

Robust Discourse Processing Considering Misrecognition in Spoken Dialogue System

Keiichi SAKAI, Yuji IKEDA and Minoru FUJITA

Media Technology Laboratory, CANON Inc.

890-12 Kashimada, Saiwai-ku, Kawasaki-shi, Kanagawa 211 JAPAN

Abstract

In this paper, we propose a new robust discourse processing compensating for misrecognition in a spoken dialogue system.

As the current speech recognition is not able to obtain adequate accuracy, a robust discourse processing which compensates for misrecognition is required in a spoken dialogue system. Here, we developed some functions for the robust discourse processing to reduce extra interactions caused by misrecognition. We incorporated the robust discourse processing functions into a spoken dialogue system and evaluated the effectiveness. The experimental result showed 27% reduction of extra interactions.

1 Introduction

The performance of a speaker independent large vocabulary speech recognition system is now improving[1, 2]. However, it's not adequate enough for a real spoken dialogue system. Unless speech recognition is perfect, what to do with misrecognition is important. Therefore, we propose a new robust discourse processing compensating for misrecognition in a spoken dialogue system in this paper.

We recently established a speaker independent spoken dialogue system, TARSAN[4]. The major mechanisms of TARSAN were: 1) a dynamic vocabulary prediction, 2) a user's intention understanding, and 3) an utterance-pair discourse processing. The state of an utterance-pair is represented as 'context' in this paper. These mechanisms enabled to deal with a large size of vocabulary and to realize a smooth conversation.

The reported experiment showed about 90% of a speech understanding rate[4]. However, we also found that mis-understanding caused a number of extra interactions. The mis-understanding mainly derives from misrecognition. Therefore, a robust discourse processing compensating for misrecognition is required.

In this paper, we first introduce the baseline system, TARSAN. Next, we investigate the misrecognition errors which prevent the user from smooth conversation. The major errors are a) keyword misrecognition,

and b) functional word misrecognition. In both cases, either the dynamic vocabulary prediction, the user's intention understanding, or the utterance-pair discourse processing have problems.

To solve these errors, we propose a new robust discourse processing with adding the following functions.

- A) misrecognized input cancelling:
- B) context-driven intention understanding:

We incorporate these robust discourse processing functions into our spoken dialogue system TARSAN and examine their effectiveness. The effectiveness is measured by comparing the results with the baseline TARSAN and the expanded one.

Finally, we conclude the paper.

2 System Description

A spoken dialogue system TARSAN is a travel guidance system based on full text retrieval using commercial CD-ROMs. The CD-ROMs contain more than 80,000 semi-structured texts about sightseeing spots classified into 24 genres. The user is able to get travel information through conversation along with one's interest. An input by keyboard typing, which covers wider expression than speech, is also available. Table 1 shows an example of a dialogue using TARSAN.

The major mechanisms of TARSAN, which enable a smooth conversation are as follows:

2.1 Dynamic Vocabulary Prediction

To deal with a large scale information in a spoken dialogue system, TARSAN uses a recognition grammar composed of a 'dynamic vocabulary'. The 'dynamic vocabulary' is predicted according to the context in a dialogue, and enable to accept the input in the specific context. This mechanism make it possible to deal with a large scale vocabulary without performance reduction on speed and accuracy.

It is reasonable to use a hierarchical structure of geographical names (Figure 1) as a discourse strategy for a travel guidance. Therefore, we adopt this strategy to realize a smooth conversation. In TARSAN, all district names are always contained as the 'static vocabulary', on the other hand, local place names are processed as the 'dynamic vocabulary'.

