Example-based Speech Intention Understanding and Its Application to In-Car Spoken Dialogue System

Shigeki Matsubara†⋆ **Shinichi Kimura**‡ **Nobuo Kawaguchi**†⋆
Yukiko Yamaguchi† and **Yasuyoshi Inagaki**‡
†Information Technology Center, Nagoya University
‡Graduate School of Engineering, Nagoya University
⋆CIAIR, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
matubara@itc.nagoya-u.ac.jp

Abstract

This paper proposes a method of speech intention understanding based on dialogue examples. The method uses a spoken dialogue corpus with intention tags to regard the intention of each input utterance as that of the sentence to which it is the most similar in the corpus. The degree of similarity is calculated according to the degree of correspondence in morphemes and dependencies between sentences, and it is weighted by the dialogue context information. An experiment on inference of utterance intentions using a large-scale in-car spoken dialogue corpus of CIAIR has shown 68.9% accuracy. Furthermore, we have developed a prototype system of in-car spoken dialogue processing for a restaurant retrieval task based on our method, and confirmed the feasiblity of the system.

1 Introduction

In order to interact with a user naturally and smoothly, it is necessary for a spoken dialogue system to understand the intentions of utterances of the user exactly. As a method of speech intention understanding, Kimura et al. has proposed a rule-based approach (Kimura *et al.*, 1998). They have defined 52 kinds of utterance intentions, and constructed rules for inferring the intention from each utterance by taking account of the intentions of the last utterances, a verb, an aspect of the input utterance, and so on. The huge work for constructing the rules, however, cannot help depending on a lot of hands, and it is also difficult to modify the rules. On the other hand, a technique for tagging dialogue acts has been proposed so far (Araki *et al.*, 2001). For the purpose of concretely determining the operations to be done by the system, the intention to be inferred should be more detailed than the level of dialogue act tags such as "yes-no question" and "wh question".

This paper proposes a method of understanding speeches intentions based on a lot of dialogue examples. The method uses the corpus in which the utterance intention has given to each sentence in advance. We have defined the utterance intention tags by extending an annotation scheme of dialogue act tags, called JDTAG (JDRI, 2000), and arrived at 78 kinds of tags presently. To detail an intention even on the level peculiar to the task enables us to describe the intention linking directly to operations of the system.

In the technique for the intention inference, the degree of similarity of each input utterance with every sentence in a corpus is calculated. The calculation is based on the degree of morphologic correspondence and that of dependency correspondence. Furthermore, the degree of similarity is weighted by using dialogue context information. The intention of the utterance to which the maximum score is given in the corpus, will be accepted as that of the input utterance. Our method using dialogue examples has the advantage that it is not necessary to construct rules for inferring the intention of every utterance and that the system can also robustly cope with the diversity of utterances.

An experiment on intention inference has been made by using a large-scale corpus of spoken dialogues. The experimental result, providing 68.9% accuracy, has shown our method to be feasible and effective. Furthermore, we have developed, based on our method, a prototype system of in-car spoken dialogue processing for a restaurant retrieval task, and confirmed the feasiblity of the system.

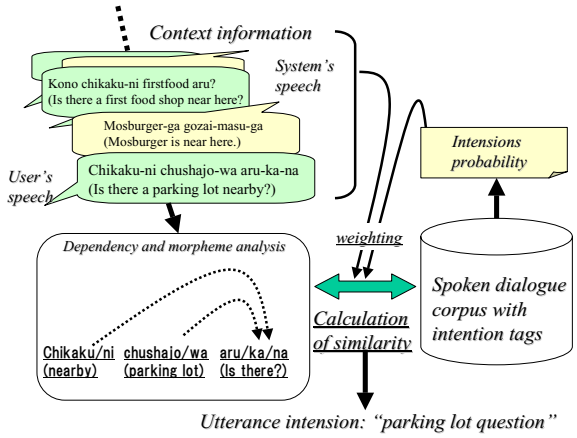

Figure 1: Flow of the intention inference processing

2 Outline of Example-based Approach

Intentions of a speaker would appear in the various types of phenomenon relevant to utterances, such as phonemes, morphemes, keywords, sentential structures, and contexts. An example-based approach is expected to be effective for developing the system which can respond to the human's complicated and diverse speeches. A dialogue corpus, in which a tag showing an utterance intention is given to each sentence, is used for our approach. In the below, the outline of our method is explained by using an inference example.

