ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

AN UNDERSTANDING STRATEGY BASED ON PLAUSIBILITY SCORE IN RECOGNITION HISTORY USING CSR CONFIDENCE MEASURE

Toshihiko ITOH†, Atsuhiko KAI‡, Tatsuhiro KONISHI‡ *and* **Yukihiro ITOH‡**

† Hokkaidou University
Kita 8 Nishi 5, Sapporo 060-0808, Japan
‡ Shizuoka University
3-5-1 Johoku, Hamamatsu, Shizuoka, 432-8011, Japan
E-mail address : t-itoh@media.eng.hokudai.ac.jp, kai@sys.eng.shizuoka.ac.jp
konishi@cs.inf.shizuoka.ac.jp, itoh@cs.inf.shizuoka.ac.jp

10.21437/Interspeech.2004-227

ABSTRACT

Although car-navigation systems attract attention as one of spoken dialogue interfaces, recognition errors due to the influence of natural speech and surrounding noise may prevent a smooth dialogue and disappoint the user. Thus, this research aims at the construction of a dialogue system which can achieve a smooth dialogue and a high degree of user satisfaction. Our system performs language understanding and response generation by using the confidence measure(CM) based on continuous speech recognizer(CSR) and the recognition history. This paper shows the spoken language understanding technique in the dialogue system. The CM, together with the speech type and the recognition history, is used for generating an integrated score. The system realizes a spoken language understanding which is more plausible for a given dialogue. As the result of evaluation experiment, it was shown that our system is more efficient (more than 15%) than a language understanding technique which simply gives priority to the higher-rank hypothesis of a speech recognition result (n-best).

1. INTRODUCTION

While some spoken dialogue systems have been developed as practical applications for a car-navigation system[6] or a robot, their usability is undermined by such problems as inability to recognize spontaneous speech, and difficulties in dealing with spoken corrections. Those problems prevent a smooth dialogue and lead to repetitive input for correcting recognition errors, which disappoints the user. The speech miss-recognition is not completely avoided in existing technology, and various research concerning this problem is in progress[1, 2, 3, 4].

We empirically know that there is almost always a correct result in n-best recognition results. Moreover the dialogue is usually coherent and each user's utterance is related with others. Therefore we propose the understanding system which presume most likely interpretation result using not only the latest n-best recognition results but also all the previous n-best recognition results in a recognition history. The most likely interpretation result is determined using calculated word scores and class scores based on the confidence measure(CM) from continuous speech recognizer(CSR). There are some researches for dialogue management using the confidence measure[5, 7], but there is little research for understanding technique using the confidence measure and recognition history.

Thus, this research aims at the construction of a dialogue system which can achieve a smooth dialogue and a high degree of user satisfaction by performing language understanding and response generation using the confidence measure based on continuous speech recognizer and the recognition history.

2. TASK AND SPEECH TYPE

We assume a dialogue task which simulates voice control of a car navigation system, where one should perform landmark setting by entering several landmark names along a driving route.

Figure 1 shows user's utterance pattern. The landmark names involve names of interchanges, stations and cities. The users input landmark names (LM) to the system. The users can supplement LM with prefectures (PR) and routes (RT). PR, RT and LM constitute a tree structure as in Figure 1. The maximum number of keywords which the user can input is 3. Each category has some classes, and the number of classes is 6. The user can input 3 categories at one time or he/she can input them separately.

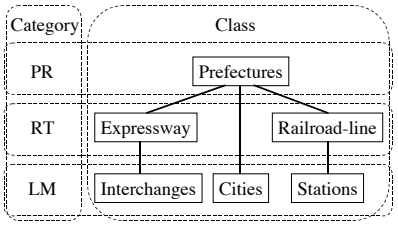


Fig. 1. Utterance pattern

The speech-type in this task consists of the followings:

(1) Narrowing: The user makes input by new or adding words which have a narrowing-relation with the previous input.

(2) Correction: The user corrects the system's recognition errors.

(3) Answer: The user answers a question from the system.

(4) Re-input: The user makes re-input in accordance with the system's re-input request.

