The Development of a Thai Spoken Dialogue System

Chai Wutiwiwatchai, Non-member

ABSTRACT

This article summarily reports the development of the first Thai spoken dialogue system (SDS), namely a Thai Interactive Hotel Reservation Agent (TIRA) . In the development, a multi-stage technique of spoken language understanding (SLU), which combined a word spotting technique for concept extraction and a pattern classification technique for goal identification, was proposed. To improve system performance, a novel approach of logical n-gram modeling was developed for concept extraction in order to enhance the SLU robustness. Furthermore, dialogue contextual information was used to help understanding and finally, an error detection mechanism was constructed to prevent unreliable interpretation outputs of the SLU. All the mentioned algorithms were incorporated in the TIRA system and evaluated by real users.

Keywords: Thai spoken dialogue system, Spoken language understanding

1. INTRODUCTION

A spoken dialogue system (SDS) is a complex system that combines technologies from several areas such as automatic speech recognition, spoken language understanding, dialogue and database management, and text-to-speech synthesis. Many research sites have developed spoken dialogue systems in various domains and languages. Well-known systems include MIT weather information [1], and AT & T operator assistance [2]. A difficulty in creating such systems depends strongly on the capability of handling mixed-initiative dialogues, which means the degree with which the system maintains an active role in the conversation. Other important factors are naturalness and the number of contents recognized by the system.

At present, there is no research site reporting a Thai spoken dialogue system. Thai writing is an alphabetic system with extra-symbols indicating five tonal levels of a syllable. There is typically no space between words and no sentence boundary marker. Although Thai has a phonetic spelling system [3], predicting the pronunciation from the orthography is not a trivial problem, due to a weak relationship between letters and phones. The shortage of efficient phonological and morphological analyzers is a major factor hindering the development of Thai speech recognition. Initiation of a spoken dialogue system for Thai is therefore a challenging task as we have to develop most of sub-components specifically for Thai.

We proposed a pioneering spoken dialogue system for Thai language. The system, namely Thai Interactive Hotel Reservation Agent (TIRA), provides customers a voice-enable interactive hotel information retrieval and reservation. The customers can talk in a natural, spontaneous speaking style. We also aim to create a mixed-initiative dialogue system, where the customers can either answer to system-directive questions or interrupt the system by asking for anything they need at any stage of the conversation.

There have been three versions of TIRA. The first version consisted of several components constructed mainly by rules. To improve the accuracy, a multi-stage SLU was developed and trained by a large size of text collected via a webpage simulating desired dialogues. The novel SLU module was integrated in the second version of TIRA. In the third version, a new algorithm namely logical n-gram modeling was introduced to improve the SLU robustness, while several ways to incorporate dialogue contextual information were tested. An error detection mechanism based on confidence scoring was also incorporated.

In this article, we summarize the development of the TIRA system as follows. The next section describes architecture of the TIRA system. Section 3 and 4 explain in details the highlighted components of TIRA version 2 and 3 respectively. Evaluations that demonstrate the development trend are given in Sect. 5 followed by discussion and conclusion in Sect. 6.

2. TIRA SYSTEM

All the three versions of TIRA have a similar basic structure as shown in Fig. 1.

- *User interface (UI)* - Users talk to the system in a push-to-talk interface. The system displays three necessary key-values, the number of guests, check-in and check-out date, on a small screen and displayed response text in a main screen. Synthesized sound is played to the users while displaying response text. The users can interrupt the system at any time they want.

Manuscript received on June 12, 2006; revised on August 20, 2006.

The author is with Human Language Technology Laboratory, National Electronics and Computer Technology Center (NECTEC) 112 Pahonyothin Rd., Klong-luang, Pathumthani, Thailand 12120; E-mail: chai@nectec.or.th