Figure 1 shows the flow of our intention inference processing for an input utterance "Chikaku-ni chushajo-wa aru-ka-na ? (Is there a parking lot nearby?)". First, morphological analysis and dependency analysis to the utterance are carried out.

Then, the degree of similarity of each input utterance with sentences in the corpus can be calculated by using the degree of correspondence since the information on both morphology and dependency are given to all sentences in the corpus in advance. In order to raise the accuracy of the intention inference, moreover, the context information is taken into consideration. That is, according to the occurrence probability of a sequence of intentions learned from a dialogue corpus with the intention tags, the degree of similarity with each utterance is weighted based on the intentions of the last utterances. Consequently, if the utterance whose degree of similarity with the input utterance is the maximum is "sono chikaku-ni chushajo ari-masu-ka? (Is there a parking lot near there?)", the intention of the input utterance is regarded as "parking lot question".

3 Similarity and its Calculation

This section describes a technique for calculating the degree of similarity between sentences using the information on both dependency and morphology.

3.1 Degree of Similarity between Sentences

In order to calculate the degree of similarity between two sentences, it can be considered to make use of morphology and dependency information. The calculation based on only morphemes means that the similarity of only surface words is taken into consideration, and thus the result of similarity calculation may become large even if they are not so similar from a structural point of view. On the other hand, the calculation based on only dependency relations has the problem that it is difficult to express the lexical meanings for the whole sentence, in particular, in the case of spoken language. By using both the information on morphology and dependency, it can be expected to carry out more reliable calculation.

Formula (1) defines the degree of similarity between utterances as the convex combination β of the degree of similarity on dependency, α_d, and that on morpheme, α_m.

$$\beta = \lambda\alpha_d + (1-\lambda)\alpha_m \qquad (1)$$

α_d: the degree of similarity in dependency
α_m: the degree of similarity in morphology
λ : the weight coefficient $(0 \leq \lambda \leq 1)$

Section 3.2 and 3.3 explain α_d and α_m, respectively.

3.2 Dependency Similarity

Generally speaking, a Japanese dependency relation means the modification relation between a *bunsetsu* and a *bunsetsu*. For example, a spoken sentence "kono chikaku-ni washoku-no mise aru? (Is there a Japanese restaurant near here?)" consists of five *bunsetsu*s of

"kono (here)", "chikaku-ni (near)", "washoku-no (Japanese-style food)", "mise (a restaurant)", "aru (being)", and there exist some dependencies such that "mise" modifies "aru". In the case of this instance, the modifying *bunsetsu* "mise" and the modified *bunsetsu* "aru" are called *dependent* and *head*, respectively. It is said that these two *bunsetsu*s are in a dependency relation. Likewise, "kono", "chikaku-ni" and "washoku-no" modify "chikaku-ni", "aru" and "mise", respectively. In the following of this paper, a dependency relation is expressed as the order pair of *bunsetsu*s like (mise, aru), (kono, chikaku-ni).

A dependency relation expresses a part of syntactic and semantic characteristics of the sentence, and can be strongly in relation to the intentional content. That is, it can be expected that two utterances whose dependency relations are similar each other have a high possibility that the intentions are also so.

A formula (2) defines the degree of similarity in Japanese dependency, α_D, between two utterances S_A and S_B as the degree of correspondence between them.

$$\alpha_d = \frac{2C_D}{D_A + D_B} \tag{2}$$

D_A: the number of dependencies in S_A
D_B: the number of dependencies in S_B
C_D: the number of dependencies in correspondence

Here, when the basic forms of independent words in a head *bunsetsu* and in a dependent *bunsetsu* correspond with each other, these dependency relations are considered to be in correspondence. For example, two dependencies (chikaku-ni, aru) and (chikaku-ni ari-masu-ka) correspond with each other because the independent words of the head *bunsetsu* and the dependent *bunsetsu* are "chikaku" and "aru", respectively. Moreover, each word class is given to nouns and proper nouns characteristic of a dialogue task. If a word which constitutes each dependency belongs to the same class, these dependencies are also considered to be in correspondence.

3.3 Morpheme Similarity

A formula (3) defines the degree of similarity in morpheme α_m between two sentences S_A and S_B.

