Open Collaborative Development of the Thai Language Resources for Natural Language Processing

Thatsanee Charoenporn[1], Virach Sornlertlamvanich[1], Sawit Kasuriya[2], Chatchawarn Hansakunbuntheung[2], and Hitoshi Isahara[1]

[1]Thai Computational Linguistics Laboratory, National Institute of Information and Communications Technology
[2]National Electronics and Computer Technology Center
Thailand Science Park, 112 Phahonyothin Rd., Klong 1, Klong Luang, Pathumthani 12120
{thatsanee@crl-asia.org, virach@crl-asia.org, sawitk@nectec.or.th, chatchawarnh@nectec.or.th, isahara@crl.go.jp}

Abstract

Language Resources are recognized as an essential component in linguistic infrastructure and a starting point of Natural Language Processing systems and applications. In this paper, we describe the achievement of the development and the use of Thai Language Resources germinated with an open collaboration platform, under the collaboration between research institutes. The resources include either text or speech. Text resources are divided into lexicon database and annotated corpus. We started developing a corpus-based Thai-English lexicon database (LEXiTRON) since 1994. It was originated from a dictionary designed for using in developing a machine translation system. Since then the Thai POS was designed and evaluated in several applications (word segmentation, machine translation, grapheme-to-phoneme, etc.) Extending the lexicon database, POS tagged corpus (ORCHID), and speech corpora for both synthesis and recognition are developed and functioned as an important part of research and development on NLP or HLT. These language resources are available for academic experiment.

1. Introduction

A large corpus plays its very essential role when stochastic and learning approaches on NLP come to their ages. Many research units put great efforts on developing the corpus for their particular purpose. But a large and complete corpus consumes a lot of man-power, time and budget. Collaboration, therefore, is established for the prompt requirement of the corpus. ORCHID, a Thai POS-tagged corpus and NECTEC-ATR Speech corpus are the concrete examples of the successfully collaborative projects related to Thai language resources.

The details of the resources are organized into 2 sections. The development of the text corpus, including lexicon database and annotated corpus will be described in Section 2. This section includes the steps of development, and design of LEXiTRON, an opened-lexicon database and the POS tagged corpus. Section 3 describes the development of speech corpus using several tools to automate the annotation process.

2. Text Corpus

Text corpus is the collection of a large database in the text level, which includes lexicon database, and annotated text. Generally, almost every research laboratory designed and constructed their own corpora according to their main project, and reused among the laboratory research fellows. Till 1996, Communications Research Laboratory, Japan, and National Electronics and Computer Technology Center, Thailand, initiated a collaborative project on building a Part-of-Speech tagged corpus. It provoked much collaboration on shared language resource construction in Thailand.

2.1 Lexicon database

We started the project of building a lexicon database, LEXiTRON, since 1994. LEXiTRON is the first Thai-English corpus-based lexicon database. The words are defined by a set of sample sentences and the usages in addition to their basic information of part-of-speech, classifier, verb pattern, synonym, antonym, and pronunciation (Palingoon, 2002). It was aimed to be a "dictionary" for writing. Most of the lexicons are originated from the database developed for Machine Translation project (the research and development of Multi-lingual Machine Translation System for Asian countries, 1987-1997). It then includes the information and word entry that are suitable for both human and machine use. The first version of LEXiTRON was launched in 1996 as a CD-ROM dictionary for human use. Recently, after a concentrated revision, the second version was launched under the Open Source concept for the contents. It is available in both stand-alone and on-line versions in http://lexitron.nectec.or.th/. The on-line version provides a free service for both English-to-Thai and Thai-to-English word lookup, and also a link to the wave file for Thai pronunciation, generated by our Thai text to speech system.