Figure 2 shows an example of dialogue.

3. SPOKEN LANGUAGE UNDERSTANDING PROCESS

An overview of the spoken language system for landmark setting is shown in Figure 3. The main components of the system are a

U1:	静岡県です。	(Shizuoka Prefecture.)	
S1:	静岡県でよろしいですか？	(Are you going to Shizuoka Prefecture?)	
U2:	浜松IC。	(Hamamatsu interchange (IC).)	[Narrowing]
S2:	静岡県の何ICですか？	(Which IC in Shizuoka are you going to?)	
U3:	浜松ICです。	(I'm going to Hamamatsu IC.)	[Answer]
S3:	静岡県の浜松西ICですか？	(Are you going to Hamamatsu-nishi IC in Shizuoka?)	
U4:	いいえ、浜松ICです。	(No, I'm going to Hamamatsu IC.)	[Correction]
S4:	もう一度言って下さい。	(Please say once again.)	
U5:	浜松ICです。	(Hamamatsu IC.)	[Re-input]
S5:	静岡県の浜松ICですね？	(Are you going to Hamamatsu IC in Shizuoka?)	
U6:	はい。	(Yes.)	
S6:	目的地に設定しました。	(The destination has been set.)	

Fig. 2. Dialogue example

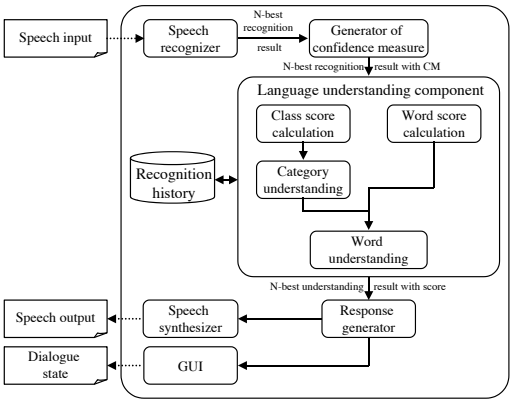


Fig. 3. Dialogue system

speech recognizer[8], a generator of the confidence measure, a language understanding component and a response generator. Among 1,600 read speech (400 * 4 person) based on the four speech types mentioned in the previous section, the SPOJUS[8] correctly recognizes 76% of the keywords.

The speech recognizer outputs ordered n-best recognition results based on acoustic probability. The confidence measure of a word w for a given speech X is estimated as an aposteriori probability $P(w|X)$, which is calculated from the likelihood scores of the n-best sentence hypotheses, $P(X|W)$.

The language understanding component calculates word scores and class scores from the confidence measures and the recognition history. Then the component predicts the category of a given keyword from class scores. Finally, the component determines word sequences which has the maximum score among all the word sequences of predicted category combinations.

The calculated scores indicate the probability of whether a given keyword is uttered in the current task. This score is also used to determine a response of the system. If the user utters a narrowing utterance, the scores of the related keywords in the recognition history are increased. On the contrary if the user utters a correction utterance, the scores of the related keywords which the system predicted are decreased and the scores of other keywords are increased. The generated word scores and class scores are preserved as the recognition history.

A category understanding process calculates a category score from the class scores of the current recognition results and the recognition history. Each category score is the sum of all class scores belonging to each category.

If the category scores (PR, RT, LM) exceed a threshold, the system predicts that the user has uttered that category. Then the understanding process determines a word combination which is allowed by the acceptable category combination and has the maximum score. Of course, the interpretation results are determined in consideration of grammatical constraints and semantic constraints.

Figure 4 and 5 show the example of the language understanding process.