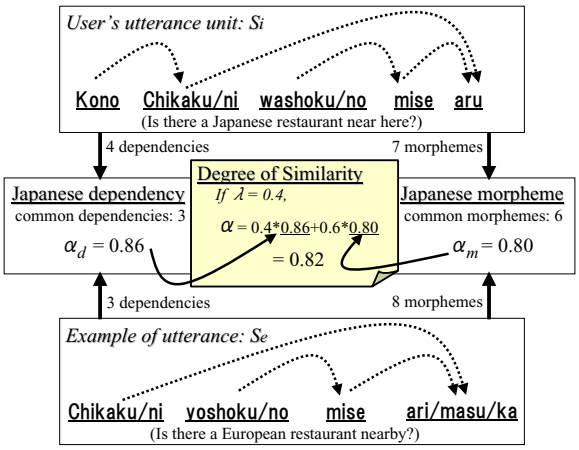

Figure 2: Example of similarity calculation

$$\alpha_m = \frac{2C_M}{M_A + M_B} \tag{3}$$

M_A: the number of morphemes in S_A
M_B: the number of morphemes in S_B
C_M: the number of morphemes in correspondence

In our research, if a word class is given to nouns and proper nouns characteristic of a dialogue task and two morphemes belong to the same class, these morphemes are also considered to be in correspondence. In order to extract the intention of the sentence more similar as the whole sentence, not only independent words and keywords but also all the morphemes such as noun and particle are used for the calculation on correspondence.

3.4 Calculation Example

Figure 2 shows an example of the calculation of the degree of similarity between an input utterance S_i "kono chikaku-ni washoku-no mise aru? (Is there a Japanese restaurant near here?)" and an example sentence in a corpus, S_e, "chikaku-ni yoshoku-no mise ari-masu-ka (Is there a European restaurant located nearby?)", when a weight coefficient $\lambda = 0.4$. The number of the dependencies of S_i and S_e is 4 and 3, respectively, and that of dependencies in correspondence is 2, i.e., (chikaku, aru) and (mise, aru). Moreover, since "washoku (Japanese-style food)" and "yoshoku" (European-style food) belong to the same word class, the dependencies

(washoku, aru) and (yoshoku, aru) also correspond with each other. Therefore, the degree of similarity in dependency α_d comes to 0.86 by the formula (2). Since the number of morphemes of S_i and S_e are 7 and 8, respectively, and that of morphemes in correspondence is 6, i.e., "chikaku", "ni", "no", "mise", "aru(i)" and "wa(yo)shoku". Therefore, α_m comes to 0.80 by a formula (3). As mentioned above, β using both morphemes and dependencies comes to 0.82 by a formula (1).

4 Utilizing Context Information

In many cases, the intention of a user's utterance occurs in dependence on the intentions of the previous utterances of the user or those of the person to which the user is speaking. Therefore, an input utterance might also receive the influence in the contents of the speeches before it. For example, the user usually returns the answer to it after the system makes a question, and furthermore, may ask the system a question after its response. Then, in our technique, the degree of similarity β, which has been explained in Section 3, is weighted based on the intentions of the utterances until it results in a user's utterance. That is, we consider the occurrence of a utterance intention I_n at a certain time n to be dependent on the intentions of the last $N-1$ utterances. Then, the conditional occurrence probability $P(I_n|I_{n-N+1}^{n-1})$ is defined as a formula (4).

$$P(I_n|I_{n-N+1}^{n-1}) = \frac{C(I_{n-N+1}^{n})}{C(I_{n-N+1}^{n-1})} \qquad (4)$$

Here, we write a sequence of utterance intentions $I_{n-N+1} \cdots I_n$ as I_{n-N+1}^{n}, call it **intentions N-gram**, and write the number of appearances of them in a dialogue corpus as $C(I_{n-N+1}^{n})$. Moreover, we call the conditional occurrence probability of the formula (4), **intentions N-gram probability**.

The weight assignment based on the intentions sequences is accomplished by reducing the value of the degree of similarity when the intentions N-gram probability is smaller than a threshold. That is, a formula (5) defines the degree of similarity γ using the weight assignment by intentions N-gram probability.

