The State of the Art in Thai Language Processing

Virach Sornlertlamvanich, Tanapong Potipiti, Chai Wutiwiwatchai and Pradit Mittrapiyanuruk
National Electronics and Computer Technology Center (NECTEC),
National Science and Technology Development Agency, Ministry of Science and Technology Environment.
22nd Floor Gypsum Metropolitan Tower 539/2 Sriayudhya Rd. Rajthevi Bangkok 10400 Thailand.
Email: {virach, tanapong, chai}@nectec.or.th, pmittrap@notes.nectec.or.th

Abstract

This paper reviews the current state of technology and research progress in the Thai language processing. It resumes the characteristics of the Thai language and the approaches to overcome the difficulties in each processing task.

1 Some Problematic Issues in the Thai Processing

It is obvious that the most fundamental semantic unit in a language is the *word*. Words are explicitly identified in those languages with word boundaries. In Thai, there is no word boundary. Thai words are implicitly recognized and in many cases, they depend on the individual judgement. This causes a lot of difficulties in the Thai language processing. To illustrate the problem, we employed a classic English example.

The segmentation of "**GODISNOWHERE**".

No.	Segmentation	Meaning
(1)	God is now here.	God is here.
(2)	God is no where.	God doesn't exist.
(3)	God is nowhere.	God doesn't exist.

With the different segmentations, (1) and (2) have absolutely opposite meanings. (2) and (3) are ambiguous that *nowhere* is one word or two words. And the difficulty becomes greatly aggravated when unknown words exist.

As a tonal language, a phoneme with different tone has different meaning. Many unique approaches are introduced for both the tone generation in speech synthesis research and tone recognition in speech recognition research.

These difficulties propagate to many levels in the language processing area such as lexical acquisition, information retrieval, machine translation, speech processing, etc. Furthermore the similar problem also occurs in the levels of sentence and paragraph.

2 Word and Sentence Segmentation

The first and most obvious problem to attack is the problem of word identification and segmentation. For the most part, the Thai language processing relies on manually created dictionaries, which have inconsistencies in defining word units and limitation in the quantity. [1] proposed a word extraction algorithm employing C4.5 with some string features such as entropy and mutual information. They reported a result of 85% in precision and 50% in recall measures. For word segmentation, the longest matching, maximal matching and probabilistic segmentation had been applied in the early research [2], [3]. However, these approaches have some limitations in dealing with unknown words. More advanced techniques of word segmentation captured many language features such as context words, parts of speech, collocations and semantics [4], [5]. These reported about 95-99 % of accuracy. For sentence segmentation, the trigram model was adopted and yielded 85% of accuracy [6].

3 Machine Translation

Currently, there is only one machine translation system available to the public, called ParSit (http://www.links.nectec.or.th/services/parsit), it is a service of English-to-Thai webpage translation. ParSiT is a collaborative work of NECTEC, Thailand and NEC, Japan. This system is based on an interlingual approach MT and the translation accuracy is about 80%. Other approaches such as generate-and-repair [7] and sentence pattern mapping have been also studied [8].

4 Language Resources

The only Thai text corpus available for research use is the ORCHID corpus. ORCHID is a 9-MB Thai part-of-speech tagged corpus initiated by NECTEC, Thailand and Communications Research Laboratory, Japan. ORCHID is available at http://www.links.nectec.or.th /orchid.

5 Research in Thai OCR

Frequently used Thai characters are about 80 characters, including alphabets, vowels, tone marks, special marks, and numerals. Thai writing are in 4 levels, without spaces between words, and the problem of similarity among many patterns has made research challenging. Moreover, the use of English and Thai in general Thai text creates many more patterns which must be recognized by OCR.

For more than 10 years, there has been a considerable growth in Thai OCR research, especially for “printed character” task. The early proposed approaches focused on structural matching and tended towards neural-network-based algorithms with input for some special characteristics of Thai characters e.g., curves, heads of characters, and placements. At least 3 commercial products have been launched including “ArnThai” by NECTEC, which claims to achieve 95% recognition performance on clean input. Recent technical improvement of ArnThai has been reported in [9]. Recently, focus has been changed to develop system that are more robust with any unclean scanning input. The approach of using more efficient features, fuzzy algorithms, and document analysis is required in this step.

At the same time, “Offline Thai handwritten character recognition” task has been investigated but is only in the research phase of isolated characters. Almost all proposed engines were neural network-based with several styles of input features [10], [11]. There has been a small amount of research on “Online handwritten character recognition”. One attempt was proposed by [12], which was also neural network-based with chain code input.

