Speech-to-Speech Translation Activities in Thailand

Chai Wutiwiwatchai, Thepchai Supnithi, Krit Kosawat
Human Language Technology Laboratory
National Electronics and Computer Technology Center
112 Pahonyothin Rd., Klong-luang, Pathumthani 12120 Thailand
{chai.wut, thepchai.sup, krit.kos}@nectec.or.th

Abstract

A speech-to-speech translation project (S2S) has been conducted since 2006 by the Human Language Technology laboratory at the National Electronics and Computer Technology Center (NECTEC) in Thailand. During the past one year, there happened a lot of activities regarding technologies constituted for S2S, including automatic speech recognition (ASR), machine translation (MT), text-to-speech synthesis (TTS), as well as technology for language resource and fundamental tool development. A developed prototype of English-to-Thai S2S has opened several research issues, which has been taken into consideration. This article intensively reports all major research and development activities and points out remaining issues for the rest two years of the project.

1 Introduction

Speech-to-speech translation (S2S) has been extensively researched since many years ago. Most of works were on some major languages such as translation among European languages, American English, Mandarin Chinese, and Japanese. There is no initiative of such research for the Thai language. In the National Electronics and Computer Technology Center (NECTEC), Thailand, there is a somewhat long history of research on Thai speech and natural language processing. Major technologies include Thai automatic speech recognition (ASR), Thai text-to-speech synthesis (TTS), English-Thai machine translation (MT), and language resource and fundamental tool development. These basic technologies are ready to seed for S2S research. The S2S project has then been conducted in NECTEC since the end of 2006.

The aim of the 3-year S2S project initiated by NECTEC is to build an English-Thai S2S service over the Internet for a travel domain, i.e. to be used by foreigners who journey in Thailand. In the first year, the baseline system combining the existing basic modules applied for the travel domain was developed. The prototype has opened several research issues needed to be solved in the rest two years of the project. This article summarizes all significant activities regarding each basic technology and reports remaining problems as well as the future plan to enhance the baseline system.

The rest of article is organized as follows. The four next sections describe in details activities conducted for ASR, MT, TTS, and language resources and fundamental tools. Section 6 summarizes the integration of S2S system and discusses on remaining research issues as well as on-going works. Section 7 concludes this article.

2 Automatic Speech Recognition (ASR)

Thai ASR research focused on two major topics. The first topic aimed to practice ASR in real environments, whereas the second topic moved towards large vocabulary continuous speech recognition (LVCSR) in rather spontaneous styles such as news broadcasting and telephone conversation. Following sub-sections give more details.

2.1 Robust speech recognition

To tackle the problem of noisy environments, acoustic model selection was adopted in our system. A tree structure was constructed with each leaf node containing speaker-, noise-, and/or SNR-specific acoustic model. The structure allowed ef-

ficient searching over a variety of speech environments. Similar to many robust ASR systems, the selected acoustic model was enhanced by adapting by the input speech using any adaptation algorithm such as MLLR or MAP. In our model, however, simulated-data adaptation was proposed (Thatphithakkul et al., 2006). The method synthesized an adaptation set by adding noise extracted from the input speech to a pre-recorded set of clean speech. A speech/non-speech detection module determined in the input speech the silence portions, which were assumed to be the environmental noise. This approach solved the problem of incorrect transcription in unsupervised adaptation and enhanced the adaptation performance by increasing the size of adaptation data.

2.2 Large-vocabulary continuous speech recognition (LVCSR)

During the last few years, researches on continuous speech recognition were based mainly on two databases, the NECTEC-ATR (Kasuriya et al., 2003a) and the LOTUS (Kasuriya et al., 2003b). The former corpus was for general purposes, whereas the latter corpus was well designed for research on acoustic phonetics as well as research on 5,000-word dictation systems. A number of research works were reported, starting by optimizing the Thai phoneme inventory (Kanokphara, 2003).

Recently, research has moved closer to real and spontaneous speech. The first task collaborated with a Thai telephone service provider was to build a telephone conversation corpus (Cotsomrong et al., 2007). To accelerate the corpus development, Thatphithakkul et al. (2007) developed a speaker segmentation model which helped separating speech from two speakers being conversed. The model was based on the simple Hidden Markov model (HMM), which achieved over 70% accuracy. Another on-going task is a collection of broadcast news video. The aim of the task is to explore the possibility to use the existing read-speech model to boot broadcast news transcription. More details will be given in Section 5.

