An Example-based Approach to Speech Intention Understanding

Shinichi Kimura†, Shigeki Matsubara‡⋆, Nobuo Kawaguchi‡⋆,
Yukiko Yamaguchi‡ and Yasuyoshi Inagaki†

†Graduate School of Engineering, Nagoya University
‡Information Technology Center, Nagoya University
⋆Center for Integrated Acoustic Information Research, 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 dependency relations between sentences, and it is weighted by the dialogue context information. An experiment on inference of utterance intentions using an in-car spoken dialogue corpus has shown 68.9% accuracy.

Keywords: spoken dialogue system, spoken language processing, dependency parsing, dialogue corpus, intention tag, in-car speech, CIAIR speech database

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 [5]. 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 [1]. 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 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 [2], 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.

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 each sentence is given a tag showing an utterance intention, 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 "kono chikaku-ni washoku-no mise aru ? (Is there a Japanese restau-

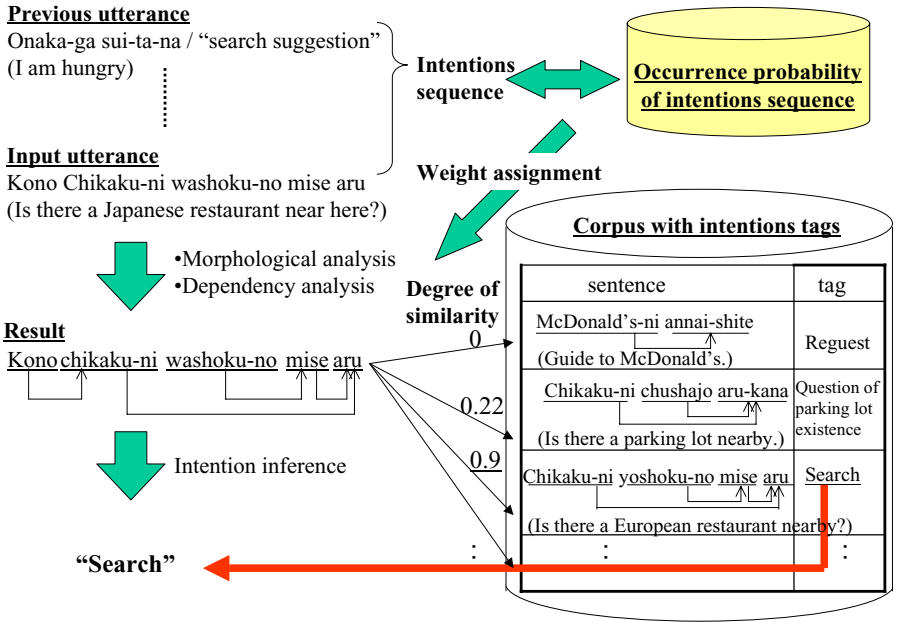

Figure 1: Flow of the intention inference processing

rant near here?)". 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, however, 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 "chikaku-ni yoshoku-no mise aru? (Is there a European restaurant nearby?)", the intention of the input utterance is regarded as "search".

3 DEGREE OF 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.

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 such a case might occur that the result of similarity calculation becomes 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.

Eq. (1) defines the degree of similarity between utterances as the convex combination β of the degree of similarity on morphemes, α_d, and that on dependency relations 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.

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.

Eq. (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* respectively, these dependency relations are considered to be in correspondence. For example, two dependency relations (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 relation belongs to the same class, these relations are also considered to be in correspondence.

Morpheme similarity

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

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

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.

Calculation example

Figure 2 shows an example of the calculation of the degree of similarity between an input utterance S_1 "kono chikaku-ni washoku-no mise aru? (Is there a Japanese restaurant near here?)" and a sentence in a corpus, S_2, "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 dependency relations of S_1 and S_2 is 4 and 3, respectively, and that of dependency relations 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 dependency relations (washoku, aru) and (yoshoku, aru) also correspond with each other. Therefore, the degree of similarity in dependency α_d comes to 0.857 by the Eq. (2). Since the number of morphemes of S_1 and S_2 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 0.8 by Eq. (3). As mentioned above, β using both morphemes and dependencies comes to 0.823 by a Eq. (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 Eq. (4).

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

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 Eq. (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 Eq. (5) defines the degree of similarity γ using the weight assignment by intentions N-gram probability.

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

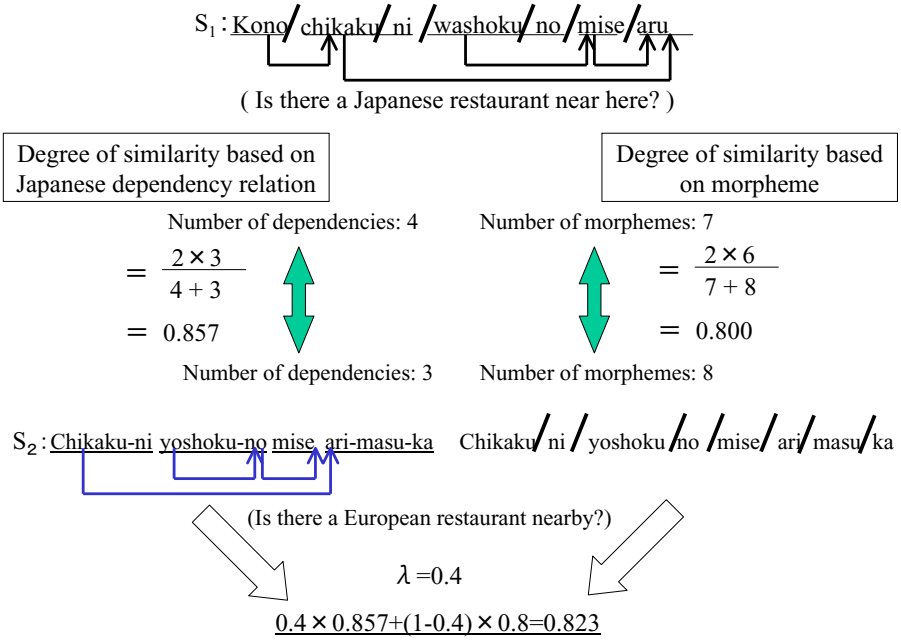

Figure 2: Example of similarity calculation

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. To 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 of 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.

