CONCEPT-DRIVEN SPEECH UNDERSTANDING INCORPORATED WITH A STATISTIC LANGUAGE MODEL

Akito Nagai *Yasushi Ishikawa*

Information Technology R&D Center, MITSUBISHI Electric Corporation
5-1-1 OFUNA, KAMAKURA, KANAGAWA 247, JAPAN
E-mail: {nagai, yasushi}@media.isl.melco.co.jp

ABSTRACT

We have proposed a method of concept-driven semantic interpretation based on general semantic knowledge of conceptual dependency. In our approach, a concept is a unit of semantic interpretation and an utterance is regarded as a sequence of concepts that convey an intention. However, a considerable number of accepted results were not syntactically meaningful. This is because the order in which linguistic features occurred in the sequence of concepts was not taken into account in constructing the whole meaning from the concepts: only semantic constraint was used to attain linguistic robustness. Therefore, we introduce a statistical language model which calculates the plausibility of a sequence of concepts from the points of view of the order in which shallow linguistic features occur. Experimental results of speech understanding for 1000-word-vocabulary spontaneous speech show that the proposed method significantly improves the system performance.

1. INTRODUCTION

It is essential that a spoken dialog system can understand spontaneous speech so that it can be used easily by a naive user whose utterances comprise a large variety of expressions, which are often ill-formed [1, 2]. Thus, to improve speech recognition, such a system must have both linguistic robustness and adequate constraint. Our method for attaining linguistic robustness involves exploiting semantic knowledge so that it represents relations between phrases by semantic-driven processing. We have proposed a method of concept-driven semantic interpretation based on general semantic knowledge of conceptual dependency [3]. In our approach, a concept is a unit of semantic interpretation and an utterance is regarded as a sequence of concepts that convey an intention (Figure 1).

The performance of this method, however, was not adequate for practical use. A considerable number of accepted results were not syntactically meaningful, although they were semantically meaningful. This was because the order in which linguistic features occurred in the sequence of concepts was not taken into account in constructing the whole meaning from concepts. Only semantic constraint was used to attain linguistic robustness.

Therefore, we introduce a statistical language model which takes into account the occurrence order of linguistic features without eliminating linguistic robustness and generality. This model calculates the plausibility of the sequence of concepts from the point of view of shallow linguistic features. This paper discusses issues concerning the incorporation of the statistical language model of shallow linguistic features into the framework of semantic-driven speech understanding. It also reports empirical evaluation of the system's performance with spontaneous speech data concerning a sightseeing task with a 1000-word vocabulary.

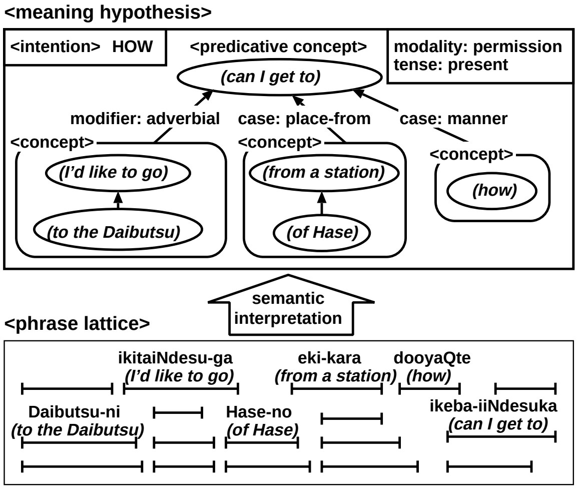

Figure 1: Concept-driven semantic interpretation.

2. LINGUISTIC FEATURES TO BE MODELED

When we express a meaning we have in our minds, various expressions of it can be produced as utterances. Among these utterances, there are some expressions which are rarely used, although they can be interpreted as the same meaning. In Japanese, it is considered that the order of phrases is basically free, but, for example, a phrase which includes a particle "wa" indicates the topic (topic case) and is often uttered at the beginning of an utterance. Such a topic case rarely occurs at the middle or at the end of an utterance. Other linguistic features like cases, attributive/adverbial modifiers and predicatives are also considered to occur according to basic principles of linguistic constraint on their order.