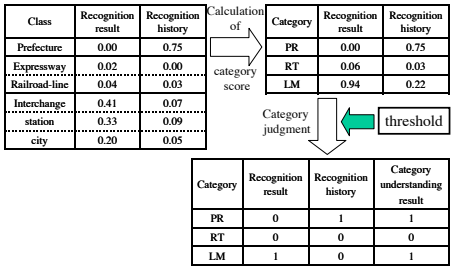


Fig. 4. Category understanding process

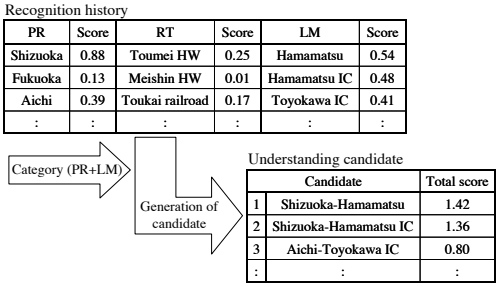


Fig. 5. Word understanding process

3.1. Calculatation process of class scores

Calculation of class scores goes as follows. (1) The language understanding system judges the speech type of the user's utterance using the relations between the recognition history and the current recognition results. (2) Narrowing and answer utterances are processed differently from correction and re-input utterances.

Table 1 shows four conditions for speech type judgement.

Table 1. Speech type judgement

Condition	Judgement
utterance after re-input request	correction or re-input
negation involved in recognition result	correction or re-input
new category introduced into the dialogue	narrowing or answer
otherwise	correction or re-input

These conditions are heuristic and provisional conditions. The rate of correct judgements is 87.7% (narrowing and answer utterance) and 91.4% (correction and re-input utterance).

3.1.1. Class score calculation for narrowing and answer utterances

The narrowing utterances add information to the previous utterances and the positive answer utterances confirm the previous recognition. If an input of the t-th user's turn is judged as these types, the equation (1) is applied to calculate class scores.

$$Score_t(c) = Score_{t-1}(c) * g_{nar} + Conf_t(c) \qquad (1)$$

$Score$: class scores of recognition history
$Conf$: confidence measures of the latest recognition results
g_{nar}: oblivion coefficient ($0.0 < g_{nar} < 1.0$)
c: class to be processed

When the user utters a narrowing utterance or a positive answer, the system can assume that the previous recognition is successful. Accordingly the class scores related to the current recognition results are added to the confidence measures of that class. Because the older the information becomes, the less trustful it becomes, the oblivion coefficient (g_{nar}) is introduced into the equation. The coefficient was experimentally calculated using a development set (DATA-A, mentioned in Section 4) so that understanding accuracy becomes the highest.

3.1.2. Class score calculation for correction and re-input utterances

The equation (2) for the correction and re-input utterances is fundamentally the same as the one for the narrowing and answer utterances, except that the confidence measures of the different class in the same category are reduced from the whole scores. Consequently score correction becomes easy when the system incorrectly recognizes the class of a given keyword.

$$\begin{aligned} Score_t(c_a) &= Score_{t-1}(c_a) * g_{cor} \\ &\quad -Conf_t(C_b) + Conf_t(c_a) \end{aligned} \tag{2}$$

$Score$: class scores of recognition history
$Conf$: confidence measures of the latest recognition results
g_{cor}:oblivion coefficient ($0.0 < g_{cor} < 1.0$)
c_a: class to be processed
c_b: class different from c_a in the same category

3.2. Calculation process of word scores

The word scores are calculated by two steps whenever a new recognition result is inputted. The first step is to check through the existing recognition history. The second step is to process the current n-best recognition results. Different strategies are employed in each step.

In the first process, the word scores are calculated based on the recognition history in consideration of the newness of a word, the contents of the previous system response, the speech type of the user's utterance and so on. The first process involves the following five strategies.

Strategy (1) Whenever a new recognition result is received, all the existing word scores are lowered. This is because the reliability of the previous recognition becomes low, as the information becomes old.

Strategy (2) When the word "A" in the history, and the word "B" in the current results have a narrowing relation, the score of the word "A" is increased.

Strategy (3) When the word "A" in the history, and the word "B" in the current results do not have a narrowing relation, the score of the word "A" is decreased.

Strategy (4) When the affirmation word is contained in the current recognition result, the score of the word contained in the previous system response is increased.

Strategy (5) When the negation word is contained in the current recognition result, the score of the word contained in the previous system response is decreased.