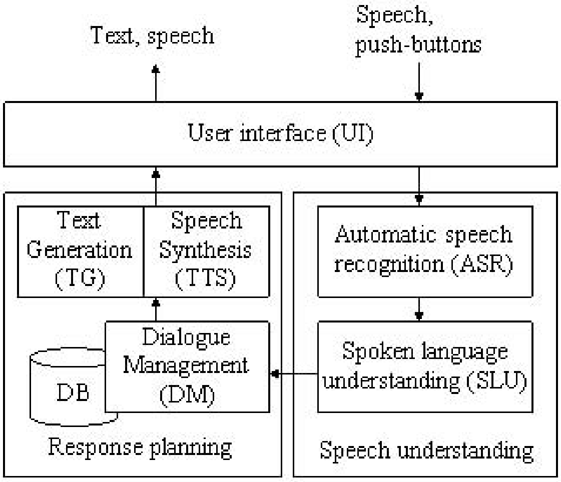


Fig.1: *The Diagram of TIRA System*

• *Automatic speech recognition (ASR)* - An tied-state triphone acoustic model was created by HTK . A word-class n-gram language model was constructed using CMU-SLM toolkit . Speech decoder is JULIUS .

• *Spoken language understanding (SLU)* - The idea behind our SLU module is similar to case grammar. Concepts (semantic slots) such as “the number of people” and “check-in date” appeared in a sentence indicate the goal of user utterance such as “giving prerequisite keys for room reservation”. The SLU hence consisted of concept extraction and goal identification sub-modules, each constructed by rules in the first version of TIRA.

• *Dialogue management (DM)* - Capability to handle mixed-initiative turns is added in a rule-based DM. According to the control rules, the system executes internal functions depending on an incoming semantic input, dialogue histories, and dialogue state variables, and produces a set of semantic responses. Some typical internal functions include “consistency verification”, in which input keys are parsed to the existing keys perceived in previous turns and “database retrieval”, where room information is retrieved from a database.

• *Text generation (TG)* and *Text-to-speech synthesis (TTS)* - Given a semantic response, a textual response is then generated using a template-based text generation. At the same time, a response sound is synthesized from the textual response using a Thai text-to-speech synthesizer provided by NECTEC, Thailand [4].

Figure 2 demonstrates a typical dialogue transcribed from a real conversation between a user and the TIRA system. It is noted that there is an open question of “What else may I help you?”, which is challenging in developing a natural spoken dialogue system.

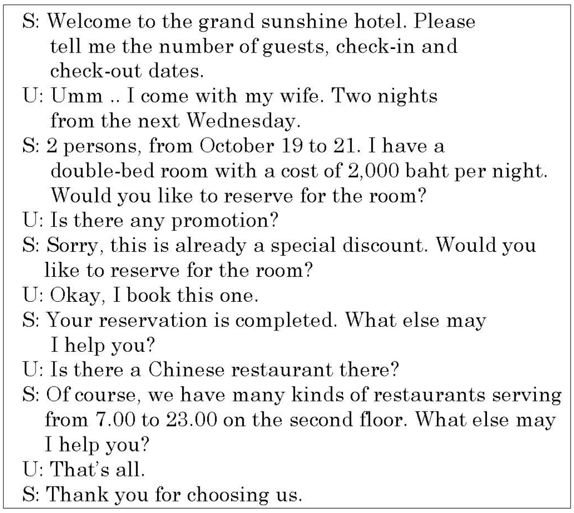

S: Welcome to the grand sunshine hotel. Please tell me the number of guests, check-in and check-out dates.
U: Umm .. I come with my wife. Two nights from the next Wednesday.
S: 2 persons, from October 19 to 21. I have a double-bed room with a cost of 2,000 baht per night. Would you like to reserve for the room?
U: Is there any promotion?
S: Sorry, this is already a special discount. Would you like to reserve for the room?
U: Okay, I book this one.
S: Your reservation is completed. What else may I help you?
U: Is there a Chinese restaurant there?
S: Of course, we have many kinds of restaurants serving from 7.00 to 23.00 on the second floor. What else may I help you?
U: That’s all.
S: Thank you for choosing us.

Fig.2: *A Typical Dialogue of the TIRA System*

3. TIRA VERSION 2

The TIRA version 1 (TIRA-1) was constructed mainly by handcrafted rules. Coding rules in several components is not a trivial task as it needs specialists or linguists who know well the behavior of spoken dialogues in the desired domain. The handcrafted rules themselves often conflict each other and hence require an efficient way to select the best one. A common approach to solve the problem is to instead implement a statistical model that can be learned from a well-annotated corpus. In the TIRA-2 system, a novel learnable multi-stage SLU approach was developed in order to solve the mentioned problem. Furthermore, the invention was suited to spoken language where sentence grammar was often loose especially for Thai.