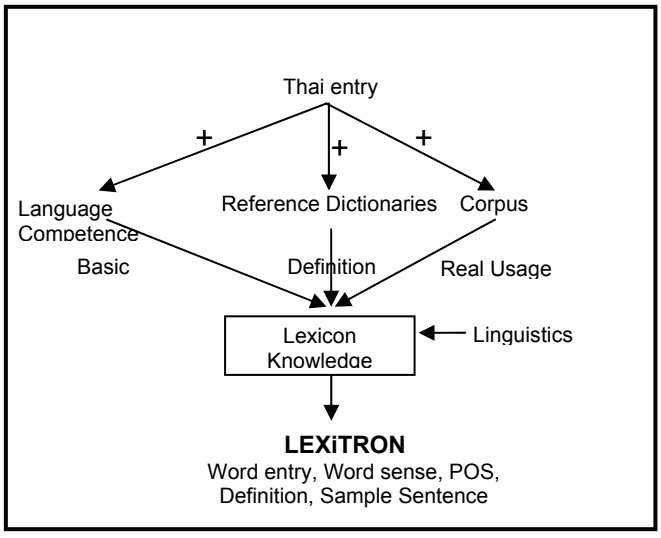

Figure 1 : A model of LEXiTRON Construction

LEXiTRON is a Thai-English corpus-based lexicon database. The finest unit is "word" which is extracted from large corpora according to the frequency of its occurrence. It contains 53,000 English entries and 35,000

Thai entries. Each entry or word is assigned its linguistic information as the followings:

- English-to-Thai: English entry, Thai equivalent word, pronunciation, head word, index for searching, synonym, antonym, Thai equivalent meaning, English sample sentence.
- Thai-to-English: Thai entry, English equivalent word, synonym, antonym, Thai meaning, classifier, and Thai sample sentence.

It is noted that the linguistic information (fields) in English-to-Thai and Thai-to-English lexicon databases differ in terms of the language features. However, they are both stored in the same designed format and linked to share the common information.

Figure 1 shows a model of LEXiTRON construction. It represents the linguistic knowledge attached to each word compiled from the reference document (dictionary), usage phenomena, and linguist competence.

In 2003, Thai Computational Linguistics Laboratory (TCL) was established as a partnership of Computational Linguistics Group, National Institute of Information and Communications Technology, Japan. One of TCL's ongoing projects is TCL's Computational Lexicon. It is developed by reusing an existing lexical database for machine translation. It contains about 69,000 Thai entries. Addition to the existing information, TCL's Computational Lexicon provides logical and semantic constraints for representing all possible semantics expressively. The logical constraints are capable of dealing with the absence of relatedness of word meanings. The semantic constraints are for discovering preferences of syntactic arguments of thematic roles. Table 1 indicates the entire constraints with their descriptions.

Logical Constraints	
Is-a (ISA)	a conceptual class of a given word
Equal (EQU)	a word that has the same or similar meaning of a given word
Not-equal (NEQ)	a word that has the opposite meaning of a given word
Part-of (POF)	a word that specifies a part of a given word
Whole-of (WOF)	a word that refers to the whole of which a given word is a part
Semantic Constraints	
Agent (AGT)	an entity that initiates the action
Object (OBJ)	an entity that is affected by the action
Instrument (INS)	an entity that is used in the action
Location (LOC)	a position or place where an event occurs
Time (TIM)	a point or period of time when an event occurs

Table 1 : Logical and semantic constraints

Figure 2 shows an example of semantic structure of the verb จ่าย 'pay' in the frame representation.

จ่าย 'pay'	
Logical constraints	
ISA	GIVE
EQU	**จ่ายเงิน** 'spend'
NEQ	**รับ** 'get'
Semantic constraints	
AGT	PERSON, ORGANIZATION
OBJ	MONETARY

Figure 2 : Semantic structure of the verb จ่าย 'pay'

2.2 Annotated corpus

ORCHID is a part-of-speech tagged corpus developed under the collaboration between the Communications Research Laboratory in Japan and NECTEC with the technical support from the Electrotechnical Laboratory in Japan, since 1996 (Charoenporn, 1997).

The original texts are a collection of technical papers of the proceedings of the National Electronics and Computer Technology Center (NECTEC) annual conferences. Each article is annotated in three levels, namely the level of paragraph, sentence and word. Figure 3 displays the procedure in annotating ORCHID. The process of annotation is semi-automatic. Separating paragraph into sentences (sentence segmentation), and post editing are conducted manually, while word segmentation and POS tagging are done automatically (Sornlertlamvanich, 1999).