In the second process, the word scores in the current recognition results are calculated from the confidence measures in consideration of the order of the n-best results, the contents of the system response, the speech type of the user's utterance, the lengths of the recognition results and so on. The second process involves the following four strategies.

Strategy (6) When the word "A" in the current recognition results, and the word "B" in the system response have a narrowing relation, the score of the word "A" is increased.

Strategy (7) If the system utters a question and the subsequent user utterance contains possible answers for the question, the relevant word scores are increased.

Strategy (8) Bonus scores based on the ranking is given to the words ranked higher in the n-best recognition results.

Strategy (9) Since longer utterances are easier to recognize, bonus scores are given based on the length of the recognized word sequence.

Those steps are applied to every pair of the words in the recognition history and the words in the current recognition results. When the current recognition results contain the word "W", which also appears in the recognition history, the word score of "W" in the recognition history is adopted regardless of the score of "W" in the recognition results.

For example, if the strategy (2) is applied to the word "W_A", the equation (3) for the word score calculation is as follows.

$$Score_t(W_A) = Score_{t-1}(W_A) + g_2 * Conf_t(W_B) \tag{3}$$

$Score$: word scores of recognition history
$Conf$: confidence measures of the latest recognition results
g_2: coefficient for strategy (2) ($0.0 < g_2 < 1.0$)
W_A: word in the recognition history
W_B: word in the current recognition result

The coefficient g_n is determined for each strategy using the development set (DATA-A) so that accuracy becomes the highest.

Word scores are calculated using an above-mentioned strategies. The calculated word scores are saved as a new recognition history scores.

By this technique, it is expected that reliable words get higher scores, and unreliable words receive low scores. As a result, regardless of the ranking of the present n-best recognition results, priority is given to a word with the highest probability in a dialogue.

The selection of the set of strategies mentioned above are heuristic, and other strategies probably also exist. The detection of other strategies and their implementation are subjects for a future work.

4. EVALUATION

The evaluation experiment was conducted in order to compare the performance of the language understanding technique shown in the paper, and the understanding technique which gives priority to the higher-rank candidate of n-best speech recognition results.

Two systems were prepared for the experiment. One is a spoken language understanding system (SYS-A) which gives top priority to the higher-rank candidate of the newest n-best recognition results, searches the recognition history for words related to the candidate, and output an understanding result. The other is a language understanding system (SYS-B) using the language understanding technique shown in this paper. The two systems have the same components except their language understanding systems.

First, we prepared 14 sentences about "Hamamatsu-nishi interchange". We asked the 5 subjects to utter each prepared sentence 3 times. Using the prepared sentences and outputed recognition results for each of them, we created 3,909 imitation dialogues for the development set (DATA-A), which are possible (natural) sequences of utterances (U1-S1-U2). All the weights required for score equations were calculated using these dialogues so that the rate of the correct understanding became the highest.

Then we asked other 10 subjects to utter the 14 sentences 3 times. We made 38,160 imitation dialogues for an evaluation set (DATA-B) from the 14 sentences and the outputed recognition results in the same way as DATA-A. Since the recognition rate of DATA-B was too much good, the noise of a car was mixed in speech data and recognition rate was lowered. We also made 52,740 imitation dialogues (DATA-C) from outputted results. The word recognition rate of each dialogue (DATA-A, DATA-B and DATA-C) is 80.5%, 93.1% and 78.4%, respectively.

Finally, the evaluation experiment using SYS-B was conducted. We asked the 10 subjects to set 2 landmarks for practice using SYS-B. Then they were asked to set 10 landmarks. Consequently, 361 utterances are collected as an actual dialogue set (DATA-D) in this experiment.

The performance of different systems (SYS-A & B mentioned above & SYS-C) are compared by inputting completely the same n-best recognition results into the systems. We evaluated each system based on the understanding of the second user utterances (U2) and how the system interprets the relations between the keywords in U2 and the recognition history (U1), except for the experiment using DATA-D.

Table 2 shows the accuracy of the understanding. The "Match" is the rate of correct understanding in which the system understands all the keywords in the user's utterances completely. The "Correct" denotes the rate of correctly understood keywords.