We utilize these basic principles of linguistic features as a constraint model in order to improve speech understanding. Our model evaluates linguistic cost according to, for example, whether a topic case is followed by an object case and whether that is usual or unusual in an utterance. Such a onstraint requires both a score which represents the plausibility of a sequence of linguistic features, and the acceptance of various expressions of spontaneous speech as knowledge sources. Therefore statistical approach is desirable. Moreover, to make the method task-independent, the units of constraint must be shallow linguistic features, rather than words which would make the method dependent on training corpus.

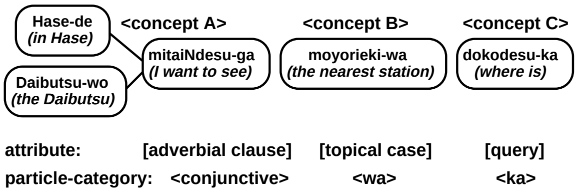

Figure 2: A sequence of particle-categories.

Therefore, we use *attributes* of the concept as the unit of the statistical model. The attributes are shallow linguistic properties that are classified, such as cases and adverbial clauses, and have a dominant role in constructing the global structure of an utterance. The attributes are decided by morphological information such as particles, parts of speech and conjugations. Attributes in Japanese are mainly represented by the *particle-category*, so we use the particle-category of the concept as the unit of the model which captures the sequence of the attributes approximately (Figure 2). The particle-category is a classification of function words, auxiliary verbs, and conjugations. Examples of correspondence between the attributes and the particle-categories are shown in Table 1.

Table 1: Examples of correspondence between the attributes and the particle-categories (partly).

Attributes	Particle-categories
topic case	"wa, Qte"
agent case	"ga"
object case	"wo"
source case	"kara"
goal case	"made"
place case	"ni, de"
conditional clause	"ba", "nara, tara"(auxiliary)
query	"ka"

3. STATISTIC LANGUAGE MODEL

We use N-gram (trigram) for modeling a sequence of the particle-categories of a concept. This model represents more global relations of linguistic features than that of word N-gram or phrase-based particle N-gram [5] because the particle-category trigram here is based on the concept which integrates phrases into a semantic unit as a dominant element for determining the meaning of an utterance.

During the training of the model, if a trigram probability is higher than a certain threshold, the trigram is given a constant value. This is because we want to extract trigram rules which have general plausibility in spontaneous speech, and to handle such trigram rules as they should have same probability. The particle-trigram model based on concept is defined in the following manner. Let $C_1, C_2, \ldots, C_n, \ldots, C_N$ be a sequence of particle-categories in an utterance which includes N concepts. The occurrence probability of a particle-trigram $P(C_n|C_{n-2}, C_{n-1})$ is trained by the formula;

$$P(C_n \mid C_{n-2},\ C_{n-1}) = \begin{cases} \frac{frequency(C_{n-2},\ C_{n-1},\ C_n)}{frequency(C_{n-2},\ C_{n-1})} & (< P_{threshold}) \\ Const. & (\geq P_{threshold}) \end{cases} \quad (1)$$

where $P_{threshold}$ $(0 < P_{threshold} < 1)$ is the threshold for extracting the trigram rule, and $Const.$ is the same probability which is given to all the trigram rules. Then, if we let $P(S)$ be the total occurrence probability of a whole utterance, and C_0, C_{N+1} be categories which mean the beginning and the end of the utterance, we have

$$P(S) = p(C_1 \mid C_0,\ C_0) \prod_{n=2}^{n=N+1} P(C_n) \qquad (2)$$

This probability is used in the process of speech understanding as the logarithmic likelihood for linguistic score S_{ngram} by

$$\begin{aligned} S_{ngram} &= -\log(P(S)) \\ &= -\log(p(C_1 \mid C_0,\ C_0)) - \sum_{n=2}^{n=N+1} \log(P(C_n)) \quad (3) \end{aligned}$$

4. TRAINING THE MODEL

The method for training the particle-category trigram and the text-based evaluation are described here.