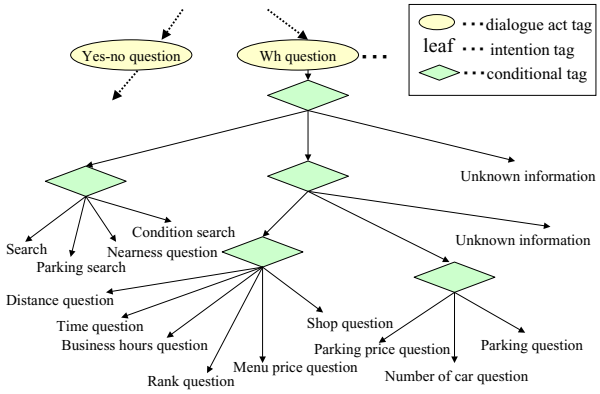

Figure 3: Decision tree of intention tag (a part)

$$\gamma = \begin{cases} \omega\beta & (P(I_n|I_{n-N+1}^{n-1}) \leq \theta) \\ \beta & (otherwise) \end{cases} \qquad (5)$$

ω: weight coefficient ($0 \leq \omega \leq 1$)
β: the degree of similarity
θ: threshold

A typical example of the effect of using intentions N-gram is shown below. For an input utterance "chikaku-ni chushajo-wa ari-masu-ka? (Is there a parking lot located nearby?)", the degree of similarity with a utterance with a tag "parking lot question" which intends to ask whether a parking lot is located around the searched store, and a utterance with a tag "parking lot search" which intends to search a parking lot located nearby, becomes the maximum. However, if the input utterance has occurred after the response intending that there is no parking lot around the store, the system can recognize its intention not to be "parking lot question" from the intentions N-gram probabilities learned from the corpus, As a result, the system can arrive at a correct utterance intention "parking lot search".

5 Evaluation

In order to evaluate the effectiveness of our method, we have made an experiment on utterance intention inference.

5.1 Experimental Data

An in-car speech dialogue corpus which has been constructed at CIAIR (Kawaguchi *et al.*,

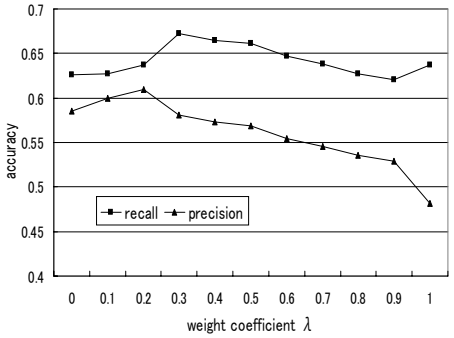

Figure 4: Relation between the weight coefficient λ and the accuracy ($\lambda = 0.3$)

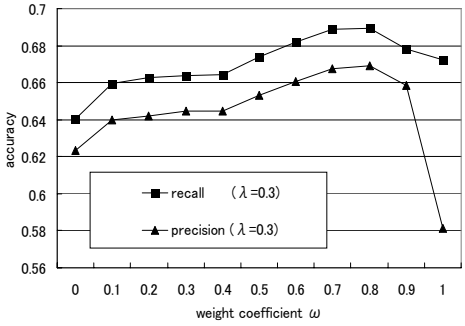

Figure 5: Relation between the weight coefficient ω and the accuracy

2001), was used as a corpus with intention tags, and analyzed based on Japanese dependency grammar (Matsubara *et al.*, 2002). That is, the intention tags were assigned manually into all sentences in 412 dialogues about restaurant search recorded on the corpus. The intentions 2-gram probability was learned from the sentences of 174 dialogues in them. The standard for assigning the intention tags was established by extending the decision tree proposed as a dialogue tag scheme (JDRI, 2000). Consequently, 78 kinds of intention tags were prepared in all (38 kinds are for driver utterances). The intention tag which should be given to each utterance can be defined by following the extended decision tree. A part of intention tags and the sentence examples is shown in Table 1, and a part of the decision tree for driver's utterances is done in Figure 3 [1].

A word class database (Murao *et al.*, 2001), which has been constructed based on the corpus, was used for calculating the rates of correspondence in morphemes and dependencies. Moreover, Chasen (Matsumoto *et al.*, 99) was used for the morphological analysis.

