A Multi-Aspect Comparison and Evaluation on Thai Word Segmentation Programs

Chaluemwut Noyunsan[1], Choochart Haruechaiyasak[2] Seksan Poltree[3], and Kanda Runapongsa Saikeaw[1]**

[1] Department of Computer Engineering, Faculty of Engineering,
Khon Kaen University 123 Mittrapap, Mueang, Khon Kaen 40002, Thailand
chaluemwut@kkumail.com, krunapon@kku.ac.th
[2] Human Language Technology Laboratory (HLT)
National Electronics and Computer Technology Center (NECTEC),
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand
choochart.haruechaiyasak@nectec.or.th
[3] Morange Solution Company Limited,
456/1 Klang Mueang, Mueang, Khon Kaen 40000, Thailand
seksan@morange.co.th

Abstract. Word segmentation is an important task in natural language processing, especially for languages without word boundaries, such as Thai language. Many Thai word segmentation programs have been developed. Researchers and developers in Thai documents usually spend a tremendous amount of time in studying and trying different Thai word segmentation programs. This paper presents the performance of six Thai word segmentation programs which include Libthai, Swath, Wordcut, CRF++, Thaisemantics, and Tlexs. Based on experimental results, we compare these programs in terms of usage, response time, time outs, and relevance.

Keywords: Word segmentation, term tokenization, software tools, natural language processing

1 Introduction

Natural Language Processing (NLP) enables computers to understand human languages. It consists of many processes such as word segmentation, part-of-speech tagging, automatic summarization, and speech synthesis. Most NLP applications require input text to be segmented into words before being processed

** Corresponding author

further. For example, in sentences similarity application, text must first be tokenized into a series of terms before being analyzed grammatically and semantically. Word segmentation is an essential part for Asian languages such as Thai, Chinese, Japanese, and Korean. This is because these languages are written without delimiter spaces for words in the same sentence.

2 Thai Word Segmentation Programs

Many researches and several programs have been developed for word segmentation. We chose six Thai word segmentation programs which were Libthai, Swath, Wordcut, CRF++, Thaisemantics, and Tlexs. They were selected because they were actively maintained and widely used. Libthai [1] is a set of C programming function to support Thai word segmentation. Swath [7] and Wordcut [6] are command-line programs. CRF++ is a tool for supporting Condiction Random Field (CRF) [2]. Thaisemantics has been developed by using Restful web service [5]. Tlexs uses CRF to train models for segmentation [4].

3 Experimental Analysis

Fig. 1 shows the system overview. Our system used the Benchmark for Enhancing the Standard of Thai language processing (BEST) corpus which has been developed by NECTEC [3] and widely accepted for Thai language processing. Our system sent an origin message to the selected Thai word segmentation programs, and then received the results from each program. Then the results from each system were compared with manually tagged words in BEST.

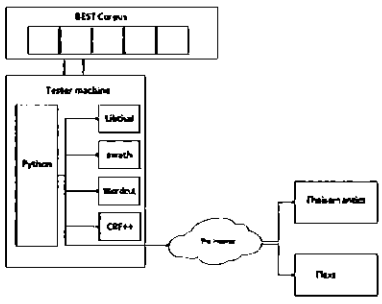

Fig. 1. System overview

Usage Table 1 summarizes the features about usage, offline support, and whether installation is needed.

Table 1. Comparison of word segmentation programs in terms of usage

Features	Libthai	Swath	Wordcut	Thaisemantics	Tlexs	CRF++
Usage	C function	Wrapper	Wrapper	REST API	SOAP API	Wrapper
Offline support	Yes	Yes	Yes	No	No	Yes
Installation needed	Yes	Yes	Yes	No	No	Yes

Response Time We ran the experiments and measured response times on the computer with Intel Dual 1.73GHz and 3 GB of RAM using Ubuntu 64bit as an operating system. The program execution times are shown in Table 2. Libthai performed the best with the response time 0.022 seconds while Swath and CRF++ had response times as 0.024 seconds. On the other hand, the responses times of Thaisemantics, Wordcut, and Tlexs were large. This is because Thaisemantics and Tlexs are programs that are used over the internet thus their response times depend on the internet bandwidth and traffic time. Wordcut is implemented using JavaScript language which may increase the response time of the program.

Table 2. Comparison of word segmentation program response times (in seconds)

Libthai	Swath	Wordcut	Thaisemantics	Tlexs	CRF++
0.022	0.024	0.203	1.520	0.144	0.024

Number of Time Outs Tlexs and Thaisemantics had several time outs because they are called over the internet. Tlexs had average 2,070 time outs while Thaisemantics had 5 time outs. Tlexs caused a large number of occurrences of time outs might be due to server settings or errors.

Relevance Messages from the BEST were sent to each program and the word segmentation output was kept in a list. After that, we compared this log list with a correct list from BEST by using the percentage of precision, recall and F-measure. We ran this test by using 5-fold cross-validation, and then computed

the average value as shown in Table 3. Both Tlexs and CRF++ have the best F-measure because they use CRF.

Table 3. F-measure of Thai word segmentation programs

Measurement	Libthai	Swath	Wordcut	CRF++	Thaisemantics	Tlexs
Precision	61.23	65.09	57.27	59.91	66.03	74.80
Recall	54.97	55.96	62.05	67.80	60.58	75.88
F-measure	**57.61**	**59.60**	**59.30**	**63.14**	**63.03**	**75.26**

4 Conclusions

This paper presents the comparison of six Thai word segment programs in terms of usage, response time, time outs, and relevance. Swath, Libthai, and CRF++ programs provide the smallest response times because they are native programs. Thaisemantics yields the largest response time because Thaisemantics is called over the internet and uses a dictionary. Although Tlexs is also called over the internet, it has better response time because it uses CRF. Both Tlexs and CRF++ give the highest F-measure because they employ CRF.

References

1. T. Karoonboonyanan, C. Silpa-Anan, P. Kiatisevi, P.Veerathanabutr and V. Ampornaramveth, *Libthai Library*, retrieved on Jul 1, 2014 from `http://linux.thai.net/projects/libthai`.
2. T. Kudo, *CRTF++: Yet Another CFT toolkit* retrieved on July 2, 2014 from `http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar`.
3. National Electronics and Computer Technology Center (NECTEC), *BEST: Bencharmk for Enhancing the Standard for Thai Language Processing* retrived on Jul 5, 2014 from `http://thailang.nectec.or.th/best/`.
4. National Electronics and Computer Technology Center (NECTEC), *TLexs: Thai Lexeme Analyser*, retrieved on Jul 3, 2014 from `http://sansarn.com/tlex/`.
5. S. Poltree, *Thaisemantics: Free Thai Language Resources and Services*, retrieved on Jul 2, 2014 from `http://www.thaisemantics.org/`.
6. V. Satayamas, *wordcut program* retrieved on Jul 3, 2014 from `https://github.com/veer66/wordcut`.
7. P. Charoenpornsawat, *SWATH - Thai Word Segmentation*, retrieved on Jul 1, 2014 from `http://www.cs.cmu.edu/~paisarn/software.html`.