3 Machine Translation (MT)

It was a long history of the NECTEC English-to-Thai machine translation (MT) which has been publicly serviced online. The "Parsit" [1] system modified from the engine developed by NEC, Japan, which was a rule-based MT (RBMT). Over 900 parsed rules were coded by Thai linguists. The system recognized more than 70,000 lexical words and 120,000 meanings.

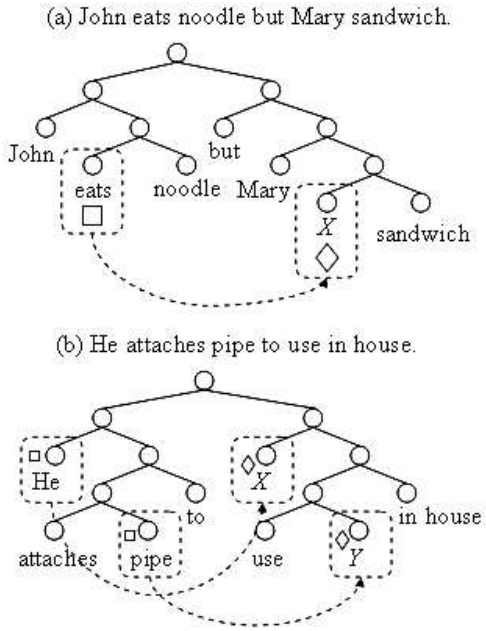

Figure 1. Examples of using MICG to solve two major problems of parsing Thai, (a) coordination with gapping and (b) verb serialization.

3.1 Thai-to-English MT

Recently, there has been an effort to develop the first rule-based system for Thai-to-English MT. The task is much more difficult than the original English-to-Thai translation since the Thai word segmentation, sentence breaking, and grammar parser are all not complete. Coding rules for parsing Thai is not trivial and the existing approach used to translate English to Thai cannot be applied counter wise. Last year, a novel rule-based approach appropriate for Thai was proposed (Boonkwan and Supnithi, 2007). The approach, called memory-inductive categorical grammar (MICG), was derived from the categorical grammar (CG). The MICG introduced memorization and induction symbols to solve problems of analytic languages such as Thai as well as many spo-

[1] Parsit MT, http://www.suparsit.com/

ken languages. In parsing Thai, there are two major problems, coordination with gapping and verb serialization. Figure 1 shows examples of the two problems with the MICG solution, where the square symbol denotes the chunk to be memorized and the diamond symbol denotes the chunk to be induced. A missing text chunk can be induced by seeking for its associated memorized text chunk.

3.2 TM and SMT

In order to improve the performance of our translation service, we have adopted a translation memory (TM) module in which translation results corrected by users are stored and reused. Moreover, the service system is capable to store translation results of individual users. A naïve user can select from the list of translation results given by various users. Figure 3 captures the system interface.

Due to powerful hardware today, research has turned to rely more on statistical approaches. This is also true for the machine translation issue. Statistical machine translation (SMT) has played an important role on modeling translation given a large amount of parallel text. In NECTEC, we also realize the benefit of SMT especially on its adaptability and naturalness of translation results. However, a drawback of SMT compared to RBMT is that it works quite well on a limited domain, i.e. translating in a specific domain. This is actually suitable to the S2S engine which has been designed to work in only a travel domain. Therefore, in parallel to RBMT, SMT is being explored for limited domains. Two parallel text corpora have been constructed. The first one, collected by ATR under the Asian speech translation advanced research (A-STAR)[2] consortium, is a Thai incorporated version of the Basic travel expression (BTEC) corpus (Kikui et al., 2003). This corpus will seed the development of S2S in the travel domain. The second parallel corpus contains examples of parallel sentences given in several Thai-English dictionaries. The latter corpus has been used for a general evaluation of Thai-English SMT. Details of both corpora will be given in the Section 5.

4 Text-to-Speech Synthesis (TTS)

Thai TTS research has begun since 2000. At present, the system utilizes a corpus-based unit-selection technique. A well-constructed phonetically-balanced speech corpus, namely "TSynC-1", containing approximately 13 hours is embedded in the TTS engine, namely "Vaja"[3]. Although the latest version of Vaja achieved a fair speech quality, there are still a plenty of rooms to improve the system. During the past few years, two major issues were considered; reducing the size of speech corpus and improving unit selection by prosody information. Following sub-sections describe the detail of each issue.