Table 1: Example of typical dialogue

usr0:	中国地方にある寺院を教えてください。 (Tell me the temples in Chūgoku-chihō.)
sys0:	岡山県に164件、広島県に151件、山口県に135件、鳥取県に87件など、合計で629件あります。 (There are 629 temples; 164 in Okayama-ken, 151 in Hiroshima-ken, 135 in Yamaguchi-ken, 87 in Tottori-ken, and so on.)
usr1:	鳥取県では。 (How about in Tottori-ken?)
sys1:	鳥取市に19件、倉吉市に12件、鹿野町に6件、東伯町に4件など、合計で87件あります。 (There are 87 temples; 19 in Tottori-shi, 12 in Kurayoshi-shi, 6 in Shikano-chō, 4 in Tōhaku-chō, and so on.)
usr2:	東伯町では。 (How about in Tōhaku-chō?)
sys2:	円応寺、倉坂神社、光徳寺、転法輪寺です。 (There are En-nō-ji, Kurasaka-jinja, Kōtoku-ji, Tempōrin-ji.)
usr3:	円応寺の作成年代は。 (When En-nō-ji was established?)
sys3:	平安時代です。 (In Heian-jidai.)
usr4:	神経痛に効く温泉を知りたい。 (I want to know the hot springs having an effect on neuralgia.)
sys4:	場所を指定して下さい。 (Please designate the place name.)
usr5:	中国地方です。 (It's Chūgoku-chihō?)
sys5:	島根県に29件、山口県に20件、岡山県に18件、広島県に13件など、合計で89件あります。 (There are 89 hot springs; 29 in Shimane-ken, 20 in Yamaguchi-ken, 18 in Okayama-ken, 13 in Hiroshima-ken, and so on.)

Proper names are also processed as the 'dynamic vocabulary' because of their numbers are huge. The proper names are added into the 'dynamic vocabulary', when the number of candidates retrieved from the CD-ROM becomes lower than the threshold.

2.2 Intention Understanding

The user's intention understanding in TARSAN is based on the input modality analysis. The action for the input is decided by the relation between the context in the utterance-pair discourse processing and the user's intention.

The major intentions related with this paper are:

- 'Question': input which ends in "教えて (Tell me...)", "知りたい (I want to know...)", and so on.
- 'Answer': input which ends in "です (desu)".
- 'Condition-change': input which ends in "では (de-ha)".

There are other intentions, such as 'Greeting', 'Yes'/'No' against confirmation, and so on[4].

2.3 Utterance-pair Discourse Processing

In a Q-A type discourse, a recursive structure of the utterances-pairs can often be seen. To deal with this structure, a discourse processing using a utterance-pair stack is proposed[3].

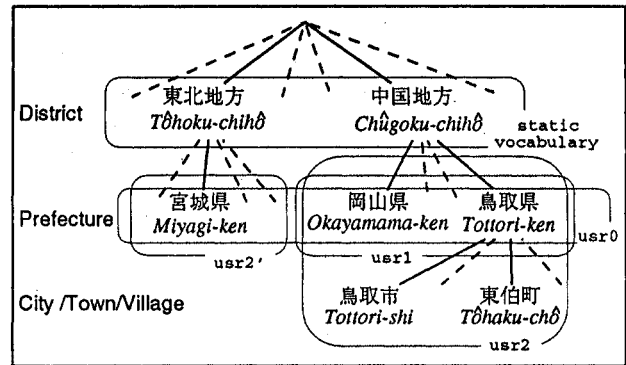


Figure 1: Hierarchical structure of geographical name

A good example of a discourse processing using the utterance-pair stack is shown in Figure 2 and Table 1 (see usr4 - sys5).

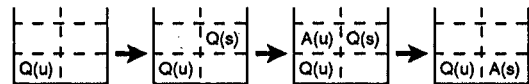


Figure 2: Utterance-pair Stack

The structure of this discourse sequence is Q(u)-Q(s)-A(u)-A(s), where Q(u) indicates that the user's input is 'question' and A(s) indicates that the output of system is 'answer'.

- 1) the intention Q(u) of usr4 is pushed on the user's stack.
- 2) the system asks for further information. The intention Q(s) of sys4 is pushed on the system's stack.
- 3) the user answers for Q(s). the intention A(u) of usr5 is pushed on the user's stack.
- 4) A(u) and Q(s) are popped because A(u) matches Q(s).
- 5) Then, Q(s) and A(u) comes at the top of the stacks.
- 6) A(s) and Q(u) are popped because A(s) matches Q(u). Finally, the discourse processing succeeds.

3 Examples of Misrecognition

We previously evaluated the performance of the baseline TARSAN[4]. From the evaluation, we found that mis-understanding caused a number of extra interactions which prevent the user from smooth conversation. The mis-understanding mainly derives from misrecognition. Therefore, we investigated the misrecognition errors in this section. The major errors in the baseline TARSAN are as follows.

3.1 Keyword Misrecognition

When the keyword (which define the retrieval conditions) is misrecognized, the context and/or the dynamic vocabulary is changed to a wrong ones, which make the user lose one's way. An example of a keyword misrecognition is shown in Table 2.