3.1 Understanding Thai Utterances

Following Panupong [5], a Thai spoken utterance is properly analyzed as shown in Fig. 3. Major meaning of the sentence is expressed in the basic part, which consists of several types of basic part such as subject, action, and object part. The sentence can be augmented by various kinds of *supplementary part.* All of these parts share the same characteristics as below.

•Word ordering is crucial within each part. Miss ordering can make confusion.

•Positions of several parts in a sentence are usually unrestricted. Thais can perceive the same meanings given various ways to form the sentence from several sentence parts.

•Often, a sentence part is separated by another part as shown in the Fig. 3.

3.2 Data-driven SLU

In the technology of trainable or data-driven SLU, two different practices based on robust semantic parsing and topic classification have been widely investigated. The first practice aims to tag the words (or group of words) in the utterance with semantic labels, which are later converted to a certain format of semantic representation. A common representation is a frame of semantic slots (or concepts) coupled optionally with their values. To generate such a semantic frame, words in the utterance are usually aligned to a semantic tree by a parsing algorithm such as a probabilistic context-free grammar or a recursive transition network, whose nodes represent semantic symbols of the words and arcs consist of transition probabilities. This approach has been successfully incorporated in several spoken dialogue systems [6, 7]. A drawback of this method is, however, the requirement of a large, fully annotated corpus, i.e. a corpus with semantic tags on every word, to ensure training reliability.

In the second practice, an understanding module is used to classify an input utterance to one of predefined user goals (if an utterance is supposed to have one goal) directly from the words contained in the utterance. In this case, the semantic representation is a semantic symbol representing the goal. A well-known application based on this method is a call routing system [2], where an incoming call is classified to a type of request. An advantage of this application is the need for training sentences tagged only with their goals, one for each utterance. However, another process is required if one needs to obtain more detailed information from the utterance.

Another different technique of SLU is based on word spotting, where only specific keywords are extracted from a word string. Keywords are often information items important for communication such as the name of places and the numbers. Word-spotting based systems are usually implemented by simple rules, finite state machines, or n-gram models [8]. It means that a word-spotting model can be constructed by either handcrafted rules or learning from tagged corpus.

For the Thai hotel reservation domain, we need an SLU system that can capture both global information (user goals) and detailed information (concepts). Since no large annotated tree bank corpus is available, an SLU model that combines the advantages of topic classification and word spotting described above would be optimum. Some systems have investigated on similar ideas of combination, i.e. constructing a multi-pass system to extract concepts or *named-entities* (NE) and a goal or a dialogue-act, but in different ways of implementation as shown in Table 1. Due to the nature of Thai as explained in the Sect. 3.1, such existing techniques remain problematic and another technique of multi-pass SLU needs to be explored.

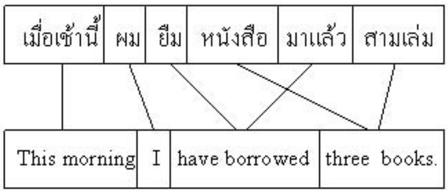
3.3 Multi-stage SLU

Based on the problems described above, we invented a new learnable SLU consisting of three sub-modules as shown in Fig. 4.

• *Concept extraction* - This sub-module aims to extract a set of *concepts* from an input utterance. The concept is actually the basic or supplementary part described in the previous sub-section. Table 2 gives some examples of concepts observed in two utterances. We implemented the concept extraction component by using weighted finite-state transducers (WFSTs). Each WFST represented substrings (word sequences) possibly expressed for the concept. This approach is sometimes called a regular grammar model, denoted in short as '**Reg**'

Table 1: *Combination of Algorithms in Multi-pass SLU*

System	**Concept/NE extraction**	**Goal/Dialogue-act identification**
AT&T HMIHY	HMM	SVM, Boostexter
CUED	Hidden vector state	Naïve Bayes
MS MiPad	CFG, n-gram	Naïve Bayes, SVM



Basic part	Subject	I	ผม
	Action	have borrowed	ยืม ... มาแล้ว
	Object	three books	หนังสือ ... สามเล่ม
Supplementary part	Time	This morning	เมื่อเช้านี้

Fig.3: *Analysis of a Thai Spoken Sentence*

• *Goal identification* - Having extracted the concepts, the goal of the utterance can be identified. The goal in our case can be considered as a derivative of the *dialogue act* coupled with additional information. As the examples show in Table 2, the goal "request_facility" means a request (dialogue act) for some facilities (additional information). Our previous work [9] showed that the goal identification task could be efficiently achieved by the simple multi-layer perceptron type of artificial neural network (ANN).