4.1 Corpus space reduction

A major problem of corpus-based unit-selection TTS is the large size of speech corpus required to obtain high-quality, natural synthetic-speech. Scalability and adaptability of such huge database become a critical issue. We then need the most compact speech corpus that still provides acceptable speech quality. An efficient way to reduce the size of corpus was recently proposed (Wutiwiwatchai et al., 2007). The method incorporated Thai phonetics knowledge in the design of phoneme/diphone inventory. Two assumptions on diphone characteristics were proved and used in the new design. One was to remove from the inventory the diphone whose coarticulation strength between adjacent phonemes was very weak. Normally, the corpus was designed to cover all tonal diphones in Thai. The second strategy to reduce the corpus was to ignore tonal levels of unvoiced phonemes. Experiments showed approximately 30% reduction of the speech corpus with the quality of synthesized speech remained.

4.2 Prosody-based naturalness improvement

The baseline TTS system selected speech units by considering only phoneme and tone context. In the past few years, analyses and modeling Thai prosodic features useful for TTS have been extensively explored. The first issue was to detect phrasal units given an input text. After several experiments (Tesprasit et al., 2003; Hansakunbuntheung et al., 2005), we decided to develop a classification and decision tree (CART) for phrase break detection.

The second issue was to model phoneme duration. Hansakunbuntheung et al. (2003) compared several models to predict the phoneme duration.

[2] A-STAR consortium, http://www.slc.atr.jp/AStar/

[3] Vaja TTS, http://vaja.nectec.or.th/

Mainly, we found linear regression appropriate for our engine as its simplicity and efficiency. Both two prosody information were integrated in our Vaja TTS engine, which achieved a better synthesis quality regarding subjective and objective evaluations (Rugchatjaroen et al., 2007).

5 Language Resources and Tools

A lot of research issues described in previous sections definitely requires the development and assessment of speech and language corpora. At the same time, there have been attempts to enhance the existing language processing tools that are commonly used in a number of advanced applications. This section explains the activities on resource and tool development.

5.1 Speech and text corpora

Table 1 summarizes recent speech and text corpora developed in NECTEC. Speech corpora in NECTEC have been continuously developed since 2000. The first official corpus under the collaboration with ATR was for general purpose (Kasuriya et al., 2003a). The largest speech corpus, called LOTUS (Kasuriya et al., 2003b), was well-designed read speech in clean and office environments. It contained both phonetically balanced utterances and news paper utterances covering 5,000 lexical words. The latter set was designed for research on Thai dictation systems. Several research works utilizing the LOTUS were reported as described in the Section 2.2.

The last year was the first-year collaboration of NECTEC and a telephone service provider to develop the first Thai telephone conversation speech corpus (Cotsomrong et al., 2007). The corpus has been used to enhance the ASR capability in dealing with various noisy telephone speeches.

Regarding text corpora, as already mentioned in the Section 3, two parallel text corpora were developed. The first corpus was a Thai version of the Basic travel expression corpus (BTEC), which will be used to train a S2S system. The second corpus developed ourselves was a general domain. It will be used also in the SMT research. Another important issue of corpus technology is to create golden standards for several Thai language processing topics. Our last year attempts focused on two sets; a golden standard set for evaluating MT and a golden standard set for training and evaluating Thai word segmentation. Finally, the most basic but essential in all works is the dictionary. Within the last year, we have increased the number of word entries in our lexicon from 35,000 English-to-Thai and 53,000 Thai-to-English entries to over 70,000 entries both. This incremental dictionary will be very useful in sustaining improvement of many language processing applications.

Table 1. Recent speech/text corpora in NECTEC.

Corpus	Purpose	Details
LOTUS	Well-designed speech utterances for 5,000-word dictation systems	- 70 hours of phonetically balanced and 5,000-word coverage sets
TSynC-1	Corpus-based unit-selection Thai speech synthesis	- 13 hours prosody-tagged fluent speech
Thai BTEC	Parallel text and speech corpora for travel-domain S2S	- 20,000 textual sentences and a small set of speech in travel domain
Parallel text	Pairs of Thai-English sample sentences from dictionaries used for SMT	- 0.2M pairs of sentences
NECTEC-TRUE	Telephone conversation speech for acoustic modeling	- 10 hours conversational speech in various telephone types

5.2 Fundamental language tools

Two major language tools have been substantially researched, word segmentation and letter-to-sound conversion. These basic tools are very useful in many applications such as ASR, MT, TTS, as well as Information retrieval (IR).

