Development of Very Large Corpora in Thailand

Rachod Thongprasirt, Thatsanee Charoenporn, Wasin Sinthupinyo and Virach Sortlertlamvanich
Information Research and Development Division
National Electronics and Computer Technology Center
539/2 Gypsum Metropolitan Building, Si-Ayudthaya Rd., Rajthevi, Bangkok, 10400, THAILAND
{rachod,thatsanee,wsinthup, virach}@nectec.or.th

Abstract

This paper describes very large corpora, developed in Thailand by the National Electronics and Computer Technology Center (NECTEC), which include text, speech and character image corpora. Since most modern techniques in natural language processing (NLP) rely heavily on large corpora, these resources become crucial for NLP researchers. In addition, these corpora can be used as a standard for comparing the performance between NLP engines. It can also reduce budget and time consumption of NLP researchers if the corpora are collaboratively developed for sharing.

1 Introduction

This paper describes large corpora, developed in Thailand by the National Electronics and Computer Technology Center (NECTEC), which include text, speech and character image corpora. These language resources are crucial for the development of a state-of-the-art natural language processing (NLP) system, since most modern techniques are based on a large corpus.

In section 2, the text corpus is presented. The steps of how to develop the POS-tagged corpus will be discussed in this section.

In section 3, details of speech corpora will be given. It includes the speech recognition corpus aiming at developing a large-vocabulary dictation system. It also mentions the XML-tagged prosody corpus used to improve naturalness of a speech synthesis system. Also, the cooperation project between NECTEC and ATR in order to create a speech corpus is also included.

In section 4, the character image corpus development is discussed. It contains Thai/English character images of 23 fonts and various resolutions. In addition, the online and offline handwritten corpora, which are under development, are also discussed.

2 Text Corpus

The text corpus is initiated with the aim at collecting a large and stylistically varied corpus of standard Thai. In this section, we present the construction of Thai part-of-speech (POS) tagged corpus (Sornlertlamvanich 1997), which is also used as a basic to construct Thai speech corpus. The POS-tagged corpus results from the collaboration between the Communications Research Laboratory (CRL) in Japan and NECTEC with the technical support from the Electrotechnical Laboratory (ETL) in Japan.

The text in the corpus is divided into line labeled with line number for data-retrieval efficiency. The tagged information is originally designed to maintain all necessary information. Currently, the tagged information is not yet committed to any standard mark-up language, but we are going to use XML as a tagging standard in the next development stage. Three levels of annotation are paragraph, sentence and word levels.

In the first process, each paragraph is manually tagged from the input text, and then each sentence in a tagged paragraph is again manually tagged. Unlike previous process, the word segmentation and POS-tagging process are automatic. Instead of separating those processes into two different routines, we combine those processes into one signal routine; i.e., we define the word segmentation and POS tagged as a process of finding the most probable combination of word segmentation and POS assignment. The probabilities of word and POS sequence are computed with trigram model, and the most probable sequence is found by using Viterbi algorithm. In addition, some illegal word segmentations are pruned off by applying a Thai

spelling rule set. The POS set in our text corpus is adapted from that of the Machine translation system. Since the original POS was designed for the MT system, we add some POS's to help clarify ambiguities. The resulting POS set includes 14 categories with 47 subcategories.
Besides adjusting the text corpus to the XML standard framework, we aim at designing a treebank-based grammatical information and annotating grammatical information like POS and syntactic information. We also establish an "internet-based lexicon databank" for compiling all Thai vocabularies and their information such as equivalent word, meaning, syntactic information and conceptual linking from Thai linguists worldwide.

3 Speech Corpus

This section covers three speech corpora: the speech corpus for speech recognition, the speech corpus for ATR and the prosody corpus. All projects have just been launched this year, and all of them are ongoing. The speech for corpus for speech recognition will be covered in detail including the sentence selection process and the statistics of sentence in the phonetically distributed set. The coverage on the speech corpus for ATR is quite superficial. For the prosody corpus, the prosodic tagged information will be given.

3.1 NECTEC Speech Corpus for Speech Recognition

Probably the most prevalent problem in developing Thai speech recognition (SR) is the lack of large standardized speech corpus making it impossible for speech researchers to develop a commercially usable Thai SR module. Fortunately, NECTEC has launched a SR project aiming at developing a large vocabulary continuous speech recognition system. The project is a cooperation between NECTEC and universities. NECTEC selects text and provides funding to universities to record these texts. The details of the text selection and recording processes will be discussed next. Prince of Songkha university (PSU) and Mahanakorn university of technology university (MU) have so far been interested in joining the project.

