論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Integration on Novel Language Understanding in a Thai 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年秋季講演論文集, Vol. , No. 1-6-26, pp. 51-52
Citation(English)	Fall Meeting, Acoust. Soc. Jpn, Vol. , No. 1-6-26, pp. 51-52
発行日 / Pub. date	2003, 9

1—6—26

Integration of Novel Language Understanding in a Thai Spoken Dialogue System*

©Chai Wutiwiwatchai and Sadaoki Furui (Tokyo Institute of Technology)

INTRODUCTION

The first Thai spoken dialogue project was started in the early of 2002. We aim at a mixed-initiative hotel reservation agent. The first user evaluation of our prototype system was reported in [1]. Operating at around 30% word error rate, the system gave approximately 54% concept accuracy. High rate of error came from up to 15% out-of-concept (OOC) turns (the turns whose concepts could not be recognized by the system). This result encouraged us to focus intensively on a systematic approach to reduce OOC. To reduce the manual effort to write the rules, a novel natural language understanding (NLU) was proposed, which could be trained by a partially annotated corpus. It is a combination of a set of weighted finite state automata and a neural network. The automata act as semantic parsers whose results are interpreted as a user intention by the neural network. An offline evaluation of the new NLU was given in [2]. The experiments showed that the proposed model achieved a considerable results on a spoken test set although it was trained only by a typed-in sentence set. This article focuses on an aspect of integrating the proposed NLU engine to our spoken dialogue system. On-site user evaluation is performed and compared with the results obtained in the previous rule-based system.

TIRA SYSTEM

Our spoken dialogue system namely Thai Interactive Hotel Reservation Agent (TIRA) consists of several components. In the ASR, a gender-dependent acoustic model was trained by a part of NECTEC-ATR speech corpus [3] and adapted to our environment by speech utterances collected in the first evaluation of our dialogue system. A word-class n-gram model was constructed from hotel reservation related sentences. After training, the ASR contains a lexicon of 850 words and 28 semantic word classes. The best recognized hypothesis is passed to the NLU module, where a set of semantic slots called *subframes* and a user intention called a *concept* are extracted. Next section explains more details about the NLU module. Given a user concept and corresponding subframes in a specific dialogue state, the dialogue manager (DM) performs some internal actions, and responds to the user via displayed text and synthesized sound. To achieve a mixed-initiative dialogue style, a large number of rules in the DM control table are required to cover every possible dialogue state. The DM control table was developed to be easily extensible by just adding in a script file. Text generation (TG) module is a rule-based system based on semantic driven templates. Given a semantic frame, the TG outputs two sets of text, one for displaying, and the other for Thai text-to-speech synthesizer [4]. The TIRA system was settled in a PC under office environment. Users can communicate with the system via a microphone in push-to-talk style.

INTEGRATION OF THE PROPOSED NLU

Language understanding issue

Overall structure of the proposed NLU system is shown in Figure 1. The first component namely *subframe extraction* is a partial semantic parser implemented by a set of non-deterministic weighted finite state automata (WFSA). A subframe such as "numperson: 1" consists of a subframe label "numperson" (number of persons) and a subframe value "1". Some subframes such as "yesnoq: -" (yes-no question) have labels, but have no value.

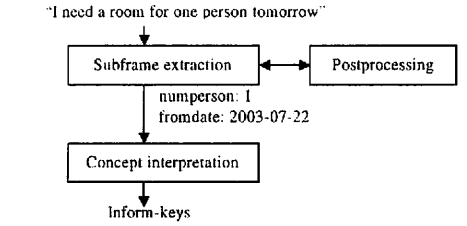

Figure 1: Structure of the language understanding.

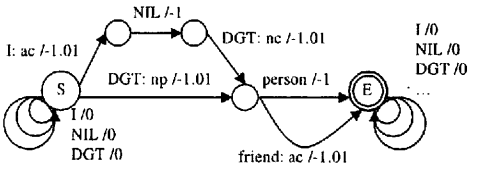

Figure 2: An example of a WFSA representing "numperson" subframe.

Figure 2 illustrates an example of a part of WFSA representing "numperson" subframe. To parse an input word string by this WFSA, some words that are not relevant to this subframe are replaced by a filler symbol "NIL", while some words that appear in predefined semantic classes are replaced by their class symbols such as "DGT" (digits). A set of substrings is accepted after parsing the preprocessed string by the WFSA. By minimizing the cumulative weight of the accepted substrings, the longest one is chosen. More details of this process were explained in [2].

The WFSA not only produces the best accepted substring, but also labels some essential keywords. For example if a substring "one person" is chosen, the symbol "one" is labeled as "np", which indicates the number of persons. Another example is a substring "until the next friday", where the word "friday" is labeled as "tw" (check-out weekday) and the word "next" is labeled as "tr" (relative word for check-out date). In the parsing step, if there are many accepted substrings that have the same cumulative weight, the one that contains more number of output labels has higher priority.

Given the output labeled words from WFSA, the postprocessing component generates a value of the subframe, for example a date value in form of "2003-07-22" for the subframe "fromdate" or "todate". By now, the postprocessing is implemented by rules. A more robust method can be achieved by a statistical model, which is one of our future works.

---

*タイ語音声対話システムにおける言語理解についての検討
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 *concept interpretation* module aims to classify a user concept given a set of accepted subframes. This module utilizes a simple multi-layer perceptron neural network. Input to the network is a vector whose elements are binary values indicating existances of subframes (1 for existing, 0 for none). More complicated input such as an average of bigram probabilities of the accepted substring was also experimented, but it achieved only a little improvement with a much higher cost of computation [2].

