# Thai Syllable-Based Information Extraction Using Hidden Markov Models

Lalita Narupiyakul[1], Calvin Thomas[1], Nick Cercone[1], and Booncharoen Sirinaovakul[2]

[1] Faculty of Computer Science, Dalhousie University 6050 University Avenue, Halifax, NS, CANADA B3H 1W5
{lalita, thomas,nick}@cs.dal.ca
[2] King Mongkut's University of Technology Thonburi 91 Pracha Uthit, Thungkru, Bangkok, THAILAND 10140
{boon@cpe.kmutt.ac.th}

**Abstract.** Information Extraction (IE) is a method which analyzes the information and retrieves significant segments or fields for insertion into tables or databases by automatic extraction. In this paper, we employ a statistical model for an IE system. Thai syllable-based information extraction using Hidden Markov Models (HMM) is our proposed method for automated information extraction. In our system, we develop a non-dictionary based method which requires a rule-based system for syllable segmentation. We employ a Viterbi algorithm, which is a statistical system for learning/testing our corpus, and extract the required fields from the information in corpus.

## 1 Introduction

In electronic communication systems, there are many ways to communicate or send information to others using hi-tech digital devices. Sending printed information via telephone lines is a basic method of communication. Electronic mail and webboard are the most popular communication tools. They are used to send various kinds of information to people such as news, advertisements and announcements. These information are broadcasted to subscribers. A tremendous amount of information is sent via electronic documents directly to people. Tools to manipulate this information are desirable. Therefore, information extraction (IE) is one of various methods proposed to analyze information and retrieve significant segments or fields for insertion into tables or databases by automatic extraction.

Developing IE in Thai is at a preliminary stage and there are a small number of researchers working on Thai IE [1],[2]. For example, Sukhahuta [1] develops information extraction strategies for Thai documents. His work is based on natural language techniques (grammar parsing, syntactic and concept analysis) to extract information from a prepared template corpus. Information extraction in Thai is difficult because Thai is distinguished from other languages including reading and writing structures. A characteristic of Thai written structures