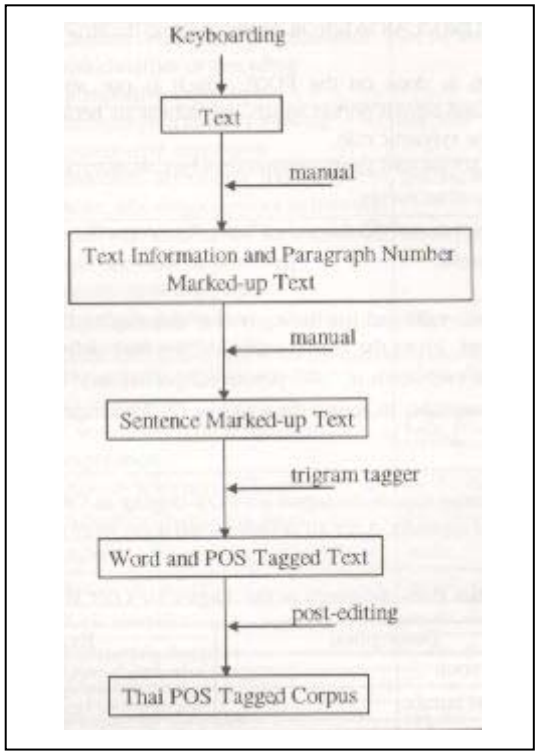

Figure 3 : Construction procedure of ORCHID

Each paragraph is manually tagged, from the input text into sentences. Each sentence in a tagged paragraph is then manually tagged with a delimiter. In the word level, word segmentation and POS-tagging processes are conducted automatically by a POS trigram word segmentation, SWATH. The POS set in ORCHID is the one that was designed for developing the MMT system. It consists of 14 categories with 47 subcategories. ORCHID is now available on http://links.nectec.or.th/orchid.

3. Speech Corpus

To develop corpus for using in the research of Thai speech recognition and synthesis, we called for collaboration from universities and research organizations to produce a large corpus. Language and acoustic models are needed in speech recognition while the models of prosody and continuous speech units are needed in natural

speech synthesis research. This speech corpus design is aimed to support both speech recognition and synthesis research. Currently there are 3 parallel projects running for NECTEC-ATR Thai Speech Database development, Thai Large vocabulary continuous speech Corpus (TLEC) development (Tarsaku, 2001; Thongprasirt, 2002; Sornlertlamvanich, 2001), and Thai Speech Corpus for Speech Synthesis (TSynC-1).

3.1 NECTEC-ATR Thai Speech Database (2001-2002) (Kanokprara, 2003; Kasuriya, 2003)

This project is the collaboration between NECTEC and ATR to develop a Thai dialogue speech corpus based on the hotel reservation task. The database consists of three sets, namely the isolated word set (DB1), the phonetic balanced sentences (DB2), and the hotel reservation dialogues (DB3).

DB1 (Isolated word set)
This set consists of three subsets. We divided the first subset into five minor subsets (D0 to D4). Each minor subset contains 1,000 words. The other two subsets are the phonetic balanced words and the extra words. The details of each subset creation are described in the followings.

- *5,000-words vocabulary subset (D0-D4):* We counted the frequency of each vocabulary from the text corpus (Thai magazines, journals, and encyclopedias) and the 5,000 most frequently used words were selected. Then they are randomly divided into five minor subsets.
- *PB word subset (D5):* Using the 5,000 words to select the PB words. All phonemes with at least amount of words and balanced occurrence were collected. Hence, the phoneme occurrences in this subset equal to the phoneme occurrences in the 5,000 words subset. A number of words in this subset are 640. Furthermore, this selection procedure is also used in PB sentence selection.
- *Extra word subset (D5):* The words that occurred in Hotel Reservation Transcription (HRT) set and did not occur in 5,000 vocabularies subset or PB word subset are called "Extra words". It contains 131 words.

DB2 (Phonetic balanced sentence set)
This set is the collection of the sentences that contains the whole set of Thai biphones. A large text corpus is required in order to extract the set of biphone. However, the current text corpus is not large enough to cover all Thai biphones. In this set, 390 phonetic balanced sentences are selected to cover all possible Thai phonemic units in the minimum number of the sentences.