In usr2', TARSAN misrecognize Tōhaku-chō as Tōhoku-chihō. As the result, all local place names

Table 2: Example of Keyword Misrecognition

usr2':	東北地方では。 (How about in Tōhoku-chihō?)
sys2':	宮城県に 302 件, 福島県に 262 件, 山形県に 249 件, 秋田県に 115 件など, 合計で 1114 件あります。 (There are 1114 temples; 303 in Miyagi-ken, 262 in Fukushima-ken, 249 in Yamagata-ken, 249 in Akita-ken, and so on.)

in the dynamic vocabulary are replaced from the city names including *Tōhaku-chō* to the prefecture names in *Tōhoku-chihō* (see *usr2* and *usr2'* in Figure 1). To return *Tōhaku-chō* into the dynamic vocabulary, the user must begin from *usr0*.

When the misrecognized keyword is a name of a genre, not only the dynamic vocabulary is thrown away, but also the wrong context pushes down the previous context into the utterance-pair stack. In this case, the dynamic vocabulary have to be recovered and the wrong context have to be removed.

3.2 Functional Word Misrecognition

In Japanese, functional words which define the input modality tend to be spoken roughly, because it appears at the end of the sentence. Misrecognition of the functional words causes mis-definition of the modality. This leads mis-understanding of the intention, because the intention is rigidly defined only from the input modality, and the context is ignored. Examples of the misrecongition of a functional word is shown in Table 3.

Table 3: Examples of Functional Word Misrecognition

usr5':	中国地方では。 (How about in Chūgoku-chihō?)
sys5':	別の言い方をして下さい。 (Other expression is required.)
usr5'':	東北地方です。 (It's Tōhoku-chihō.)
sys5'':	福島県に 133 件, 秋田県に 94 件, 山形県に 73 件, 青 森県に 72 件など, 合計で 470 件あります。 (There are 470 hot springs; 133 in Fukushima- ken, 94 in Akita-ken, 73 in Yamagata-ken, 72 in Aomori-ken, and so on.)
usr6'':	中国地方です。 (It's Chūgoku-chihō.)
sys6'':	別の言い方をして下さい。 (Other expression is required.)

In *usr5'*, TARSAN expects for intention 'answer', however, the intention of the input is 'condition-change' because of misrecognition. Thus, the intentions at the top of the utterance-pair stack conflicts. As the result, TARSAN rejects that input as (*sys5'*). This is because TARSAN misrecognizes “です (*desu*)” to “では (*de-ha*)”; the former is a functional word which indicates 'answer', while the latter indicates 'condition-change'.

Here is another example of the intention misunderstanding. In *usr5''*, TARSAN first misrecognizes the keyword, *Chūgoku-chihō* to *Tōhoku-chihō*. Thus, TARSAN replies a wrong result as *sys5''* which is against to the user's request. And again, in *usr6''*, the user intends to correct the keyword by expressing “It's *Chūgoku-chihō*” whose intention is 'answer'. However, this intention conflicts with the context of *sys5''*, and TARSAN rejects the *usr6''*.

4 Robust Discourse Processing

The robust discourse processing is required, because the misrecogntions cannot be avoided in a current spoken dialogue system. Here, we propose a robust discourse processing against misrecogntions mentioned above.

4.1 Misrecognized Input Cancelling

To prevent the user from being lost his way by keyword misrecognition, we add a function to recover the context and the dynamic vocabulary by cancelling the misrecognized input.

In order to realize this function for a robust discourse processing, we expand the utterance-pair processing by incorporating a history mechanism of the context and the dynamic vocabulary.

By this function, the user can recover the context and the dynamic vocabulary easily. In the case of *usr2'*, the user can input *usr2* again after using this function. On the other hand, the baseline system requires the user to begin from *usr0*, which two interactions at least are required to recover the desired dynamic vocabulary.

4.2 Context-Driven Intention Analysis

To prevent the user from being irritated by functional word misrecognition, we add a context-driven intention understanding mechanism. This mechanism decides the intention not only from the input modality, but also from the context.

When the intention decided from the input modality analysis conflicts with the ones which belongs to the context, the intention of the input is assumed to be mis-analyzed because of the misrecognition. In the new intention analysis, the intention of the input is changed according to the context and the confusion probabilities of recognition among the functional words.

Using this mechanism, for example, the intention of *usr5'* is changed from 'condition-change' to 'answer', and *usr5''* from 'answer' to 'condition-change'. After all, either input is accepted as the user intended, and *sys5* is correctly replied (see Table 1 & 3).