• *Concept-value recognition* - Some key concepts contain values such as "the number of people" and "facility". The aim of this sub-module is to find out the value inside the word sequence of a concept. During parsing an input sentence by concept WFSTs in the concept extraction sub-module, important keywords indicating concept-values are also marked. The marked keywords are converted to the concept-values

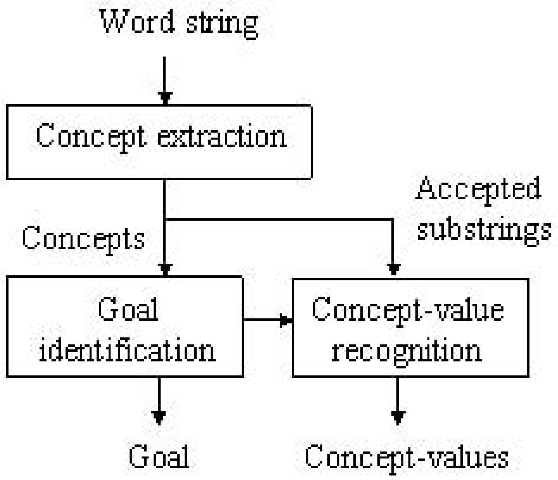


Fig.4: The Structure of Multi-stage SLU

using simple rules.

Table 2: Examples of Goals, Concepts, and Concept-Values

"from the sixth two nights to the eighth of July"		
Goal	inform_prerequisitekeys	
Concept	*Concept-value*	*Corresponding word sequence*
fromdate	July-6	from the sixth .. of July
todate	July-8	to the eight of July
numnight	2	two nights
"there is a pool, right?"		
Goal	request_facility	
Concept	*Concept-value*	*Corresponding word sequence*
reqprovide	-	there is .. right
facility	pool	pool
yesnoq	-	right

The concept WFSTs and the ANN goal identifier can be trained by a partially annotated text corpus. To collect a large corpus, a specific webpage simulating expected dialogues was created and Thai natives were requested to answer to the questions by tying. By this way, we could obtain more than 10,000 sentences (dialogue turns) from over 200 Thai natives within one month. The sentences were filtered, verbalized, word-segmented, annotated, and used to train the ASR language model as well as the SLU module.

4. TIRA VERSION 3

To improve the overall performance of TIRA system, three major issues were considered. The first one was to enhance robustness of the concept extraction component in the SLU module. The second issue was how to incorporate the knowledge of dialogue states into the SLU module, and the last issue is how to prevent errors made by the speech understanding part. These three issues were solved and implemented in the TIRA-3 system. The followings explain in details.

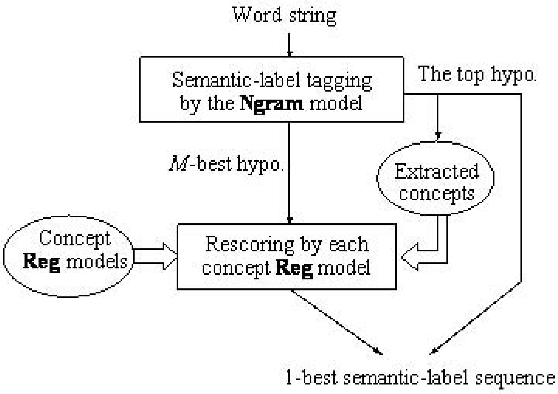

Word string
Semantic-label tagging by the Ngram model
The top hypo.
M-best hypo.
Extracted concepts
Concept Reg models
Rescoring by each concept Reg model
1-best semantic-label sequence

Fig.5: Algorithm of Logical N-gram Modeling

4.1 Logical N-gram Modeling

The concept WFST, the **Reg** model, trained by an annotated corpus accepts only substrings seen in the corpus. To make it more robust to unseen spoken sentences, a statistical model is preferable [10-12]. It can be noted that probabilistic models proposed in several works utilized n-gram probabilities with different designs of semantic units. Since our way to define concepts differs from other systems, a specific design of semantic units is needed.