5.2 Experiment

5.2.1 Outline of Experiment

We have divided 1,609 driver's utterances of 238 dialogues, which is not used for learning the intentions 2-gram probability, into 10 pieces equally, and evaluated by cross validation. That is, the inference of the intentions of all 1,609 sentences was performed, and the recall and precision were calculated. The experiments based on the following four methods of calculating the degree of similarity were made, and their results were compared.

1. Calculation using only morphemes
2. Calculation using only denpendencies
3. Calculation using both morphemes and denpendencies (With changing the value of the weight coefficient λ)
4. Calculation using intentions 2-gram probabilities in addition to the condition of 3. (With changing the value of the weight coefficient ω and as $\theta = 0$)

5.2.2 Experimental Result

The experimental result is shown in Figure 4. 63.7% as the recall and 48.2% as the precision were obtained by the inference based on the above method 1 (i.e. $\lambda = 0$), and 62.6% and 58.6% were done in the method 2 (i.e. $\lambda = 1.0$). On the other hand , in the experiment on the method 3, the precision became the maximum by $\lambda = 0.2$, providing 61.0%, and the recall by $\lambda = 0.3$ was 67.2%. The result shows our technique of using both information on morphology and dependency to be effective.

When $\lambda \leq 0.3$, the precision of the method 3 became lower than that of 1. This is because the user speaks with driving a car (Kawaguchi *et al.*, 2000) and therefore there are much comparatively short utterances in the in-car speech corpus. Since there is a few dependencies per

[1] In Figure 3, the description in condition branches is omitted.

Table 1: Intention tags and their utterance examples

intention tag	utterance example
search	Is there a Japanese restaurant near here?
request	Guide me to McDonald's.
parking lot question	Is there a parking lot?
distance question	How far is it from here?
nearness question	Which is near here?
restaurant menu question	Are Chinese noodles on the menu?

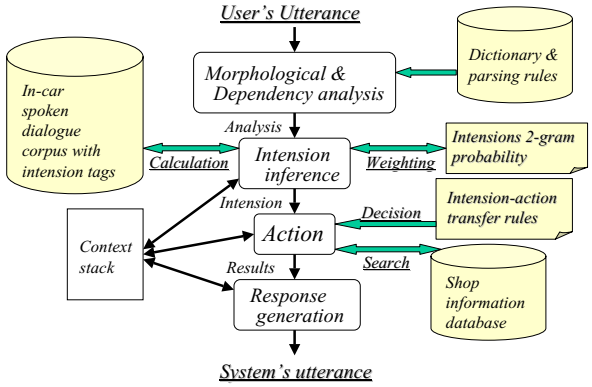

Figure 6: Configuration of the prototype system

one utterance, a lot of sentences in the corpus tend to have the maximum value in inference using dependency information.

Next, the experimental result of the inference using weight assignment by intentions 2-gram probabilities, when considering as $\lambda = 0.3$, is shown in Figure 5. At $\omega = 0.8$, the maximum values in both precision and recall were provided (i.e., the precision is 68.9%). This shows our technique of learning the context information from the spoken dialogue corpus to be effective.

6 In-car Spoken Dialogue System

In order to confirm our technique for automatically inferring the intentions of the user's utterances to be feasible and effective for task-oriented spoken dialogue processing, a prototype system for restaurant retrieval has been developed. This section describes the outline of the system and its evaluation.

6.1 Implementation of the System

The configuration of the system is shown in Figure 6.

Table 2: Comparison between the results on inferred intentions and those on given intentions

Intentions	Inferred		Given	
	num.	rate	num.	rate
Correct	31	51.7%	42	70.0%
Partially corr.	5	8.3%	4	6.7%
Incorrect	7	11.7%	2	3.3%
No action	17	28.3%	12	20.0%

1. **Morphological and dependency analysis**: For the purpose of example-based speech understanding, the morphological and dependency analyses are given to each user's utterance by referring the dictionary and parsing rules. Morphological analysis is executed by Chasen (Matsumoto *et al.*, 99). Dependency parsing is done based on a statistical approach (Matsubara *et al.*, 2002).

2. **Intentions inference**: As section 3 and 4 explain, the intention of the user's utterance is inferred according to the degree of similarity of it with each sentence in a corpus, and the intentions 2-gram probabilities.