The understanding performance of the SYS-B is better than SYS-A in both imitation dialogues and actual dialogues. The evaluation results shows that the system proposed in this paper is effective.

From detailed result, we have noticed that insertion error and deletion error have lowered the whole performance. Our technique predicts an utterance category first and outputs the understanding result consistent with the category. Therefore, if presumption of an utterance category goes wrong, the right understanding cannot be outputted by any means. We experimentally tried a condition in which the correct categories are given apriori and investigated how the whole performance improved (SYS-C).

Since the performance improved significantly, it turns out that it is important to improve the category understanding performance in our framework. In addition, although the rate of category understanding was made 100%, deletion error did not become 0%. This is because the system searched for the words which were consistent with the category understanding. When there is no consistent words, deletion error occurs.

Table 2. Performance of correct understanding for spoken dialogue task

Data	System	#Utt	#Match(%)	#keyword	#Correct(%)
DATA-A	SYS-A	3909	2649(67.8)	9537	8623(90.4)
	SYS-B		3225(83.3)		8830(92.6)
DATA-B	SYS-A	38160	25591(67.1)	93960	90823(96.7)
	SYS-B		36129(94.7)		91863(97.8)
	SYS-C		37903(99.3)		93532(99.5)
DATA-C	SYS-A	52740	35362(67.1)	128790	112342(87.2)
	SYS-B		43304(82.1)		118441(92.0)
	SYS-C		47826(90.7)		119827(93.0)
DATA-D	SYS-A	361	208(57.6)	901	625(69.4)
	SYS-B		278(77.0)		795(88.2)
	SYS-C		315(87.3)		833(92.5)

5. CONCLUSION

This paper showed the spoken language understanding technique in which the confidence measures and the history of speech recognition results were used. In the technique, a score is calculated for all the words of an old recognition result based on the relation of speech recognition results with two or more candidates called n-best, their confidence measures, a recognition history, and the contents of the system response. The most probable understanding result is generated from that score. From the evaluation experiment it was shown that the understanding accuracy of our technique is more than 15% higher than the language understanding technique which gives top priority to the higher-rank candidate of n-best.

As a future work, the framework of category understanding should be improved so that the whole performance of the language understanding will be raised.

6. REFERENCES

[1] S.Oviatt, M.MacEachern and G-A.Levow, Predicting hyperarticulate speech during human-computer error resolution, Speech Communication, Vol.24, pp.87-110 (1998)

[2] M.Swerts, D.Litman and J.Hirschberg, Correction in spoken dialogue systems, Proc. ICSLP'2000, pp.615-618 (2000)

[3] A.Kai, Y.Nonomura, T.Itoh, T.Konishi and Y.Itoh, "Influence of different dialogue situations on user's behavior in spoken corrections", Proc. ICSLP'2002, pp.1189-1192 (2002)

[4] J.Hirasawa, N.Miyazaki, M.Nakano and K.Aikawa, "New Feature Parameters For Detecting Misunderstandings in a Spoken Dialogue System", Proc. ICSLP'2000, Vol.2, pp.154-157 (2000)

[5] Y.Niimi and Y.Kobayashi, "Modeling Dialogue Control Strategies to Relieve Speech Recognition Erros", Proc. European Conf. on Speech Communication and Technology, pp.1177-1180 (1995).

[6] X.Pouteau, E.Krahmer and J.Landsbergen, "Robust Spoken Dialogue Management for Driver Information Systems", Proc. Eurospeech'97, pp.2207-2210 (1997)

[7] D.J.Litman, M.A.Walker and M.J.Kearns, "Automatic detection of poor speech recognition at the dialogue level", Proc. 37th Annual Meeting of the Association of Computational Linguistics (ACL99), pp.309-316 (1999)

[8] S.Nakagawa and A.Kai, "A Context-Free Grammar-Driven, One-Pass HMM-Based Continuous Speech Recognition Method", Systems and Computers in Japan, Vol.25, No.4, pp.92-102 (1994)