The concept extraction process could be viewed as a sequence labeling task, where a label sequence, L, is determined given a word string, W. Each label indicates which concept the word lied in. Finding the most probable sequence is equivalent to maximizing the joint probability $P(W,L)$, which can be simplified using n-gram modeling (denoted as **Ngram** modeling).

The n-gram model can assign a likelihood score to any input utterance, it however cannot distinguish between valid and invalid grammar structure. Thus, another probabilistic approach that improved the n-gram model was proposed, namely l*ogical n-gram modeling* (**LNgram**). The algorithm is shown in Fig. 5.

The **LNgram** model, motivated mainly by [13], combines the statistical and structural models in two-pass processing. Firstly, the conventional n-gram model is used to generate M-best hypotheses of label sequences given an input word string. The likelihood score of each hypothesis is then enhanced once its word-and-label syntax is permitted by the regular grammar model. By rescoring the M-best list using the modified scores, the syntactically valid sequence that has the highest n-gram probability is reordered to the top. Even if no label sequence is permitted by the regular grammar, the hybrid model is still able to output the best sequence based on the original n-gram scores. This idea can be implemented efficiently in the framework of WFST.

4.2 Incorporating System Beliefs

the state of dialogue can help much in the understanding process. Incorporating the dialogue-contextual information (system belief) into the SLU component of the TIRA system was performed in two ways. The first way was to develop a dialogue-state dependent semantic model in the concept extraction module. This was simple by training separated **Ngram** models for each dialogue-state and used an interpolation or maximization over all models during parsing.

The second way was to rescore N-best goal hypotheses produced by the ANN goal identifier using system-belief scores. The belief score could be computed as a probability conditioned on the latest system prompt and the dialogue history such as the previous user turn. Rescoring could be performed simply by linear interpolation or, in our proposed method, non-linear estimation such as ANN and Support vector machines (SVM). We found an advantage of using the non-linear estimator when linear-interpolation weights could not be reliably estimated.

4.3 Understanding-error Detection

The final issue implemented in the TIRA system was to incorporate an understanding-error detection mechanism. Given a goal identified by the ANN goal identifier, the mechanism computed confidence score based on several potential features such as the best ANN output value, a difference between the first and the second ANN output values, the number of ANN outputs whose values were greater than a threshold, an average probability of **LNgram** parser, etc.

An accept/reject classifier was used to determine whether the SLU output was reliable based on the confidence scores. The non-linear estimator used for N-best goal rescoring mentioned in the previous subsection could be used at the same time as the accept/reject classifier by just extending the number of estimator inputs to cover the confidence features. This process did not help in improving the SLU accuracy, but prevented harmful errors made by the SLU.

5. EXPERIMENTAL RESULTS

Table 3 briefly summarizes some important characteristics of each version. Figure 6 presents an improvement of ASR performance in terms of word error rate (WER), perplexity of test sets (PP), and out-of-vocabulary rate (OOV). The major reason of WER improvement came from the larger text used for language modeling. Since there was no change from TIRA-2 to TIRA-3 in language modeling, WER results of the two versions were almost equal. It was observed that in this hotel reservation domain, the vocabulary size of 850 in the latter two versions almost covered all words expressed by the users.