The overall NLU system was trained by a set of 6,198 typed-in sentences, which mostly came from our web-based simulated dialogues. 40 different concepts and 78 kinds of semantic subframes were derived.

Dialogue management issue

The improved DM must be able to handle a larger contents supplied by the new NLU. Thanks to the extensible DM system, we can add much more rules to the dialogue control table without difficulty. The previous TIRA system contains a DM with 3 major internal functions; the *consistency verification*, *database retrieval*, and *relaxation* function. In the current version, two internal functions need to be additionally programmed. To handle an anaphora phenomenon, the *reference solving* is applied to some concepts such as "inform-number", in which the users give only a digit sequence in their utterances. The *reference solving* makes decision what the number refers to. Although the NLU system can produce a variety of concepts, some concepts can be implicitly interpreted as the same purpose. For example a concept "req-promotion" and a concept "req-cost" with a subframe "cost: lower" state that the user is more or less unsatisfied with the offered price. We can use a unique response to these concepts, making the conversation more compact. This task is carried out by a *concept clustering* function, which clusters the coming concept and subframes to a proper concept before sending to the rest of the DM module.

In the current system, the number of DM rules expands to 112. The number of semantic responses is 34.

USER EVALUATION

Our focus is to inspect the performance of the newly proposed NLU when working in a real conversation. 22 Thai native speakers, 17 males and 5 females, acted as customers using this system. The ASR was first speaker adapted by a few sentences recorded from the user. Each speaker was requested to perform 3 conversational dialogues separately, and filled out a questionaire at the end. This resulted 66 dialogues with 698 utterances and 3,907 words in total.

The 850-vocabulary ASR achives 76.6% word accuracy with only 1.4% out-of-vocabulary rate. It is worth noting that out of 8,266 sentences used for ASR language modelling, there was only a small number of sentences that came from the transcription of spoken utterances (403 transcriptions from the TIRA1 evaluation). The rest came from a typed-in collection. With a larger size of spontaneous speech transcription, the ASR would be able to achieve a better result.

The NLU was evaluated by either recognized or manually transcribed word sequence. There are three levels of evaluation, concept accuracy, subframe label detection (only subframe label e.g. "numperson"), and subframe value detection (both subframe label and value e.g. "numperson: 1"). Table 1 reports the evaluation results. With high reduction of the out-of-concept (OOC) rate from 14.9% in the TIRA1 evaluation [2] to 5.3% in the current evaluation, the concept accuracy of the recognized input improved from 54.2% to 71.6%. The increasing accuracy not only came from elimination of OOC, but also from more robustness of the proposed NLU. There was a gap of around 15% concept accuracy between the transcribed and the recognized input, which must be reduced by improving the ASR performance. The performance of subframe label detection was considerable, but that of subframe value detection was rather low. The most crucial problem is that the longest matching strategy used in the subframe extraction and the rule-based postprocessing for interpreting a subframe value are somewhat sensitive to the recognition errors. A further analysis on the concept-error utterances showed that up to 80% of utterances were misclassified as the other concepts, whereas the rest was unable to be classified because of no subframe occurred.

Table 1. Evaluation results of the NLU system

Measure (%)	Trans	Recog
Concept accuracy	86.1	71.6
Subframe label precision	80.2	68.0
Subframe label recall	84.5	66.8
Subframe value precision	73.7	56.9
Subframe value recall	77.7	55.9

In terms of dialogue success, from the total of 66 dialogues, 20 dialogues were improperly suspended. Out of 20 dialogues, the users gave up for 11 dialogues, whereas the rest was terminated because the system misunderstood as a request to stop. In terms of mixed-initiative realization, about half of the user turns complied with the system-directive question. The users made 24% mixed-initiative turns i.e. non-compliant turns. The rest was the turns after an open-ended question "What else can I help you?". In total, 22 users provided 43 different concepts, while 32 concepts were coded in our NLU system. An evaluation by a questionaire showed that the system achieved 3.8 of 5 grade satisfaction.

CONCLUSION

This article reported an improvement of TIRA, the first Thai spoken dialogue system in a domain of hotel reservation. A newly proposed NLU was integrated into the system with some modifications such as a postprocessing component of the NLU and some additional internal functions in the DM. In the on-site evaluation, concept accuracy improvement of over 17% came from not only high reduction of OOC, but also more robustness of NLU. Our future work includes an investigation on another WFSA-based statistical approach for subframe extraction, which is less sensitive to a recognition error. An approach for NLU confidence estimation is another important issue to prevent any dialogue failure caused by misunderstanding.

Acknowledgements

The authors would like to thank ATR and NECTEC for the Thai speech corpus and the Thai text-to-speech synthesizer.

References

[1] Wutiwiwatchai, C. and Furui, S., "Pioneering a Thai Language Spoken Dialogue System", *Spring Meeting of Acoustic Society of Japan*, 2-4-15, pp.87-88, March 2003.

[2] Wutiwiwatchai, C. and Furui, S., "An Efficient Language Understanding Approach for Mixed-Initiative Spoken Dialogue Systems", *IEICE Technical Report*, SP2003-29, pp.19-24, May 2003.

[3] Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Jitsuhiro, T., Kikui, G., and Sagisaka, Y., "Thai Speech Database for Speech Recognition (NECTEC-ATR Thai Speech Database)", submitted to *O-COCOSDA 2003*.

[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.