論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Pioneering a Thai Language Spoken Dialogue System
著者(和文)	古井 貞熙
Authors(English)	Chai Wutiwiwatchai, Sadaoki Furui
出典(和文)	日本音響学会 2003年春季講演論文集, Vol. , No. 2-4-15, pp. 87-88
Citation(English)	Spring Meeting, Acoust. Soc. Jpn, Vol. , No. 2-4-15, pp. 87-88
発行日 / Pub. date	2003, 3

Pioneering a Thai Language Spoken Dialogue System*

◎*Chai Wutiwiwatchai and Sadaoki Furui (Tokyo Institute of Technology)*

ABSTRACT

A pioneering spoken dialogue system with Thai language interaction is introduced in this paper. The system aims at an automatic hotel reservation task in a mixed-initiative scheme. With an existing small set of speech corpus, an attempt to initialize the system for data collection is reported. Except a speech recognizer and synthesizer, the other submodules including a language understanding module, a dialogue manager, and a text generation module are invented as easily extensible, rule-based components. Descriptions of each submodule are given briefly. The first user's evaluation achieves medium user satisfaction and is a good indicator for us to rapidly improve the current system.

INTRODUCTION

According to the author experience, Thai speech recognition technology is still in an initial step of continuous speech recognition, i.e. there is no speech corpus with considerable size neither for a general dictation system nor any domain specific spoken dialogue system. In the topic of spoken dialogue technology, many languages have been investigated starting from a simple voice command, a system-directive dialogue system, toward a mixed-initiative system, which is widely researched at present. This paper introdudes a pioneering Thai speech interaction dialogue system with a mixed-initiative capability added in. Hotel reservation task is our case study since there has been a small Thai speech corpus related to this domain, created by NECTEC under a sponsor of ATR [1]. With this initial data, we invented the first version of hotel reservation system. The first version aims to be used for data collection and preliminary user evaluation. The results are also reported in this paper.

SYSTEM DESCRIPTION

The system consists of several submodules described below. Initialization of each subsystem on a limitation of data spareness in a new language is the most important issue. Most parts of the system have been developed based on extensible ability, i.e. developers can add new functions by adding in a script file. In the following, we explain the roles of invention of each modules.

Speech Recognizer – An initial set of ATR's hotel reservation corpus for Thai language was prepared for training acoustic models using HTK toolkit, and the models are further speaker adapted using MLLR technique during system evaluation. Word-class n-gram was created by CMU-SLM toolkit and embedded in the JULIUS decoder. A combination of two sentence sets, one from the same ATR's corpus and the other from artificial written text, was used for training. This includes 1,780 sentences which cover about 250 vocabularies.

Language Understanding – An understanding module with unrestricted grammar parser is suitable for a flexible grammar language like Thai. Moreover, any garbage word occurred either by recognizer's errors or by spontaneous speech characteristics, can be eliminated easier. We adopted a strategy based on the idea of case grammar [2]. The module contains two submodules; a *subframe extraction* and a *concept interpretation*. In the former submodule, a set of subframes (or called slots) is extracted from an input text. Each subframe is modeled by a finite state automaton whose arc represents either an isolated word or a word class. The next submodule then interprets the set of subframes, yielding a concept of the input sentence. Figure 1 illustrates an example of understanding process. A set of subframes such as "numperson 2" and "fromdate 2003-1-15" is interpreted as a concept of "inform_keys". At the time of writing this paper, there are 9 concepts and 16 subframes, all written manually.

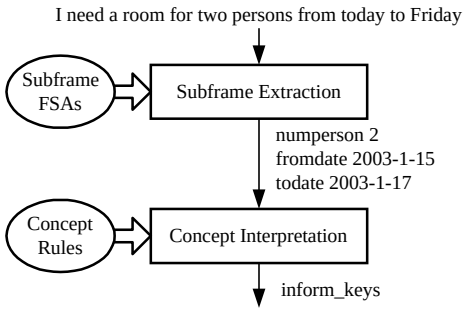

Figure 1. An example of the understanding process

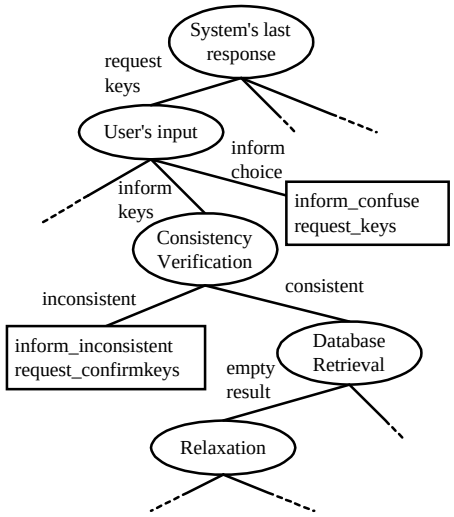

Figure 2. A part of tree representation of rules for the dialogue manager

Dialogue Manager – To achieve a mixed-initiative dialogue system, users must be allowed to ask or inform any related topic at any dialogue state. The system must be able to replace the main task by a requested subtask, and return to the main task after the subtask is solved. This capability has been added to our rule-based dialogue manager. In this module, the rules control the system to execute several submodules depending on an incoming semantic input, dialogue histories, and dialogue state variables. Figure 2 demonstrates a part of dialogue manager's rules in a form of tree representation.