3.1.1 Overview of the Corpus

The SR corpus contains four sets: the phonetically distributed set (PD), the training set (TR), the development test set (DT), and the evaluation test set (ET). The PD set is created by first forming the phonetically balanced set (PB), the set which covers all possible Biphones, and then adding to the PB set until the Biphone distribution are similar to the reference text. The PD set is used for training an initial acoustic model (AM). The coverage of Biphone in the PB and PD sets includes Biphones occurring within syllable, between syllables and between words. In constructing the PB and PD set, no consideration has been made to a tone.

The TR set is used to train an additional AM and a language model (LM). The number of coverage vocabularies is initially aimed at 5,000 vocabularies. Each sentence must have medium length and medium perplexity. The number of sentences in the TR set is at least 1,000.

The DT and ET sets are used to test the LM at the development time and evaluation time, respectively. Both of these sets have the same properties as those in the TR set except that they contain at least 500 sentences each.

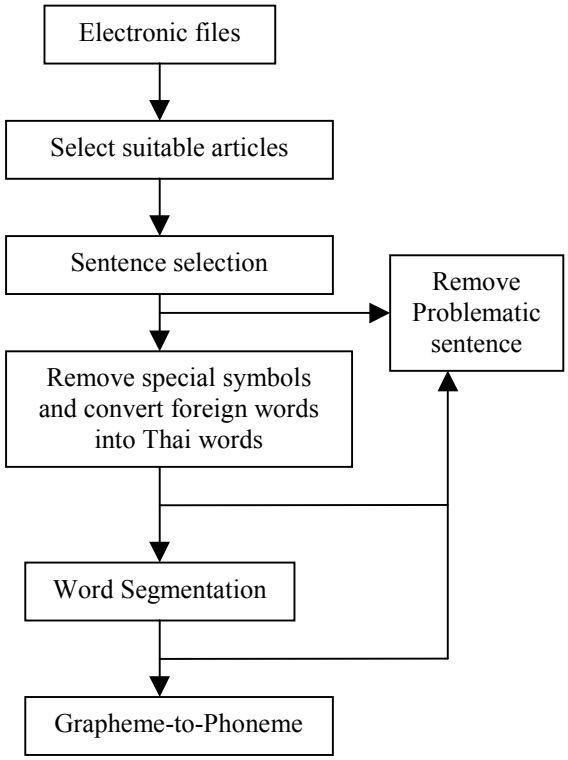

Figure 1. The text management phase.

Table 1 shows the sentence distribution for each speaker. There are six different groups. The overall number of speakers is 248. The first three groups are assigned to two universities, and the last three groups are assigned to NECTEC. The NECTEC group acts as a pilot project to determine in advance any problems that may occur during the corpus creation. It can be seen that each speaker must speak 100 sentences with 20 sentences coming from the PD set. Furthermore, each speaker can only utter sentences from only one of the TR, DT and ET sets.

Table 1. The distribution of sentence to each institute. PSU and MU stand for Prince of Songkha University and Mahanakorn University of Technology, respectively.

Institute	Number of Speakers	Number of sentences/speaker			
		PD	TR	DT	ET
PSU	100	20	80	-	-
MU	50	20	-	40	-
MU	50	20	-	-	40
NECTEC	16	20	80	-	-
NECTEC	16	20	-	40	-
NECTEC	16	20	-	-	40

3.1.2 Database Creation

The database creation is divided into three phases: the text management phase, the PD sentence selection phase, and the TR-DT-ET sentence selection phase. The text management phase is shown in Figure 1 below. Firstly sentences are extracted automatically and then manually selected from the reference text; any problematic sentences are excluded from the database at this phase. Each selected sentence is then modified into its equivalent non-verbal sentence by removing and/or changing any special symbol such as hyphen, command, question mark, repeater symbol. In addition, any English or foreign words are changed to their corresponding Thai words, and sentences or words inside parenthesis are discarded. After selected all usable sentences, the text is sent to the automatic word segmentation, and then rechecked manually. The last process is to map the grapheme for every sentences into their corresponding phonemes. Again, this process is first done automatically and then manually.

In order to construct the PD set, the list of all possible Biphones has been formed. Theoretically there are 2,568 possible Biphones (Luksaneeyanawin 1993). However, by eye inspection, some Biphones can never occur, and it is found that the number of all possible Biphones is only 1,616. Having known all possible Biphones, one can first create the PB set and then the PD set with the method shown in (Shen et al. 1999). Figure 2 summaries all process in this phase.

