to the small size of the text corpus, only ten original queries have been chosen for each type of experiment. Each query possesses a specific topic in order that we can judge the relevant results after. The experiment processes are as following:

1. Input ten original queries into the both systems, and calculate the precisions and recalls. All queries are selected from the different topics, and each query contains at least an expandable word corresponding to its expansion type.

2. Expand the original queries according to their expansion type. Then input the ten expanded

[6] http://www.google.com/webmasters/tools

[7] Khmer Unicode typing: http://www.khmeros.info/download/KhmerUnicodeTyping.pdf

Table 2: Results of Spelling Variants and Synonyms Expansions

	Spelling Variants			Synonyms		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Google	39.79%	46.72%	37.35%	38.91%	39.99%	34.66%
Proposed Sys.	62.09%	47.63%	50.78%	46.74%	47.03%	44.71%
Google & QE	46.83%	53.04%	46.13%	44.32%	58.99%	45.28%
Proposed Sys. & QE	60.73%	64.01%	58.38%	48.34%	64.59%	51.66%

Table 3: Results of Derivative Word and Reduplicative Word Expansions

	Derivative Words			Reduplicarive Words		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Google	28.35%	56.04%	36.56%	21.14%	42.61%	26.16%
Proposed Sys.	41.07%	51.07%	44.10%	31.69%	39.05%	29.00%
Google & QE	29.18%	60.41%	38.32%	24.69%	48.58%	26.35%
Proposed Sys. & QE	33.93%	62.38%	41.71%	34.28%	56.98%	36.25%

queries into the both systems and recalculate the precisions and recalls.

The relevance judgments were manually done based on the content of each document obtained by the both IR systems. Since there is no Khmer digital thesaurus yet, the expansions were manually done by respecting the query syntax of Lucene and Google[8]. Moreover, as Google web search engine cannot tokenize Khmer words, the tokenization was also done manually. For example: the query តុលាការខ្មែរក្រហម "Khmer Rough tribunal" consists two words តុលាការ "tribunal" and ខ្មែរក្រហម "Khmer Rough". We know that the synonym of តុលាការ "tribunal" is សាលាក្តី "tribunal". So the expanded query for our proposed system is "តុលាការខ្មែរក្រហម OR សាលាក្តីខ្មែរក្រហម". On the other hand, the expanded query for Google is "[តុលាការ OR សាលាក្តី] AND ខ្មែរក្រហម".

5.3 Results and Discussion

Table 2 and 3 show the results of precision, recall and F-Measure before and after implementing each QE techniques to the proposed system and Google web search engine. The improvement in recall of our proposed system is 16.38%, 17.56%, 11.31%, 17.93% after applying the respective QE techniques, while the increase in recall at 6.32%, 19.00%, 4.37%, 5.97% after applying the QE techniques respectively to Google web search engine.

In addition, comparing our proposed system with QE to Google web search engine without QE, the recall improvement is 17.29%, 24.60%, 6.34%, 14.37% respectively, while to Google web search engine with QE, the recall improvement is 10.97%, 5.60%, 1.97%, 8.40% respectively.

As a summary, the search results using our proposed system with QE techniques is significantly better than the conventional Google search results. This can be seen clearly from the improvement of F-Measure 21.03%, 17.00%, 5.15%, 10.09% respectively.

6 Conclusion and Future Works

In this research, we have investigated four types of QE technique based on Khmer linguistic characteristics. These QE techniques are specifically designed to improve the retrieval performance of Khmer IR System. The experiments have demonstrated the improvement of retrieval performance of the proposed system and the Google web search engine after applying the proposed QE techniques. However, the improvement in precision after utilizing our proposed QE techniques is not so significant. In the case of derivative words, it even shows some slight decrement. This is one of the

[8] http://www.google.com/support/websearch/bin/answer.py?answer=136861

main problems that we will tackle in our future research to reduce non-relevant contents by semantically analyzing Khmer content. At the moment the size of the corpus is very small, and we are actively dealing with this issue in hope to provide a good Khmer language resource for the future research in Khmer language processing.

In addition, due to the lack of research in Khmer IR as well as in Khmer language processing, a lot of aspects can still be worked on in order to improve the system performance. For example, the improvement of Khmer word segmentation and the building of Khmer thesaurus for IR system, which are expected to improve the IR system performance, are also in the priority tasks of our future works.

References

Beaulieu, Micheline and Susan Jones. 1998. Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. *Interacting With Computers*, 10(3):237--248.

Buddhist Institute. 1967. វចនានុក្រមខ្មែរ "*Khmer Dictionary*". Buddhist Institute, Phnom Penh, Cambodia.

Chea, Sok-Huor, Rithy Top, Pich-Hemy Ros, Navy Vann, Chanthirith Chin and Tola Chhoeun. 2007. Word Bigram Vs Orthographic Syllable Bigram in Khmer Word Segmentation. *http://www.panl10n.net/english/OutputsCambodia1.htm*, (Last retrieved 30 April 2010).

Hatcher, Erik, Otis Gospodnetić and Michael McCandless. 2009. *Lucene in Action, Second Edition*. Manning Publications, Connecticut, USA.

Ide, Nancy 1998. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In *Proceedings of the First International Language Resources and Evaluation Conference*, pp. 463-70.

Ide, Nancy, Patrice Bonhomme and Laurent Romary 2000. XCES: An XML-based standard for linguistic corpora. In *Proceeding of Second Language Resources and Evaluation Conference (LREC)*, pp. 825--830, Athens. Greece.

Jenner, Philip Norman. 1969. Affixation in modern Khmer. *A dissertation submitted to the graduate division*, University of Hawaii.

Khin, Sok. 2007. វេយ្យាករណ៍ភាសាខ្មែរ "*Khmer Grammar*". Royal Academy of Cambodia, Phnom Penh, Cambodia.

Long, Siem. 1999. បញ្ហាវេចនសព្ទវិទ្យាខ្មែរ "*Khmer Lexicology Problem*". National Language Institute, Royal University of Phnom Penh, Phnom Penh, Cambodia.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze 2009. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

Nou, Chenda and Wataru Kameyama 2007. Khmer POS Tagger: A Transformation-based Approach with Hybrid Unknown Word Handling. In *Proceeding of International Conference on Semantic Computing (ICSC)*, pp. 482--492. Irvine, USA.

Ourn, Noeurng and John Haiman. 2000. Symmetrical Compounds in Khmer. *Studies in Language*. 24(3), pp. 483--514.

Singhal, Amit. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35--43.