＊タイ語音声対話システムの構築に関する検討
ウッティウィワッチャイ・チャイ、古井貞熙（東京工業大学）

The *consistency verification* subsystem is accessed to check consistency between incoming subframes and the previous system's belief. When three essential keys, "numperson", "fromdate", and "todate" are completely collected, the *database retrieval* component creates an SQL command to retrieve a room list from the hotel reservation database. With empty output from the database retrieval, the *relaxation* submodule then relax the query constraint, e.g. allowing one person staying in a double room, and access the database again. Executation of each subsystem results a value stored in a state variable. All state variables, semantic input, as well as dialogue histories are passed among subsystems via a file called an *internal form* [3]. This method makes the dialogue manager easily extensible. At present, this module consists of 51 rules, i.e. 51 tree leaves and 26 sematic responses.

Text Generation and Text-to-Speech Synthesis – Output from the dialogue manager triggers a template-based text genration module to write some output sentences in Thai language. The template sentences are stored in a script file, which is easily editted and expanded. In parallel, this module generates another sentence used by a Thai text-to-speech synthesizer [4] to create a response sound.

In addition to the core modules described above, a user interface design is also an important part of the system. Our system accepts not only speech, but also some push button inputs. Users can either utter new key values or push a button to change values. They also can push an "undo" button to step back to the previous dialogue state. The system reflexes its belief on the perceived key values on a part of screen, while displaying its response text on the main screen. A push-to-talk style is conducted for this system with a barge-in capability, i.e. ability to interrupt the system's response at any time.

USER EVALUATION

For evaluation, 10 Thai native speakers, 5 males and 5 females, acted as customers using this system. The system is settled on a PC in office environment. After a brief introduction to the system, each speaker was requested to perform 3 conversational dialogues independently, and filled out a questionaire at the end. This resulted 30 dialogues with 404 utterances in total. Table 1 shows evaluation results indicating the performance of each submodule.

In order to evaluation the system, all collected utterances were manually transcribed and parsed by the understanding module. Then manual correction of subframes and concepts was performed for each utterance. Consequently, there exist two sets of results for the understanding process, one from the recognized text input and the other from the transcribed text input. As seen in the table, a high rate of word error rate at around 30% comes from the high OOV rate of 10.0%. During the user evaluation, we also observed a considerable frequency of spontaneous speech phenomena, including hesitation, false start, self repair, repetition, and exclamation. This became another main reason of the observed errors. Almost half of concept error rate comes from 14.9% out-of-concept (OOC) rate. The rest comes from the limitation of restricted finite state automata representing subframes, as well as the capability of the rule-based concept interpretation. The number of "notud" responses indicate how often the system could not understand the user's utterances. This is caused by either the concept interpretation error or any over-expected mixed-initiative utterrances. However, the speech interactive system impressed the users about its intelligence as shown in the questionaire results. Although the users seem to satisfy the current user interface, we observed that there was a very low frequency in using the non-speech input, i.e. push buttons. Two of the users suggested to implement a speech-only input version.

Table 1. Evaluation results from 10 Thai native speakers

General statistics	
No. of utterances (= No. of concepts)	404
No. of words	2693
No. of subframes	503
Speech recognizer	
Trigram perplexity	16.4
OOV rate	10.0%
Word accuracy	69.7%
Language understanding	
Out-of-concept (OOC) rate	14.9%
Concept accuracy (recog / trans)	54.2 / 60.6%
Subframe accuracy (recog / trans)	55.9 / 69.4%
Dialogue manager	
No. of turns	425
No. of "notud" turns	137 (32.2%)
No. of non-speech turns	21 (4.9%)
Average time per dialogue	4 min 29 sec
Questionaire results	(1-5, 5 = best)
User interface	3.9
System response	3.1
Intelligence	3.0
Overall satisfaction	3.5

CONCLUSION

This paper has briefly explained the research on a pioneer spoken dialogue system using Thai language with a hotel reservation task as a case study. Lack of data for training and analysis was the problem of system invention. We adopted a small data set from ATR combined with some artificial sentence set to initialize the speech recognizer and enhanced its performance by the speaker adaptation approach. The language understanding, the dialogue manager, and the text generation are all handcrafted rule-based modules with easily extensible schemes. The first version could achieve an acceptable user satisfaction rate. The evaluation results firstly encourage us to improve the system by decreasing the OOV as well as the OOC. Next, we are planing to find a way to automatically insert more concepts and subframes to the understanding module from a larger training set. Moreover, the dialogue flow will be revised to cover more contents and to enhance the naturalness by adding more rules in the script file.

Acknowledgements
The authors would like to thank for a part of ATR's Thai speech corpus and the text-to-speech tool given by NECTEC, Thailand.

References

[1] Kasuriya, S.and Cotsomrong, P., "Research and Development of Thai Speech Corpus for ATR", *Tech. report RDI-2002, NECTEC*, 2002. (in Thai)

[2] Bennacef, S.K., Bonneau-Maynard, H., Gauvain J.L., Lamel, L., and Minker, W., "A Spoken Language System for Information Retrieval", *Proc. ICSLP 94*, pp. 1271-1274, 1994.

[3] Pieraccini, R., Levin, E. and Eckert, W., "AMICA: the AT&T Mixed Initiative Conversational Architecture," *Proc. of EUROSPEECH 97*, Greece, Sept. 1997.

[4] Mittrapiyanuruk, P., Hansakunbuntheung, C., Tesprasit, V., and Sornlertlamvanich, V., "Issues in Thai Text-to-Speech Synthesis: The NECTEC Approach*", 12th NECTEC Annual Conference*, Bangkok, pp. 483-495, 2000.