Experimental data

An in-car speech dialogue corpus which has been constructed at CIAIR [4], was used as a corpus with intention tags, and analyzed based on Japanese dependency grammar [6]. 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

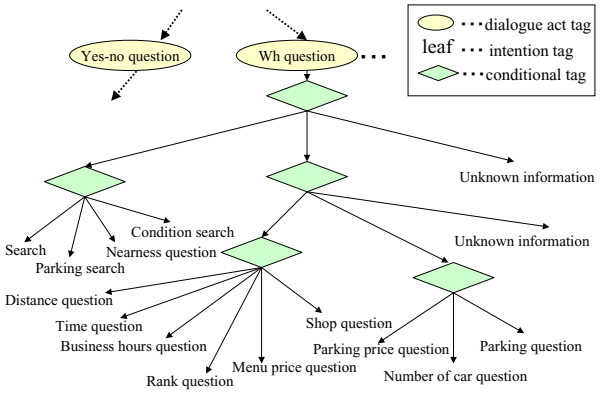

Figure 3: Decision tree of intention tag (a part)

from the sentences of 174 dialogues in them. The standard for assigning the intention tags was established by extending the decision tree proposed [2] as a dialogue tag scheme. 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 [8], which has been constructed based on the corpus, was used for calculating the rates of correspondence in morphemes and depen-

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

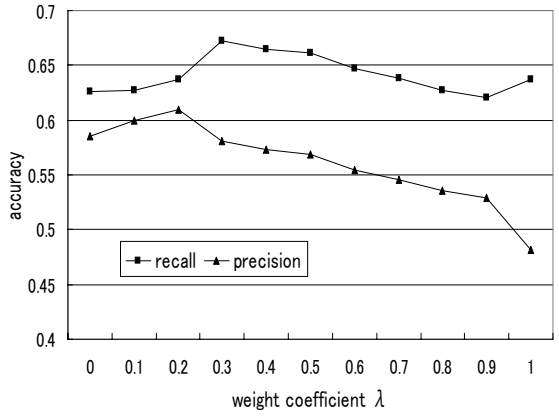

Figure 4: Relation between the weight coefficient λ and the accuracy

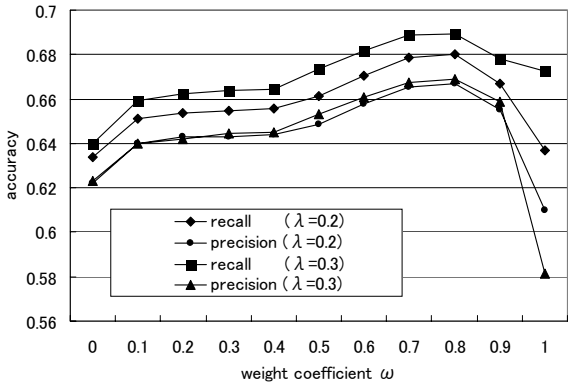

Figure 5: Relation between the weight coefficient ω and the accuracy

dencies. Moreover, Chasen [7] was used for the morphological analysis.

Experiment

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 denpendency relations** (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 $\theta = 0$)

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 more low than that of 1. This is because the user speaks with driving a car [3] and therefore there are much comparatively short utterances in the in-car speech corpus. Since there is a few dependency relations per 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 $\lambda = 0.2$ and 0.3, is shown in Figure 5. At $\omega = 0.8$, the maximum values in both precision and recall were provided. This shows our technique of learning the context information from the spoken dialogue corpus to be effective.

6 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 has shown the feasibility of example-based speech intention understanding and the effectiveness of our technique for the intention inference

Acknowledgement: The authors also would like to thank Dr. Hiroya Murao, Sanyo Electric Co. LTD. for his helpful advice. The analysis of dependency structures using the Japanese spoken language corpus has been carried out by all members of Spoken Language Processing Group in our laboratory. 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.

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?

References

[1] M. Araki, Y. Kimura, T. Nishimoto and Y. Niimi: "Development of a Machine Learnable Discourse Tagging Tool", *Proceedings of 2nd SIGdial Workshop on Discourse and Dialogue*, 2001, pp. 20–25.

[2] The Japanese Discouse Research Initiative JDRI: Japanese Dialogue Corpus of Multi-level Annotation, *Proceedings of 1st SIGdial Workshop on Discourse and Dialogue*, 2000.

[3] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura and Y. Inagaki: "Construction of Speech Corpus in Moving Car Environment", *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP-2000)*, Vol. III, 2000, pp. 362–365.

[4] N. Kawaguchi, S. Matsubara, K. Takeda and F. Itakura: "Multimedia Data Collection of In-car Speech Communication", *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, pp. 2027–2030 (2001).

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

[6] S. Matsubara, T. Sato, N. Kawaguchi and Y. Inagaki: "Robust Parsing of Spoken Language based on Statistical Dependencies", *IPSJ SIG Notes*, Vol. 2001, No. 54, 2001, pp. 63–68. (In Japanese)

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

[8] H. Murao, N. Kawaguchi, S. Matsubara and Y. Inagaki: "Example-based Query Generation for Spontaneous Speech" *Proceedings of 2001 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-2001)*, 2001.