4.1. Conditions

16 particle-categories are defined, including eleven particle-categories for representing cases ("wa, Qte, ga, wo, mo, ni, de, kara, made, e, ka"), three for representing adverbial clauses (<conditional>, <conjunctive-1>, <conjunctive-2>), one for representing the others (<else>) and one symbol ("-") for the beginning/end of an utterance. The particle-category trigram was trained with 1091 text data concerning a sightseeing dialog. The texts were manually segmented into units of a concept, then automatically tagged with the 16 particle-categories. The threshold for extracting the trigram rule is 0.05 to give the same probability of 1.0.

Table 2: Results of training the model.

variety of trigrams	375 kinds
#training samples	3955 samples
#possible trigrams	4096 (16^3)

Table 3: Examples of trigram rules extracted from a sequence of particle-categories (frequency).

<-/-/wa> (160), <wa/ka/-> (125), <-/-/de> (119), <else/ka/-> (115), <wo/else/-> (98), <-/-/else> (92), <-/-/conditional> (82), <else/else/-> (79), <-/-/Qte> (55), <-/-/wo> (54), <ga/ka/-> (44), <-/wa/ka> (42), <de/wo/else> (36), <wa/ni/ka> (22)
not found: <wa/made/mo>, <ga/ga/ka>, <wo/made/wa>, <ni/kara/de>

4.2. Results

Table 2 shows the results of training the model. The number of training samples was almost equal to the number of possible trigrams, so we think that the amount of training data is reasonable for extracting plausible trigram rules. Plausible trigram rules which occurred with high frequency and trigrams which were not found are shown in Table 3. This suggests that the model captured linguistic features of spoken dialog fairly successfully. To evaluate linguistic constraints of the model, linguistic scores for both the training corpus and for texts of understanding errors were calculated by equation (3) (the lower the score, the better). A score of 0.1-2.0 was given to the training corpus and a score of 4.5-13.1 was given to the errors. These results convinced us that reasonable constraints can distinguish between plausible texts and errors.

5. SPEECH UNDERSTANDING EXPERIMENTS

We have used this trigram model to carry out speech-understanding experiments for 1000-word-vocabulary spontaneous speech in human to machine communication.

5.1. Conditions

162 speech data from six males were used for the evaluation. These were collected in an office environment with an experimental spoken dialog system that was used for the sightseeing task. We told the subjects to ask questions freely. These speech data did not include types of utterance that consist of only one concept, for example, yes/no-responses or fragments like "ashita-desu (tomorrow)". In the process of speech recognition, phrase spotting used intra-phrase networks that had a vocabulary of 1005-word which included 23 filled-pauses. Speaker-independent syllable-HMMs were used. In the semantic interpretation, 102 concepts and 13 types of conceptual dependency were used. 40 linguistic penalty rules [4] for heuristic constraint were also used. The semantic interpretation outputs the understanding results as N-best meaning hypotheses.

A total score of each meaning hypothesis is calculated by combining three scores; acoustic likelihood of phrase spotting ($S_{acoustic}$), penalty score ($S_{penalty}$) and linguistic score of the particle-category trigram (S_{ngram}). We define this total score S_{total} by the following formula;

$$S_{total} = S_{acoustic} + W_1 * S_{penalty} + W_2 * S_{ngram} \quad (4)$$

where W_1 and W_2 are weighting parameters. We treats W_1 as an experimental constant value, because the penalty score is based on heuristic penalty rules and it represents the plausibility of the whole utterance. As for W_2, it is required to be normalized by the length of the utterance because the score S_{ngram} deteriorates as the utterance becomes longer. Then, normalization by the number of concepts n was done as follows;

$$W_2 = C/(n+1) \tag{5}$$

where C is a weighting constant.