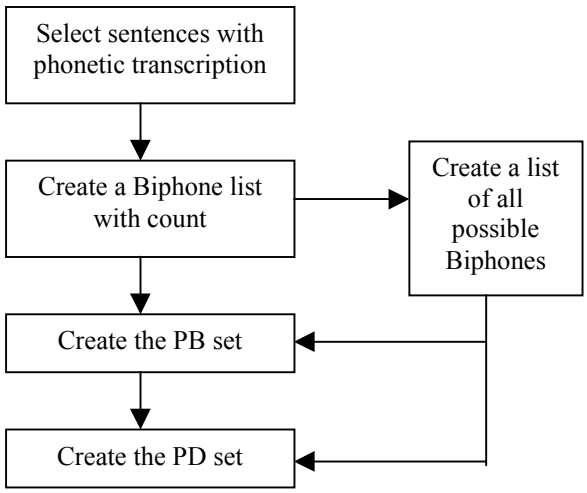

Figure 2. The creation of PD phase.

The construction of the TR, DT, and ET sets are shown in Figure 3 below. Firstly, selected sentences are divided into two sets. The first set, about 90% of text, is used to create an N-gram language model, and to calculate the word frequency list. A set of 5,000 most frequent word list is then formed. A sentence in the second set, the rest 10%, will be selected into the TR, DT or ET set provided that it satisfies all of the following conditions:

1. All words in the sentence are within the 5,000 most frequent word list.
2. The sentence has medium perplexity calculated from the language model created from the first set.
3. The sentence has medium length.

All of these sentences is then distributed into the TR, DT, and ET sets with constraints that each set has to cover all 5,000 frequent words, and the smallest allowable number of sentences

for the TR, DT, and ET are 1,000, 500 and 500, respectively.

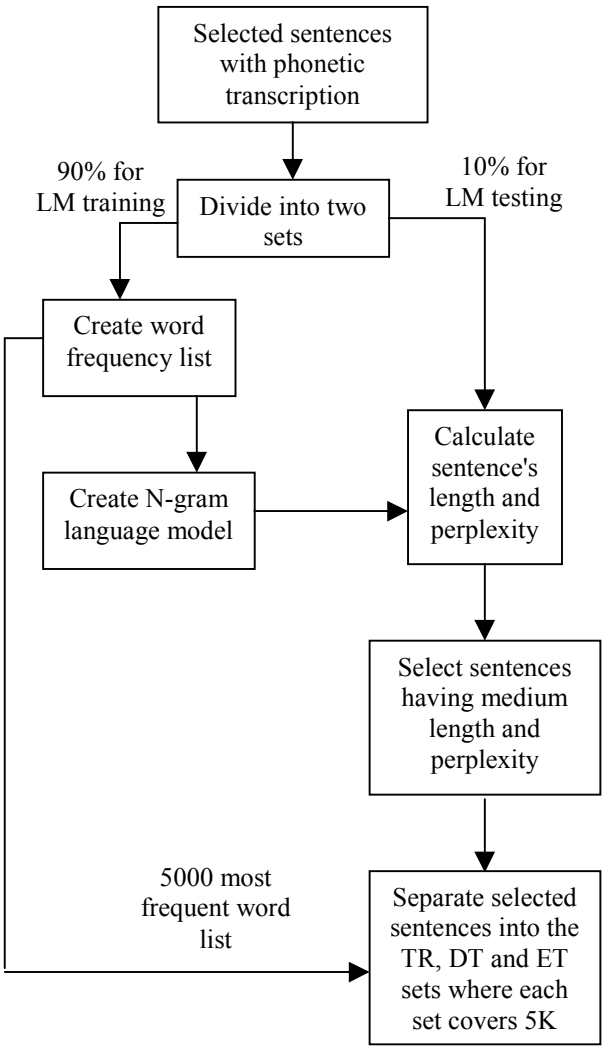

Figure 3. The selection of the TR, DT and ET sets.

3.1.3 Recording Environment

Three recording environments are the clean speech (CL) environment and two office environments (OF1 and OF2). For the clean speech environment, speech is recorded with a high quality dynamic close-talk microphone in a quasi-quiet room. For the office environments, two microphones—a medium quality close-talk and a unidirectional microphones—are used to record speech simultaneously. Each speech is first recorded into DAT with 48 kHz sampling rate and 16-bit quantization levels. The speech waveform is then converted into 16 kHz sampling rate with 16-bit linear Quantization, and it is then recorded as a PC file using the NIST SPHERE header with shorten compression technique.