Table 3: *Characteristics of 3 Versions of TIRA*

Characteristics	*TIRA-1*	*TIRA-2*	*TIRA-3*
ASR Acoustic modeling Language modeling #Lexical words	4.5 hrs. 12k wrd. 250	6 hrs. 59k wrd. 850	6 hrs. 59k wrd. 850
SLU #Concepts #Goals	16 9	78 40	78 40
DM & TG #Control rules #Response templates	51 27	112 33	114 33
Improving techniques	-	Multi- stage SLU	LNgram, System beliefs, Error detection

Figure 7 shows a trend of SLU improvement by comparing results of concept F-measures (ConF), goal accuracies (GAcc), concept-value accuracies (CAcc), and out-of-goal rates (OOG), i,e, the goals not trained in the SLU. Figure 7(a) presents results when test utterances are manually transcribed, whereas Fig. 7(b) is for automatically speech-recognized test utterances. It is noted that in the case of TIRA-3, goal accuracies are computed after goal rescoring and before error detection. The high improvement of the goal accuracy came from significant reduction of out-of-goals. The improvement trends of the concept F-measure were similar to the concept-value accuracy. However, a few increment of the concept F-measure could highly improve the goal accuracy. This might be due to recovery of dominant concepts. TIRA-3 incorporated the error-detection mechanism, which resulted in 15.6% correct rejection and 5.4% false rejection. This reduced the actual goal accuracy from 83.3% to 78.8% with a benefit of 15.6% error rejection. The correctly rejected turns were not only the ones misclassified by the goal identifier, but also the ones containing serious errors of speech recognition or concept extraction.

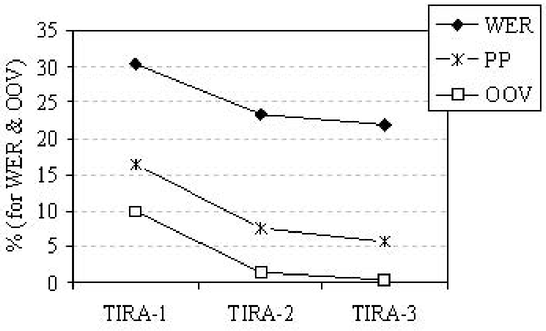


Fig.6: *Improvement of ASR Performance*

Finally, Fig. 8 illustrates an improvement in the

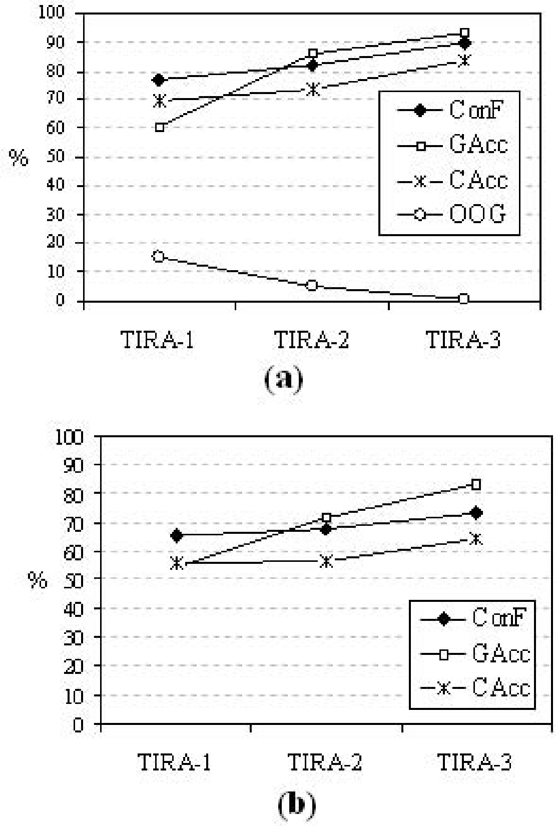


Fig.7: *Improvement of SLU Performance*

term of dialogue success rate separated between na?ve and non-na?ve users. The enhancement of dialogue success in TIRA-2 came from the high improvement of goal and concept-value accuracies due to the larger coverage of concepts and goals in the new data-driven SLU. The success rate was further enlarged in TIRA-3 not only by the higher goal and concept-value accuracies achieved by the logical n-gram model, but also by the use of error-detection mechanism. As expected, na?ve users produced much smaller success rates than non-na?ve users.

6. CONCLUSION

A pioneering Thai spoken dialogue system in the domain of hotel reservation was created under limited resources of fundamental tools and corpora. Speech understanding was mainly focused as it was found to highly affect the overall performance and Thai specificity could improve the performance. Two main breakthroughs include the invention of learnable multi-stage SLU which, trained by a large set of typed-in sentences, achieved a significant increment of SLU accuracy. The other breakthrough was the introduction of the logical n-gram modeling, which highly increased SLU robustness.