DB3 (Hotel Reservation Transcriptions, HRT)
A hotel reservation system is one of well-known speech recognition application. Developing the corpus is the major problem in this application because there are multifarious speaking styles, several different dialogues, various types of hotel reservation procedure. The transcriptions used in this project are translated from the set 50 dialogues in HRT of the Spoken Language Translation Research Laboratories (SLT), Advanced Telecommunication Research International Institute (ATR), Kyoto, Japan. The dialogues have been translated to more than two languages such as English and Japanese. All utterances are recorded in quasi-quiet room. The qualities of them are around 20 dB. A number of speakers are 20 males and 20 females (18 to 40 years old).

3.2 Thai Large Vocabulary Continuous Speech Corpus (ORCHID-SPEECH CORPUS) (2001-)

ORCHID-SPEECH CORPUS is the collaborative project between NECTEC, Faculty of Electrical Engineering, Mahanakorn University of Technology (MUT), and Faculty of Computer Engineering, Prince of Songkhla University. The contents of this corpus consist of two sets, namely (1) the phonetically distributed (PD) sentence set and (2) a 5,000-word-vocabulary sentence set.

Attribute	PD set	TR set	DT set	ET set
No. of sentences	802	3,007	500	500
No. of vocabularies	2,269	5,000	1,622	1,630
No. of words	7,847	55,504	8,076	8,290

Table 2 : Summary of PD sentence set and 5,000-words vocabulary sentence set

The utterances are recorded in two environments: the clean speech environment (CS) and the office environment (OF). These environments are separated by the signal to noise ratio (SNR) Moreover, the SNR of CS and OF are around 30 dB and 20 dB respectively. All utterances are recorded according to reading styles. A number of speakers are 248 speakers (PSU: 100, MUT: 100, and NECTEC: 48).

Phonetically distributed sentence (PD) set
To initial acoustic model efficiently, phonetically balanced sentences (PB) are usually used for training. PB is the smallest set of sentences covering all phonemic units in the language. In our case, the phonemic unit is biphone. PD is the extension of PB. It does not only cover all biphone, but the text distribution is also similar to the daily used context (ORCHID corpus in this case).

The PD selection process starts from PB construction. In PB construction process, the sentence containing mostly unselected biphone is chosen one by one until all biphones are included in the PB set. Before constructing PD set, the biphone distributions of ORCHID are calculated. Then, some sentences are added to PB to change the distribution as same as distributions of ORCHID. The number of adding sentences should be kept at minimum while the biphone distribution of PD set is closest to ORCHID's distribution.

5,000 word vocabulary set
The objective of this set is to collect the structure of Thai language for language model (LM) construction. This set is divided into three subsets: the training set (TR), the development test set (DT), and the evaluation test set (ET). The TR set is used to train language models. The DT and ET sets are used for testing in development and evaluation phases respectively.

Firstly, the words of all sentences are listed and sorted. There are 43,255 vocabularies. The sentences containing the first 5,000 most frequently occurred vocabularies are selected. These sentences (11,202 sentences) are chosen to the next step. The TR set (3,007 sentences) is selected by collecting the minimum amount

of sentences that pertains 5,000 vocabularies. The remaining sentences are divided into two sets: set A and B, for language model construction (5,000 sentences) and DT, ET selection (3,195 sentences), consecutively. In addition, the set B is selected by calculating the sentence scores of each sentences and choosing the 3,195 sentences that are the highest sentence scores. On the other set, the tri-gram language model is created by 5,000 sentences and 3,007 sentences (TR set). There are 8,007 sentences that use for LM construction. And LM is used for calculating the perplexity of each sentence in set B. In the next procedure, the 1,000 sentences that have the medium perplexity (around 100 to 300), are selected and randomly divided into DT and ET sets.

3.3 NECTEC's Thai Speech Corpus for Speech Synthesis (TSynC-1) (Hansakunbuntheung, 2003)

The aims of this corpus are (1) to construct a chunk of speech unit candidates for developing a unit selection speech synthesis system and (2) to build a speech corpus with linguistic tags and acoustic information for conducting research on Thai reading-style prosodic model.