3.1.4 Transcription and Lexicon Files

There are two transcription files for each sentence in the PD set: the word and phonetic transcriptions that has the following format:
Word transcriptions:

<Start><End><Word><\n>

Phonetic transcriptions:

<Start><End><Phonetics><Tone><\n>

where the <Start> and <End> indicates the starting and ending samples of each word or each phoneme, the <Tone> is the tone indicator, and <\n> denotes the newline character.

Like a sentence in the PD set, each sentence in the TR, DT and ET sets also has two transcription files. The only difference is that there is no timing label, i.e. the <Start> and <END>, in the transcription files.

The Lexicon file that contains information of 5,000 most frequent used words. The format of the information of each word in the Lexicon is:

<Word><Pos><Pronunciation><\n>

where <Pos> is a comma separated list of the part-of-speech of the word (denoted by <Word>) defined in (Sornlertlamvanich 1998), and <Pronunciation> is a list of phonetic transcription together with tonal labeling.

3.1.5 Statistics of the PD Set

We have so far completed selecting the PD set. This section contains some of the statistics from the PD set. The PD set is composed of 802 sentences. Tables 2 shows the minimum, maximum, mean and standard deviation of number of syllables, words and Biphones in the PD set.

Table 3 shows the comparison between the number of theoretical Biphone pairs (2,568) and the actual possible Biphone pairs (1,616) in Thai language. Only about 63% of the theoretical Biphone pairs actually occurs. (1,616 out of 2,568).

Table 2. The statistics of the number of syllables, words and Biphones of sentences in the PD set.

	Number of Syllables	Number of Words	Number of Biphones
Min	7	1	9
Max	117	52	130
Mean	17.07	9.78	36.13
Std.	11.60	5.91	19.48

Table 3 The list of the number of theoretical Biphones and the number of actual Biphones. C_i, V and C_f denote the initial consonant, vowel and final consonant, respectively.

Biphone pairs	Num. of theoretical Biphone pairs	Num. of actual Biphone pairs
C_iV	912	580
VC_f	288	152
C_fC_i	456	556
VC_i	912	328
Total	2,568	1,616

3.2 Speech Corpus for ATR

The ATR Spoken Language Translation Research Laboratories has subsidized NECTEC to create a Thai speech corpus to incorporate into its engine. The project is scheduled to finish at the end of this November. This corpus is divided into three sets: the isolated-word speech database, the phonetically balanced sentence database and the script-scheduled speech database for the hotel reservation task (HRT). The number of speakers is 40 (20 males and 20 females), and all of the recording was done in an office environment.

The isolated-word speech database contains 5,000 daily-used words, the phonetically balanced words (PBW) and some extra words such as credit-card type, money units and Thai province's name. The PBW set is uttered by all forty speakers. The 5,000 daily-used word set is divided into five groups of 1,000 words. Forty speakers are divided into five groups of eight speakers, and each group is then assigned to only one 1000-word group. The 390 phonetically balanced sentences are the PB set described in the previous corpus. The PB set is uttered by all 40 speakers. The HRT database contains 50 conversations between a hotel clerk and a customer. The conversations have been translated from English scripts into Thai scripts, and each person is randomly assigned to utter five different conversations.

Table 4. The description of Prosody speech corpus hierarchy.

At the word level	
Fields	**Descriptions**
Duration	word's duration
POS	Part-of-speech
Position	Relative position of word in the phrase such as beginning, inside or ending
At the syllabic level	
Fields	**Descriptions**
Fo values and Time	The values of fundamental frequency at specified times
Parameters	Pitch contour's parameters
Duration	syllable's duration
Loudness	Loudness in dB
Stress	Stress or non stess syllable
Position	Relative position of syllable in the word
At the phoneme level	
Fields	**Descriptions**
Ci	The initial consonant
V	Vowel
Cf	Final consonant

3.3 Prosody Corpus for Speech Synthesis

The prosody corpus is developed with the aim at improving quality of speech synthesizer. Sentences in the corpus are selected to cover all possible Tri-phone and Tri-tone combinations. In addition, each sentence must have at least seven syllables to avoid any short pause. The number of sentences in the corpus is 5,000. Each sentence is spoken by an experience female narrator. The prosodic information is tagged into each sentence using the extensible markup language (XML). Basically, the tagged

information is divided into levels such as sentence, phrase, word. syllable and phoneme. Table 4 shows some examples of the tagged fields and their descriptions.