5.2. Results

Figure 3 illustrates understanding error rates of the first rank, within the third and the tenth. Each error rate is the average of the six males. The standards for judging an answer to be correct are that intention, concepts and their boundaries, conceptual dependency, and semantic values of phrase candidates are correctly extracted. These results show that the particle-category trigram reduced the error rate of the first rank from 23.5% to 19.1% and halved it within the third and the tenth rank in comparison with the penalty score. It was also confirmed that this model was particularly helpful in suppressing errors of unreasonable inversion between cases, and unmatched dependencies between an adverbial clause and an attribute of a predicative concept.

As for the weighting parameters, we found that; (1) 0.2 for W_1 of the penalty score was the best, (2) the system performance deteriorated slightly with an increase in W_2 which was not normalized, but that the normalized W_2 canceled this degradation and performed better than W_2 which had not been normalized.

5.3. Discussion

By examining in detail the errors in the N-best outputs, we obtained results of error analysis which is listed in Table 4. The results show that two thirds of all errors are semantically plausible. This means that our method is likely to be promising to reasonably understand one utterance, although there still remains some errors to be suppressed. Thus, further improvement will focus on utilizing dialog context in speech understanding rather than solving the issues of an utterance, e.g., senseless coexistence among concepts.

Table 4: Types of understanding errors at a higher rank than that of a correct answer.

#errors	90	
plausible	61 (68%)	
inplausible	higher knowledge required	12 (13%)
	coexistence among concepts	17 (19%)

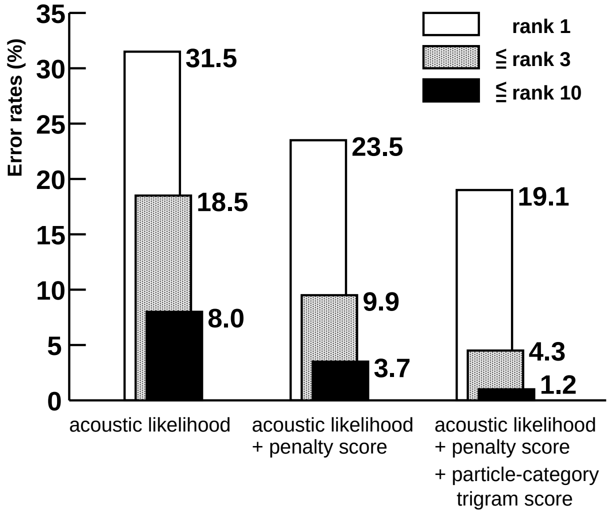

Figure 3: Understanding error rates (%).

6. CONCLUSION

We propose a statistical language model for capturing plausible sequences of linguistic features based on concept-driven speech understanding. Experimental results convinced us that this model is effective in attaining a high degree of accuracy in understanding spontaneous speech. Future works will include the use of information regarding dialog context for more precise understanding.

7. REFERENCES

1. Paolo Baggia and Claudio Rullent, “Partial Parsing as Robust Parsing Strategy,” Proc. ICASSP’93, Minneapolis (U.S.A.), pp. 123–126, Apr. 1993.
2. Wayne Ward and Sheryl R. Young, “Flexible Use of Semantic Constraints in Speech Recognition,” Proc. ICASSP’93, Minneapolis (U.S.A.), pp. 49–50, Apr. 1993.
3. Akito Nagai and Yasushi Ishikawa, “Speech Understanding Based on Integrating Concepts by Conceptual Dependency,” Proc. EuroSpeech’97, Rhodes (Greece), pp. 2747–2750, Sep. 1997.
4. Akito Nagai, Yasushi Ishikawa, and Kunio Nakajima, “Integration of Concept-Driven Semantic Interpretation with Speech Recognition,” Proc. ICASSP’96, Atlanta (U.S.A.), pp. 431–434, May 1996.
5. Ryosuke Isotani and Shigeki Sagayama, “Speech Recognition Using Particle N-grams and Content-Word N-grams,” Proc. EuroSpeech’93, Berlin (Germany), pp. 1955–1958, Sep. 1993.