There are many issues left for research especially for Thai. There has been a very little effort in incorporating Thai-specific knowledge in the ASR. Furthermore, Passing either 1-best or N-best hypotheses from the ASR to the SLU was only a simple method. There should be a tighter connection approach. The hotel reservation domain is such a complicated task which requires more extensive research and development to achieve a practical system.

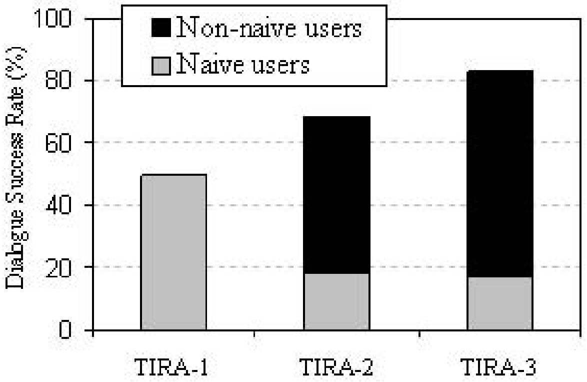


Fig.8: *Improvement of Dialogue Success*

References

[1]] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L., "Jupiter: A telephone-based conversational interface for weather information", *IEEE Trans. Speech Audio Processing*, vol. 8, no.1, pp. 85-96, 2000.

[2] Gorin, A., Riccardi, G., Wright, J., "How may I help you?", *Speech Communication*, vol. 23, pp. 113-127, 1997.

[3] Luksaneeyanawin, S., "Speech computing and speech technology in Thailand", *Proc. SNLP 1993*, pp. 276-321, 1993.

[4] Mittrapiyanurak, P., Hansakunbuntheung, C., Tesprasit, V., Sornlertlamvanich, V., "Issues in Thai text-to-speech synthesis: the NECTEC approach", *Proc. of NECTEC Annual Conference*, Bangkok, pp. 483-495, 2000.

[5] Panupong, V., *The structure of Thai*, Master Thesis, Faculty of Arts, Chulalongkorn University, 1982. (in Thai)

[6] Seneff, S., "TINA: A natural language system for spoken language applications", *Computational Linguistics*, vol. 18, no. 1., pp. 61-86, 1992.

[7] Potamianos, A., Kwang, H., Kuo, J., "Statistical recursive finite state machine parsing for speech understanding", *Proc. ICSLP 2000*, vol. 3, pp. 510-513, 2000.

[8] Kono, Y., Yano, T., Sasajima, M., "BTH: An efficient parsing algorithm for word-spotting", *Proc. ICSLP 1998*, pp. 2067-2070, 1998.

[9] Wutiwiwatchai, C., Furui, S., "Combination of finite state automata and neural network for spoken language understanding", *Proc. EUROSPEECH 2002*, pp. 2761-2764, 2002.

[10] Miller, S., Bobrow, R., Ingria, R., Schwartz, R., "Hidden understanding models of natural language", *Proc. ACL 1994*, pp. 25-32, 1994.

[11] Minker, W., Bennacef, S., Gauvain, J. L., "A stochastic case frame approach for natural language understanding", *Proc. ICSLP 1996*, pp. 1013-1016, 1996.

[12] He, Y., Young, S., "A data-driven spoken language understanding system", *Proc. ASRU*, 2003.

[13] B?chet, F., Gorin, A., Wright, J., and Tur, D. H., "Named entity extraction from spontaneous speech in How May I Help You", *Proc. ICSLP 2002*, pp. 597-600, 2002.

Chai Wutiwiwatchai received his B.Eng. (the first honor goal medal) and M.Eng. degrees in electrical engineering from Thammasat and Chulalongkorn University in Thailand in 1994 and 1997, respectively. In 2001, he received a scholarship from the Japanese government to pursue a Ph.D. at the Furui laboratory at Tokyo Institute of Technology in Japan and graduated in 2004. He has been conducting research in the National Electronics and Computer Technology Center (NECTEC) in Thailand since 1997 and now acts as the Chief of Human Language Technology Laboratory. His research interests include speech and speaker recognition, natural language processing, and human-machine interaction.