5 Experiment

We perform an experiment to examine the effectiveness of the proposed robust processing mechanism. The effectiveness is measured by comparing the results with the baseline and the expanded TARSAN, using the dialogues which we used in [4]. The dialogues are uttered by 20 male subjects by giving typical conversation templates.

Each of the three templates contains 8 interactions; totally 480 interactions unless misrecognition happens. The results are shown in Table 4.

In Table 4, M.R. indicates the misrecognitions. R.I. indicates the interactions for recovering the context and the dynamic vocabulary. E.I. indicates the total extra interactions ($E.I = M.R. + R.I$). T.I. indicates total interactions ($T.I = 480 + E.I$) for the user's to finish the prepared templates.

The reduction of the extra interaction is 27% (from 77 inputs to 56), the misrecognition is 19% (from 57 inputs to 46), and the recovering interaction is 50% (from 20 inputs to 10). Here, we count the recovering interaction of misrecognized input cancelling as one interaction.

Keyword misrecognition

There are 15 keyword misrecognitions in the baseline system. In order to recover the context and the dynamic vocabulary from these misrecognitions, 20 interactions are required, as shown in Table 5. The table indicates that 4 misrecognitions don't have to recover, while 5 misrecognitions need one interaction to recover, 3 need two interactions and 3 need three.

In the expanded system, two keyword misrecognitions is removed and the total misrecognitions decrease to 13. This was because these inputs appear after the recovered inputs. 10 interactions are omitted and The total was reduced to 10.

Functional word misrecognition

There are 12 functional word misrecognitions in the baseline system. In the expanded system, 9 cases out of 12 are improved to be acceptable by the context-driven intention understanding.

The other three functional word misrecognitions are not able to deal with the proposed method, yet. This is because the intention decided by misrecognition does not conflict with the context.

Other misrecognition

There are 30 other misrecognitions. These can be classified into: 1) 11 input repetitions of the same misrecognition, 2) 8 inputs which don't affect the context, 3) 6 input rejections in speech recognition process, 4) and so on.

As for these misrecognitions, a new method should be studied.

Table 4: Results of the experiment

	M.R.	R.I.	E.I.	T.I.
baseline	57	20	77	557
expanded	46	10	56	536

Table 5: Extra interactions by keyword misrecognition

	0	1	2	3	Total
baseline	4	5	3	3	20
expanded	3	10	0	0	10

6 Conclusion & Discussion

In this paper, we proposed a robust discourse processing which compensates for misrecognition in a spoken dialogue system. We incorporated the proposed method into TARSAN and performed an evaluation. The result showed 27% reduction of extra interactions compared to the baseline TARSAN.

As for evaluation, we count the numbers of interactions, however, this evaluation does not take into account of the quality of the dialogue. The comfortableness of the user will differ, even if the counts of interactions are the same. We insist that misrecognition input cancelling is better than utilizing a conversation strategy to recover the context and the dynamic vocabulary, because the the former should only know a single natural function. However, this difference of the quality does not appear in the evaluation that we performed. To evaluate the quality in a spoken dialogue, a new measure is required.

Acknowledgement

The authors wish to thank Dr. Hideyuki TAMURA, Head of the Intelligent Media Div., for giving the opportunity of this study, Dr. Yasuhiro KOMORI for constructive and suitable advice, Mr. Takaya UEDA, Mr. Fumiaki ITOH, and Mr. Masayuki YAMADA for useful discussions. Also thanks to the members of the Media Technology Laboratory for their participation in our experiments.

References

- [1] K.F.Lee, H.W.Hon and R.Reddy, "An Overview of the SPHINX Speech Recognition System," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.38(No.1), pp.34-45 (1990).
- [2] M. Bates, R. Bobrow, P. Fung, R. Ingria, F. Kubala, J. Makhoul, L. Nguyen, R. Schwartz and D. Stallard, "The BBN/HARC Spoken Language Understanding System," Proc. ICASSP93, pp.II.111-II.114 (1993).
- [3] 宮地泰造, 井草ひとみ, 近藤省造, 太細孝, 古川康一, "話題管理機能を持つ対話システムの試作," IPSJ-WGAI, 知識工学と人工知能 38-7 (1985).
- [4] K. Sakai, M. Yamada, F. Itoh, Y. Komori, T. Ueda, Y. Ikeda, "Speech Guidance System with Full Text Retrieval for Texts on CD-ROMs and Its Evaluation," IEICE, Vol. J 77-A No.2, pp.232-240 (1994)(in Japanese).