4 Character Image Corpus

One cannot deny the importance of Thai and English character image corpus for Thai OCR research. For example, the well-known English OCR engine, Omnifont OCR, was developed with a large training set of multi-font character bitmaps. An efficient Thai OCR, therefore, requires great amount of data for training and testing its module. NECTEC, as a national research center, strives to help Thai OCR researchers (Tanprasert et. al 1997) to save their time preparing data. We have developed Thai and English character image corpus and make it publicly available at http://203.150.243.150.

The current Thai and English character image corpus is a set of training material for printed character recognition. It is composed of 78 Thai, 52 English and 34 special characters, each having four character styles (normal, italic, bold and italic-bold), 23 fonts (AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC, FreesiaUPC, IrisUPC, JasmineUPC, SV Jittra, SV Kanokraykha, SV Kantima, TS PeePee, AgfaTom, DB75Narai, DB95ThaiText, DBFongNam, DBNarai, DBSurawong, DBThaiText, EACChuanPim, EACEact, EACPemai, PSL-Text), three scanning resolutions (200 dpi, 300 dpi, 400 dpi), six font sizes (10, 12, 14, 16, 18, 20) and two scanning types (sharp and normal). The development process is divided into three phases. First, Windows and Macintosh fonts are prepared and printed on a laser printer. In the second phase, the printed fonts are scanned and co-ordinates of each character are kept in the database. In the final phase, an internet application program is developed for searching the database for a character and returning the corresponding bitmap file.

Three more databases are now under development. The first one is a testing Thai OCR database. It consists of 10 types of document (85 newsletter pages, 120 book pages, 100 newpapers pages, 100 magazine pages and 60 journal pages) and 4 types of image resolution (200 dpi black&white, 200 dpi gray-scale, 300 dpi black&white and 300 dpi gray-scale). It contains 930 image files including area labeling information and corrected text. The second one is an online Thai handwritten database that is written on a digitizer from 200 writers. It contains approximately 72,000 character images (60,000 isolated characters and 3,000 words) with handwritten strokes. The last one is an offline Thai handwritten database. It consists of hand-printed samples from 1,000 writers covering approximate 160,000 character images (100,000 isolated characters and 1,500 words) with 300-dpi, gray-scale image resolution.

5 Summary

We have covered the development corpora: the text corpus, the speech corpus, and the character image corpus. When these corpora are available, it would be of great facility for Thai NLP researchers. However, many more corpora would have to be developed in the future. For example, the speech corpora are only limited with reading style; many more styles would be needed in the near future especially conversation speech.

Acknowledgements

The first author would like to thank Chai Wutiwiwatchai who has played an important role in developing the speech corpus for speech recognition. I also appreciate Supphanat Kanokphara on his effort for finding statistics of the PD set. Many thanks to my colleagues at NECTEC as well.

References

Luksaneeyanawin, Sudaporn. 1993. Speech Computing and Speech Technology in Thailand. *Proceeding of the Symposim on Natural Language Processing*, pages 276–321, Bangkok, Thailand.

Shen, Jia-lin, Hsin-min Wang, Ren-yuan Lyu, and Lin-shan Lee. 1999. Automatic Selection of Phonetically Distributed Sentence Sets for Speaker Adaptation with Application to Large Vocabulary Mandarin Speech Recognition, *Computer Speech and Language*, 13:79–98.

Sornlertlamvanich, Virach, Thatsanee Charoenporn, and Hitoshi Isahara. 1997. ORCHID: Thai Part-of-Speech Tagged Corpus.

In NECTEC Technical Report: ORCHID Corpus. pages 5–19.

Sornlertlamvanich, Virach, Naoto Takatoshi, and Hitoshi Isahara. Thai Part-of-Speech Tagged Corpus: ORCHID. *Proceedings of O-COCOSDA Workshop*, pages 131–138, Tsukuba, Japan.

Tanprasert, Chularat, Wasin Sinthupinyo, Premnath Dubey, and Thitipong Tanprasert. 1997. Improved Mixed Thai & English OCR using Two-step Neural Net Classification. *Proceedings of the Internaltional Conference on Neural Information Processing (ICONIP'97)*. 2:1227–1230, New Zealand.