1. Specification

- 5,200 sentences collected from ORCHID Thai POS tagged corpus
- Text corpus in multilevel XML format: document, paragraph, sentence and word levels
- Linguistic tag and acoustic information: POS, Tone, Phone boundary, Syllable boundary, Word boundary, Phrase boundary, syllable position in word and phrase, Energy, F0, Duration, Voiced/unvoiced region, tone/toneless region
- Covering of tri-phone, tri-vowel and tri-tone unit
- Covering entire Thai and foreign (loaned) phones
- 14-hour speech sound of one female voice with standard Thai accent
- Recording environment:
  - 44kHz sampling rate, 16 bit per samples
  - SNR > 46 dB (silent room)
  - Using DAT

2. Development Procedure

The procedure in developing the corpus is divided into two parts as the followings:

2.1 Text selection

- Convert grapheme to phoneme transcription using Probabilistic GLR parser
- Define tri-phone, tri-vowel and tri-tone unit
- Score each unit using occurring probabilities
- Select sentences that cover the defined unit using greedy algorithm

2.2 Speech Tagging

- Use automatic phone segmentation tool based on HTK
- Revise the phone marks and insert phrase marks by linguists
- Mark syllable boundaries automatically using phone labels
- Locate syllable position in words and phrases using phone labels, word marks and phrase marks
- Mark voiced/unvoiced region
- Mark tone/toneless region using voiced/unvoiced marks and phone labels
- Extract energy and F0 curve

4. Conclusion

For years, we have spent a lot of efforts to develop the linguistics resources. At the time the corpora are complete and available, it will be managed to support the research in Thai NLP and speech technology research. However, many aspects relating to the corpus construction should be developed for increasing its capacity in NLP researches such as the size, the variety of the raw data, the annotated tag, tools and so on. We are planning to widen the coverage and the collaboration to fulfill the need. Should the standard for data exchange be designed and shared among the research communities in the near future.

5. References

Hansakunbuntheung, C., Tesprasit, V., Sornlertlamvanich, V. (2003). Thai Tagged Speech Corpus for Speech Synthesis. In Proceedings of O-COCOSDA 2003 (pp. 97-104). Singapore.

Tarsaku, P., Sornlertlamvanich, V., Thongprasirt, R. (2001). Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser. In the Proceedings of Eurospeech (Vol. 2, pp. 1057-1060). Scandinavia.

Palingoon, P., Chantanapraiwan, P., Theerawattanasuk, S., Charoenporn, T., Sornlertlamvanich, V. (2002). Qualitative and Quantitative Approaches in Bilingual Corpus-Based Dictionary. In the Proceedings of The Fifth Symposium on Natural Language Processing 2002 & Oriental COCOSDA Workshop 2002 (SNLP-O-COCOSDA 2002) (pp. 152-158). Thailand.

Thongprasirt, R., Sornlertlamvanich, V., Cotsomrong, P., Subevisai, S., Kanokphara, S. (2002). Progress Report on Corpus Development and Speech Technology in Thailand. In the Proceedings of The Fifth Symposium on Natural Language Processing 2002 & Oriental COCOSDA Workshop 2002 (SNLP-O-COCOSDA 2002) (pp. 300-306). Thailand.

Kanokprara, S., Testprasit, V., Thongprasirt, R. (2003). Pronunciation Variation Speech Recognition without Dictionary Modification on Sparse Database. In the Proceedings of International Conference on Acoustic, Speech, and Signal Processing (Vol. 1, pp. 764-767). Hong Kong.

Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Jitsuhiro, T., Kikui, G., Sagisaka, Y. (2003). Thai Speech Database for Speech Recognition (NECTEC-ATR Thai Speech Database). In the Proceedings of O-COCOSDA 2003 (pp. 105-111). Singapore.

Charoenporn, T., Sornlertlamvanich, V. Isahara, H. (1997). Building A Large Thai Text Corpus---Part-Of-Speech Tagged Corpus: ORCHID---. In the Proceedings of the Natural Language Processing Pacific Rim Symposium, (pp.). Thailand.

Sornlertlamvanich, ., Takahashi, N., Isahara, H. (1999) Building a Thai Part-Of-Speech Tagged Corpus (ORCHID). The Journal of the Acoustical Society of Japan (E) (Vol.20, No.3, pp. 189-140).

Sornlertlamvanich, V., Thongpresirt, R. (2001). Speech Technology and Corpus Development in Thailand, In the Proceedings of O-COCOSDA2001 (pp.67-70). Korea.