Since Thai writing has no explicit word and sentence boundary marker. The first issue on processing Thai is to perform word segmentation. Our baseline morphological analyzer determined word boundaries and word part-of-speech (POS) simultaneously using a POS n-gram model and a predefined lexicon. Recently, we have explored Thai named-entity (NE) recognition, which is expected to help alleviating the problem of incorrect word segmentation. Due to the difficulty of Thai word segmentation, we initiated a benchmark evaluation on Thai word segmentation, which will be held in 2008. This will gather researchers who are inter-

ested in Thai language processing to consider the problem on a standard text corpus.

The problem of incorrect word segmentation propagates to the letter-to-sound conversion (LTS) module which finds pronunciations on the word basis. Our original LTS algorithm was based on probabilistic generalized LR parser (PGLR). Recently, we proposed a novel method to automatically induce syllable patterns from a large text with no need for any preprocessing (Thangthai et al., 2006). This approach largely helped alleviating the tedious work on text corpus annotation.

Another important issue we took into account was an automatic approach to find pronunciations of English words using Thai phonology. The issue is particularly necessary in many languages where their local scripts are always mixed with English scripts. We proposed a new model that utilized both English graphemes and English phonemes, if found in an English pronunciation dictionary, to predict Thai phonemes of the word (Thangthai et al., 2007).

6 Speech-to-Speech Translation (S2S)

In parallel to the research and development of individual technology elements, some efforts have been on the development of Thai-English speech-to-speech translation (S2S). Wutiwiwatchai (2007) already explained in details about the activities, which will be briefly reported in this section.

As described briefly in the Introduction, the aim of our three-year S2S project is to develop an S2S engine in the travel domain, which will be given service over the Internet. In the last year, we developed a prototype English-to-Thai S2S engine, where major tasks turned to be the development of English ASR in the travel domain and the integration of three core engines, English ASR, English-to-Thai RBMT, and Thai TTS.

6.1 System development

Our current prototype of English ASR adopted a well-known SPHINX toolkit, developed by Carnegie Mellon University. An American English acoustic model has been provided with the toolkit. An n-gram language model was trained by a small set of sentences in travel domain. The training text contains 210 patterns of sentences spanning over 480 lexical words, all prepared by hands. Figure 2 shows some examples of sentence pattern.

```
Call <DIGIT>
A CONTAINER of DRINK
Check the bill
I come from COUNTRY
I want to go to PLACE
I want to go to the nearest STOP
I would like to have FOOD
What time does VEHICLE go
```

Figure 2. Examples of sentence patterns for language modeling (uppercases are word classes, bracket means repetition).

In the return direction, a Thai ASR is required. Instead of using the SPHINX toolkit[4], we built our own Thai ASR toolkit, which accepts an acoustic model in the Hidden Markov toolkit (HTK)[5] format proposed by Cambridge University. The "iSpeech"[6] toolkit that supports an n-gram language model is currently under developing.

The English ASR, English-to-Thai RBMT, and Thai TTS were integrated simply by using the 1-best result of ASR as an input of MT and generating a sound of the MT output by TTS. The prototype system, run on PC, utilizes a push-to-talk interface so that errors made by ASR can be alleviated.

6.2 On-going works

To enhance the acoustic and language models, a Thai speech corpus as well as a Thai-English parallel corpus in the travel domain is constructing as mentioned in the Section 5.1, the Thai version of BTEC corpus. Each monolingual part of the parallel text will be used to train a specific ASR language model.

For the MT module, we can use the parallel text to train a TM or SMT. We expect to combine the trained model with our existing rule-based model, which will be hopefully more effective than each individual model. Recently, we have developed a TM engine. It will be incorporated in the S2S engine in this early stage.