6 Thai Speech Technology

Regarding speech, Thai, like Chinese, is a tonal language. The tonal perception is important to the meaning of the speech. The research currently being done in speech technology can be divided into 3 major fields: (1) speech analysis, (2) speech recognition and (3) speech synthesis. Most of the research in (1) done by the linguists are on the basic study of Thai phonetics e.g. [13].

In speech recognition, most of the current research [14] focus on the recognition of isolated words. To develop continuous speech recognition, a large-scale speech corpus is needed. The status of practical research on continuous speech recognition is in its initial step with at least one published paper [15]. In contrast to western speech recognition, topics specifying tonal languages or tone recognition have been deeply researched as seen in many papers e.g., [16].

For text-to-speech synthesis, processing the idiosyncrasy of Thai text and handling the tones interplaying with intonation are the topics that make the TTS algorithm for the Thai language differrent from others. In the research, the first successful system was accomplished by [14] and later by NECTEC [15]. Both systems employ the same synthesis technique based on the concatenation of demisyllable inventory units.

References

[1] V. Sornlertlamvanich, T. Potipiti and T. Charoenporn. *Automatic Corpus-Based Thai Word Extraction with the C4.5 Learning Algorithm.* In forthcoming Proceedings of COLING 2000.

[2] V. Sornlertlamvanich. *Word Segmentation for Thai in Machine Translation System Machine Translation.* National Electronics and Computer Technology Center, Bangkok. pp. 50-56, 1993. (in Thai).

[3] A. Kawtrakul, S. Kumtanode, T. Jamjunya and A. Jewriyavech. *Lexibase Model for Writing Production Assistant System.* In Proceedings of the Symposium on Natural Language Processing in Thailand, 1995.

[4] S. Meknavin, P. Charoenpornsawat and B. Kijsirikul. *Featured Based Thai Word Segmentation.* In Proceedings of Natural Language Processing Pacific Rim Symposium, pp. 41-46, 1997.

[5] A. Kawtrakul, C. Thumkanon, P. Varasarai and M. Suktarachan. *Autmatic Thai Unknown Word Recognition.* In Proceedings of Natural Language Processing Pacific Rim Symposium, pp. 341-347, 1997.

[6] P. Mitrapiyanurak and V. Sornlertlamvanich. *The Automatic Thai Sentence Extraction.* In Proceedings of the Fourth Symposium on Natural Language Processing, pp. 23-28, May 2000.

[7] K. Naruedomkul and N. Cercone. *Generate and Repair Machine Translation.* In Proceedings of the Fourth Symposium on Natural Language Processing, pp. 63-79, May 2000.

[8] K. Chancharoen and B. Sirinaowakul. *English Thai Machine Translation Using Sentence Pattern Mapping.* In Proceedings of the Fourth Symposium on Natural Language Processing, pp. 29-36, May 2000.

[9] C. Tanprasert and T. Koanantakool. *Thai OCR: A Neural Network Application*. In Proceedings of IEEE Region Ten Conference, vol.1, pp.90-95, November 1996.

[10] I. Methasate, S. Jitapankul, K. Kiratiratanaphung and W. Unsiam. *Fuzzy Feature Extraction for Thai Handwritten Character Recognition.* In Proceedings of the Forth Symposium on Natural Language Processing, pp.136-141, May 2000.

[11] P. Phokharatkul and C. Kimpan. *Handwritten Thai Character Recognition using Fourior Descriptors and Genetic Neural Networks.* In Proceedings of the Fourth Symposium on Natural Language Processing, pp.108-123, May 2000.

[12] S. Madarasmi and P. Lekhachaiworakul. *Customizable Online Thai-English Handwriting Recognition.* In Proceedings of the Forth Symposium on Natural Language Processing, pp.142-153, May 2000.

[13] J. T. Gandour, S. Potisuk and S. Dechongkit. Tonal Coarticulation in Thai, Journal of Phonetics, vol 22, pp.477-492, 1994.

[14] S. Luksaneeyanawin, et al. *A Thai Text-to-Speech System*. In Proceedings of Fourth NECTEC Conference, pp.65-78, 1992. (in Thai).

[15] P. Mittrapiyanuruk, C. Hansakunbuntheung, V. Tesprasit and V. Sornlertlamvanich. *Improving Naturalness of Thai Text-to-Speech Synthesis by Prosodic Rule*. In forthcoming Proceedings of ICSLP2000.

[16] S. Jitapunkul, S. Luksaneeyanawin, V. Ahkuputra, C. Wutiwiwatchai. *Recent Advances of Thai Speech Recognition in Thailand*. In Proceedings of IEEE Asia-Pacific conference on Circuits and Systems, pp.173-176, 1998.