Chapter XXIV
Word Segmentation in Indo-China Languages for Digital Libraries

Jin-Cheon Na
Nanyang Technological University, Singapore

Yin-Leng Theng
Nanyang Technological University, Singapore

Tun Thura Thet
Nanyang Technological University, Singapore

Schubert Foo
Nanyang Technological University, Singapore

Dion Hoe-Lian Goh
Nanyang Technological University, Singapore

Paul Horng-Jyh Wu
Nanyang Technological University, Singapore

ABSTRACT

This chapter introduces word segmentation methods for Indo-China languages. It describes six different word segmentation methods developed for the Thai, Vietnamese, and Myanmar languages and compare different approaches in terms of their algorithms and results achieved. The discussion and comparison of these word segmentation methods will provide underlying views about how word segmentation can be achieved and employed in Indo-China languages to support search functionality in digital libraries.

INTRODUCTION

Digital libraries are not really digitized libraries (Witten & Bainbridge, 2003). They essentially changes the way information is used in the world. A digital library is about new ways of dealing with knowledge. One of the many advantages digital libraries have over traditional libraries is the ability to search information efficiently and effectively. Users can conduct various searches, ranging from a simple title search query to a complex advanced query. It is pretty straightforward for documents written in well established western languages such as English but when it comes to languages

such as Thai, Vietnamese, and Myanmar, it can be significantly challenging, mainly due to the nontrivial task of segmenting words. Word segmentation is an essential preprocess for the full-text indexing of documents written in these languages in order to support search functionality in digital libraries.

Unlike English, there are no white spaces between words in these languages. Dissimilar to Chinese, a syllable can be composed of multiple characters where a word can contain multiple syllables. It seems to be quite effortlessly easy for a native speaker to determine where word boundaries are in a sentence or a document. For the very same reason, people expect computers to determine the word boundaries automatically. Unless the issue of word segmentation is addressed properly, the indexing of terms for search functions in digital libraries will not be feasible. From the information processing perspective, it is important to index and search words in documents' contents rather than just metadata, such as titles and descriptions. In addition, for a document in one language to be translated into another language, the first step is always to segment words before doing any further processing. Therefore, word segmentation is basic yet essential in order to carry out any further information processing for documents written in these languages.

The following sections discuss word segmentation methods for the Thai language, the Vietnamese language, and the Myanmar language, and the comparison of them. The last section summarizes the current status and discusses future work.

WORD SEGMENTATION

The Thai Language

The Thai script is a member of the Indic family of scripts, descended from Brahmi. In the Thai language, a "word" is difficult to define, as it does not exhibit explicit word boundaries. Like many other Asian languages, the Thai language does not use white spaces for word boundaries. Each Thai letter is a consonant possessing an inherent vowel sound as well as inherent tones. Both the inherent vowel and tone can be modified by means of vowel signs and tone marks attached to the base consonant letter. All of the tone marks and some of the vowel signs are rendered in the script as diacritics attached above or below the base consonant. In the Unicode memory representation, these combining signs and marks are encoded after the modified consonant (Unicode Consortium, 2004). One main cause of the problems in Thai word segmentation is the lack of a clear definition of a Thai "word" (Wirot, 2002). Traditional methods of Thai word segmentation are based on unclear criteria and procedures, and have several limitations. Most of the word segmentation approaches use a dictionary for segmenting running texts.

A study conducted by Sornlertlamvanich, Potipiti, and Charoenporn (2000) used automatic corpus-based word extraction. It employed the C4.5 decision tree induction program (Quinlan, 1993) as a learning algorithm for word extraction. The induction algorithm evaluates the content of a series of attributes and interactively builds a tree. The leaves of the decision tree represent the values of the goal attributes. The method used C4.5 to prune the entire decision tree in order to reduce the effect of over fitting. It recursively traveled to each subtree to determine if the leaf or branch could reduce the expected error rate. The attributes of the learning algorithm are mutual information, entropy, frequency, and string length. Evaluation of the method was carried out with a corpus of size 1 MB, consisting of 75 articles from various fields. Thirty thousand strings were manually tagged and compared with the results produced by the method, which recorded a 84.1% accuracy for the test dataset.

Another study conducted by Wirot (2002) used a two-part approach: a syllable-based trigram model for syllable segmentation, and maximum