Plangsarn: Thai Romanization Library Application

Srichan Chancheewa, THAMMASAT UNIVERSITY
Dr. Choochart Haruechayiyasak, NECTEC, THAILAND
Why do we need this application?

before

after
• Romanization is the representation of a written word or spoken speech with the Roman (Latin) script.

• Two main methods of romanization are

  (1) Transliteration: for representing written text,

  (2) Transcription: for representing the spoken word,
      (2.1) phonemic transcription: records the phonemes or units of semantic meaning in speech
      (2.2) phonetic transcription: records speech sounds with precision.
Background and Motivation

• Phonetic transcription attempts to depict all phones in the source language.

• We adopt the International Phonetic Alphabet (IPA).

• IPA is an alphabetic system of phonetic notation based primarily on the Latin alphabet.

• IPA was devised by the International Phonetic Association as a standardized representation of the sounds of oral language.

• IPA symbols are composed of one or more elements of two basic types, letters and diacritics.
Background and Motivation

• For library cataloging, there is a guideline for preparing romanization:
  
  — ALA-LC Romanization Tables: Transliteration Schemes for Non-Roman Scripts, which is approved by the Library of Congress and the American Library Association.

  — ALA-LC is a set of standards for romanization, or the representation of text in other writing systems using the Latin alphabet.
Objective

• Design and implement a system for automatically romanizing Thai texts with the focus on bibliographic data.

• The system is intended to support librarians who perform the cataloguing task.

• The system is called Plangsarn (แปลงสาส์น).
  — plang (แปลง) = transform
  — sarn (สาส์น) = text, message
Plang'sarn’s System Architecture

- Lexicon contains Thai word list with romanization.
- G2P (Grapheme-to-Phoneme) converts grapheme unit into phoneme unit.
- Tokenizer performs the word segmentation on Thai written texts.
1. Get Thai text input from the user via browser interface.
2. Perform word segmentation on the input text.
3. For each word token, look up in the lexicon for existing romanized representation.
4. If the romanized representation already exists, then use it. Otherwise send the word to G2P module.
5. Concatenate all romanized representations and send the final result to the user via the browser interface.
Example of Plangsarn process

Input Thai text: การเดินทางไปมหาวิทยาลัย

Tokenized text: การเดินทาง ไป มหาวิทยาลัย

Romanized text: kāndoēnthāŋ pai mahāwitthayālai
Plangsarn’s Web interface:
http://164.115.23.167/plangsarn/

ข้อความไทย
การเดินทางไปมหาวิทยาลัย

ผลลัพธ์
kāndoenthāng pai mahāwitthayālai

ผลลัพธ์การแปล
Travelling to the University
User Evaluation

• Plangsarn is under the process of user testing and evaluation.

• Two main reported problems so far are
  — Incorrect word segmented unit.
    • ห้องครัว (Kitchen) => ห้อง khrūa (should be ห้อง khrūa)
  — Unable to recognize and capitalize named entities such as the names of person, organization, place
    • กรุงเทพ (Bangkok) => krung thēp (should be Krungthēp)

• To solve the problem, the system allows authorized users to manually include words with its correct romanization into the lexicon.
Plangsarn’s Summarized Features

• Automatic romanization of Thai text using the G2P (Grapheme-to-Phoneme) technique
• Use the lexicon-based Thai word segmentation to tokenize the text
• Allow authorized users to manually add and edit the list of romanized lexicon
• Support both IPA and OCLC Connexion
• Automatic capitalization for named entities (e.g., names of person, organization, place)
• Provide the machine translation of Thai to English
PLANGSARN IS NOW AVAILABLE WITHIN THAI GOVERNMENT CLOUD SERVICE

HTTP://164.115.23.167/PLANGSARN/
SPECIAL THANKS TO

Dr. Choochart Haruechayiasak
Senior Researcher
Speech and Audio Technology Laboratory (SPT)
National Electronics and Computer Technology Center (NECTEC)
112 Phahonyothin Rd., Klong 1, Klong Luang,
Pathumthani 12120
THAILAND

gmail: choochart.haruechayiasak@nectec.or.th
Tel. (+66) 2 564-6900 Ext. 2240
Fax. (+66) 2 564-6873