In the part of TTS, several issues have been researched and integrated in the system. On-going works include incorporating a Thai intonation

[4] CMU SPHINX, http://cmusphinx.sourceforge.net/
[5] HTK, Cambridge University, http://htk.eng.cam.ac.uk/
[6] iSpeech ASR, http://www.nectec.or.th/rdi/ispeech/

model in unit-selection, improving the accuracy of Thai text segmentation, and learning for hidden Markov model (HMM) based speech synthesis, which will hopefully provide a good framework for compiling TTS on portable devices.

7 Conclusion

There have been a considerable amount of research and development issues carried out under the speech-to-speech translation project at NECTEC, Thailand. This article summarized and reported all significant works mainly in the last few years. Indeed, research and development activities in each technology element, i.e. ASR, MT, and TTS have been sustained individually. The attempt to integrate all systems forming an innovative technology of S2S has just been carried out for a year. There are many research and development topics left to explore. Major challenges include at least but not limited to the following issues:

- The rapid development of Thai-specific elements such as robust Thai domain-specific ASR and MT
- Migration of the existing written language translation to spoken language translation

Recently, there have been some initiations of machine translation among Thai and other languages such as Javi, a minor language used in the southern part of Thailand and Mandarin Chinese. We expect that some technologies carried out in this S2S project will be helpful in porting to the other pairs of languages.

Acknowledgement

The authors would like to thank the ATR, Japan, in initiating the fruitful A-STAR consortium and in providing some resources and tools for our research and development.

References

Boonkwan, P., Supnithi, T., 2008. *Memory-inductive categorial grammar: an approach to gap resolution in analytic-language translation*, To be presented in IJCNLP 2008.

Cotsomrong, P., Saykham, K., Wutiwiwatchai, C., Sreratanaprapahd, S., Songwattana, K., 2007. *A Thai spontaneous telephone speech corpus and its applications to speech recognition*, O-COCOSDA 2007.

Hansakunbuntheung, C., Tesprasit, V., Siricharoenchai, R., Sagisaka, Y., 2003. *Analysis and modeling of syllable duration for Thai speech synthesis*, EUROSPEECH 2003, pp. 93-96.

Hansakunbuntheung, C., Thangthai, A., Wutiwiwatchai, C., Siricharoenchai, R., 2005. *Learning methods and features for corpus-based phrase break prediction on Thai*, EUROSPEECH 2005, pp. 1969-1972.

Kanokphara, S., 2003. *Syllable structure based phonetic units for context-dependent continuous Thai speech recognition*, EUROSPEECH 2003, pp. 797-800.

Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Jitsuhiro, T., Kikui, G., Sagisaka, Y., 2003a. *NECTEC-ATR Thai speech corpus*, O-COCOSDA 2003.

Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Kanokphara, S., Thatphithakkul, N., 2003b. *Thai speech corpus for speech recognition*, International Conference on Speech Databases and Assessments (Oriental-COCOSDA).

Kikui, G., Sumita, E., Takezawa, T., Yamamoto, S., 2003. *Creating corpora for speech-to-speech translation*, EUROSPEECH 2003.

Tesprasit, V., Charoenpornsawat, P., Sornlertlamvanich, V., 2003. *Learning phrase break detection in Thai text-to-speech*, EUROSPEECH 2003, pp. 325-328.

Rugchatjaroen, A., Thangthai, A., Saychum, S., Thatphithakkul, N., Wutiwiwatchai, C., 2007. *Prosody-based naturalness improvement in Thai unit-selection speech synthesis*, ECTI-CON 2007, Thailand.

Thangthai, A., Hansakunbuntheung, C., Siricharoenchai, R., Wutiwiwatchai, C., 2006. *Automatic syllable-pattern induction in statistical Thai text-to-phone transcription*, INTERSPEECH 2006.

Thangthai, A., Wutiwiwatchai, C., Ragchatjaroen, A., Saychum, S., 2007. *A learning method for Thai phonetization of English words*, INTERSPEECH 2007.

Thatphithakkul, N., Kruatrachue, B., Wutiwiwatchai, C., Marukatat, S., Boonpiam, V., 2006. *A simulated-data adaptation technique for robust speech recognition*, INTERSPEECH 2006.

Wutiwiwatchai, C., 2007. *Toward Thai-English speech translation*, International Symposium on Universal Communications (ISUC 2007), Japan.

Wutiwiwatchai, C., Saychum, S., Rugchatjaroen, A., 2007. *An intensive design of a Thai speech synthesis corpus*, To be presented in International Symposium on Natural Language Processing (SNLP 2007).