Building a Khmer Text Corpus

Channa VAN and Wataru KAMEYAMA
Graduate School of Information and Telecommunication, WASEDA University
{channa@fuji.waseda.jp, wataru@waseda.jp}

1 Introduction

Text corpus is very important for many fields of research particularly in Natural Language Processing (NLP). While most of the major languages already have its own text corpus, Khmer does not have one yet. The absence of Khmer text corpus has been an obstacle and a big issue for many researchers who have worked on Khmer NLP. This research is the first attempt to construct a text corpus for Khmer. In order that the corpus can be used widely and gives its benefit to the future research, we have focused on every aspect of the corpus design from encoding to annotation.

2 System Design and Implementation

Figure 1 shows the processing steps of building the proposed Khmer Text Corpus. There are four main stages in the system: text collection, preprocessing tasks, annotation and encoding. We begin by manually collecting Khmer Unicode texts from various Khmer websites and blogs on the Internet. Then, all texts are cleaned by removing the unnecessary features such as images and unwanted texts. After that, each text is assigned with its description in the labeling process.

The next step is the annotation which consists of sentence annotation, word annotation and Part-of-Speech (POS) annotation. Finally, the annotation information is encoded in XML which is conformed to the XCES [5] coding scheme.

2.1 Text Collection

Collecting Khmer digital texts is a challenging task. It is not possible to scan or extract Khmer texts from neither books nor other paper sources since there is no implementation of Khmer optical character recognition yet. Therefore, all texts in this corpus have been manually collected from websites and blogs.

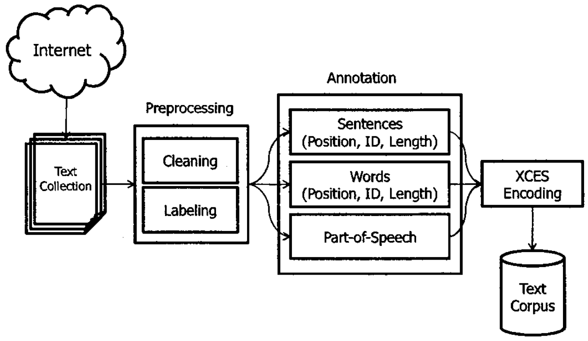

FIGURE 1: System Architecture

2.2 Preprocessing Tasks

The texts collected from the Internet usually are not clean and unstructured. It may contain unwanted elements such as images, links and 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 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.

2.3 Corpus Annotations

Corpus annotation information is very useful for the corpus-based research. Therefore, this corpus also includes the common annotations which are sentence annotation, word annotation and POS annotation.

2.3.1 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 last character of a sentence in a text.
2. Identification: the sequence number of a sentence within a text file.
3. Length: the number of characters of a sentence.

In order to annotate these information, all the sentences must be separated. A series of characters exist in Khmer which are used to mark the end of different kind of sentences [3]:

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

By using these characters as boundaries, each sentence can be separated. However, sentences which do not have any special characters at the ends exist. For example, the sentences which appear as title, section's name, subsection's name and etc. In these cases, a special character is manually put at the end of the sentence.

2.3.2 Word Annotation

Like sentence, the same information is also annotated for each word:

1. Position: it is defined by the position of the first character and last character of a word in a text.
2. Identification: the sequence number of a word within a text file.
3. Length: the number of characters of a word.

To obtain the above information of each word, all words must be separated. Khmer text is a non-segmented. There is no separator between words which makes them very difficult to segment. In order to do that, we have used Khmer Word Segmentation Tool [2] developed by the team of PAN Localization Cambodia [3]. Their work is based on statistical model: Word Bigram Model.

2.3.3 Part-of-Speech Annotation

To enchance the usefulness of the corpus, we also include Part-of-Speech annotation. We have used a Khmer POS Tagger which is based on the work of Nou et al. [4] where a transformation-based approach with hybrid unknown word handling for Khmer POS tagger is proposed. 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.

2.4 XCES Encoding

To assure the extensibility of the corpus for future work, this corpus is encoded in eXtensible Corpus Encoding Standard(XCES) [5]. XCES is an XML based standard to codify text corpus. It is highly based on previous Corpus Encoding Standard(CES) [6] but using XML as the markup language. It supports many types of corpora especially the annotation corpora. The encoding of annotation files and text description files are conformed to XCES's schemas version 1.0.4 [7].

3 Current Result

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

Domain	**# of Article**	**# of Sentence**	**# of Word**
Newspaper	667	14830	467109
Magazine	52	1335	41018
Medical	3	76	1950
Technology	15	607	15629
Culture	33	1178	41961
Law	43	5146	97299
History	9	276	7363
Agriculture	29	1484	28887
Essay	8	304	8001
Story	108	5642	189859
Novel	78	12012	221541
Other	5	134	5382
Corpus Total	**1050**	**43024**	**1125999**

TABLE 1: Current Result

4 Conclusion

This corpus is the first attempt to build a Khmer text corpus which includes different kinds of annotation information. It is still in the early stage compared to the other languages. There are a lot of things needed to be improved in the future. The corpus size is relatively small and there is still room for improvement for word segmentation efficiency. In addition, more works can be done to improve the POS tagger accuracy. These are the issues that we would like to tackle as our future works.

The building of this corpus is considered as a big contribution to the research of Khmer NLP. We are going to make it available to everyone via the Internet so that this corpus can be used for evaluation, comparison and testing in the future research of Khmer NLP. More research on Khmer NLP such as grammar checking, spell checking, information retrieval, machine translation, text to speech and etc. can fully take the advantage of this research.

References

[1] Khin, S. Khmer Grammar First Edition 2007 (In Khmer). Royal Academy of Cambodia.

[2] Chea, S., Top , R. Ros ,P. Word Bigram Vs Orthographic Syllable Bigram in Khmer Word Segmentation. http://www.panl10n.net/english/OutputsCambodia1.htm. (Last retrieved 7 January 2010).

[3] PAN Localization Cambodia. http://www.pancambodia.info. (Last retrieved 30 Demcember 2009).

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

[5] Ide, N., Bonhomme, P. and Romary, L. (2000) XCES: An XML-based standard for linguistic corpora. In *Proceeding of Second Language Resources and Evaluation Conference (LREC), pp 825–830*. Athens, Greece.

[6] Ide, Nancy, Greg Priest-Dorman, and Jean V´eronis (1996). Corpus Encoding Standard. http://www.cs.vassar.edu/CES/. (Last retrieved 30 December 2010).

[7] XCES Schemas. http://www.xces.org/schema/. (Last retrieved 30 December 2009).