3. **Action**: The transfer rules from the user's intentions to the system's actions have been made so that the system can work as the user intends. We have already made the rules for all of 78 kinds of intentions. The system decides the actions based on the rules, and executes them. After that, it revises the context stack. For example, if a user's utterance is "kono chikaku-ni washoku-no mise arimasu-ka (Is there a Japanese restaurant near here?)", its intention is "search". Inferring it, the system retrieves the shop information database by utilizing the key-

words such as "washoku (Japanese restaurant)" and "chikaku (near)".

4. **Response generation**: The system responds based on templates which include the name of shop, the number of shops, and so on, as the slots.

6.2 Evaluation of the System

In order to confirm that by understanding the user's intention correctly the system can behave appropriately, we have made an experiment on the system. We used 1609 of driver's utterances in Section 5.2.1 as the learning data, and the intentions 2-gram probabilities learned by 174 of dialogues in Section 5.1. Furthermore, 60 of driver's utterances which are not included in the learning data were used for the test. We have compared the results of the actions based on the inferred intentions with those based on the given correct intentions. The results have been classified into four groups: correct, partially correct, incorrect, and no action.

The experimental result is shown in Table 2. The correct rate including partial correctness provides 76.7% for the giving intentions and 60.0% for the inferred intentions. We have confirmed that the system could work appropriately if correct intentions are inferred.

The causes that the system based on given intentions did not behave appropriately for 14 utterances, have been investigated. 6 utterances are due to the failure of keywords processing, and 8 utterances are due to that they are out of the system's expectation. It is expected for the improvement of the transfer rules to be effective for the former. For the latter, it is considered to turn the responses such as "I cannot answer the question. If the questions are about $\cdots$, I can do that."

7 Concluding Remarks

This paper has proposed the example-based method for inferring speaker's intention. The intention of each input utterance is regarded as that of the most similar utterance in the corpus. The degree of similarity is calculated based on the degrees of correspondence in both morphemes and dependencies, taking account of the effects of a sequence of the previous utterance's intentions. The experimental result using 1,609 driver's utterances of CIAIR in-car speech corpus has shown the feasibility of example-based speech intention understanding. Furthermore, we have developed a prototype system of in-car spoken dialogue processing for a restaurant retrieval task based on our method.

Acknowledgement: The authors would like to thank Dr. Hiroya Murao, Sanyo Electric Co. LTD. for his helpful advice. This work is partially supported by the Grand-in-Aid for COE Research of the Ministry of Education, Science, Sports and Culture, Japan. and Kayamori Foundation of Information Science Advancement.

References

Araki, M., Kimura, Y., Nishimoto, T. and Niimi, Y.: Development of a Machine Learnable Discourse Tagging Tool, *Proc. of 2nd SIGdial Workshop on Discourse and Dialogue*, pp. 20–25 (2001).

The Japanese Discouse Research Initiative JDRI: Japanese Dialogue Corpus of Multilevel Annotation, *Proc. of 1st SIGdial Workshop on Discourse and Dialogue* (2000).

Kawaguchi, N., Matsubara, S., Iwa, H., Kajita, S, Takeda, K., Itakura, F. and Inagaki, Y.: Construction of Speech Corpus in Moving Car Environment, *Proc. of ICSLP-2000*, Vol. III, pp. 362–365 (2000).

Kawaguchi, N., Matsubara, S., Takeda, K. and Itakura, F.: Multimedia Data Collection of In-car Speech Communication, *Proc. of Eurospeech-2001*, pp. 2027–2030 (2001).

Kimura, H., Tokuhisa, M., Mera, K., Kai, K. and Okada, N.: Comprehension of Intentions and Planning for Response in Dialogue, *Technical Report of IEICE*, TL98-15, pp:25–32 (1998). (In Japanese)

Matsubara, S., Murase, T., Kawaguchi, N. and Inagaki, Y.: Stochastic Dependency Parsing of Spontaneous Japanese Spoken Language, *Proc. of COLING-2002* (2002).

Matsumoto, Y., Kitauchi, A., Yamashita, T. and Hirano, Y.: Japanese Morphological Analysis System Chasen version 2.0 Manual, *NAIST Techinical Report*, NAIST-IS-TR99009 (1999).

Murao, H., Kawaguchi, N., Matsubara, S. and Inagaki, Y.: Example-based Query Generation for Spontaneous Speech, *Proc. of ASRU-2000* (2001).