A Linguistic Knowledge Base for Applying Semantic Information
to a Speech Understanding System

Naomi INOUE, Tsuyoshi MORIMOTO, Kentarou OGURA

ATR Interpreting Telephony Research Laboratories
Seika-cho, Soraku-gun, Kyoto, JAPAN

ABSTRACT
We have developed a knowledge base which describes the semantic co-occurrence between words and the inference mechanism on it. Using this knowledge base, it is possible to predict the next words likely to be uttered in the context. Also, this knowledge base is automatically created from ATR's linguistic database and can easily be constructed for a large vocabulary in a specific domain. This paper describes the structure of the knowledge base and the inference method.

1. INTRODUCTION
There are various problems in integrating speech recognition with natural language processing because the speech recognition system outputs various candidates for each speech input but, on the other hand, the natural language processing is difficult to treat a large number of word sequences for sentences. Therefore, the effort to reduce the number of sentence hypotheses from the speech recognition system must be made before the natural language processing. In this paper, the term "speech understanding" means acquiring the most likely sentence hypothesis among the candidates from the speech recognition system.

There have been many speech recognition systems using linguistic information, especially syntactic information, for example finite state grammars [1] or context free grammars [2],[3],[4]. However, syntactic restrictions only is not enough to reduce the ambiguity because there remain candidates which are syntactically, but not semantically, correct. Then, parsing approaches using caseframe as semantic information have been proposed. [5],[6] In these approaches, the caseframe parser, which had been developed for natural language processing, is modified to handle ambiguous speech output from the speech recognition system and select semantically more likely word sequences in many sentence hypotheses.

The basic idea of the caseframe is to define the semantic relationship between predicates and words modifying them ("CASE"). Caseframe parsing makes it possible to check out candidates which can not semi-technically co-occurs with predicates. The caseframe approach also uses a category to encode a set of words modifying predicates at an abstract level because it is hard work to extract all words modifying them. Either semantic features or thesauri is usually used in natural language processing as a category. However, it is very difficult to make universal categories because they are essentially cognitive.

ATR has constructed a large linguistic database which consists of a large number of conversational sentences and linguistic information attached to them, such as the semantic co-occurrence between words, the comparison of Japanese and English words, etc. [7] In this paper, we propose a more practical approach using the knowledge base which describes the semantic co-occurrence between words and its statistical information extracted from the above linguistic database.

2. NEED FOR STATISTICAL MEASURES
The following problem arises in integrating speech recognition with natural language processing.

• many hypotheses for one sentence
A speech recognition system outputs various candidates for a single phrase. Sentence hypotheses increase exponentially as the sentence becomes longer. The output from the speech recognition system probably include a large number of syntactically and semantically incorrect hypotheses.

Both caseframe and semantic categories have been used to check semantic co-occurrence. However, as shown below, they are not enough to reduce the ambiguity even in a specific domain. When the next Japanese sentence "Moushikomiyoushi shikyuu okuraseteitadakimasu" ("We will send you the registration form immediately.") is input into our Japanese phrase speech recognition system, as shown in Figure 1, the most likely hypotheses for each phrase are "moushikomiyoushi" ("registration form"), "shikyuu" ("immediately") and "kiraseteitadakimasu" ("send"). Here, italics indicate the inflection ("asete") and the polite expression ("itadakimasu") in Japanese and need not be considered. The correct candidate for the third phrase, "okuraseteitadakimasu" ("send"), is the third most likely. If "moushikomiyoushi" ("registration form") more frequently co-occurs with "okuraseteitadakimasu" ("send") than with "kiraseteitadakimasu" ("cut"), this is, in general, right in the domain "Secretary service of the international conference", then the frequency may contribute to give a better score to the sentence hypothesis "moushikomiyoushi shikyuu okuraseteitadakimasu" than "moushikomiyoushi shikyuu kiraseteitadakimasu". This can probably not be performed by the caseframe filler check, because "moushikomiyoushi" ("registration form") can be a caseframe filler

*1 The current ATR speech understanding scheme is applied to a Japanese phrase input not to continuous input. However, the problem seems to be applied to the continuous speech understanding system.

*2 Here, "semantically" means that the caseframe restriction check is right.

*3 In this paper, semantic category means both semantic features and thesauri.
for both "ekuraseteitadakimasu" ("send") and "kiraseteitadakimasu" ("cut").
Therefore, it is necessary to use a statistical measure in addition to the caseframe.

3. KNOWLEDGE BASE STRUCTURE

Our knowledge base describes the semantic relationship between words in the form of a semantic network with specified types of nodes and weighted links. The weight is used as a statistical measure. This semantic network is created from KAKARIUKE information described in ATR's linguistic database. KAKARIUKE includes CASE and the semantic relationship between nouns. An example of KAKARIUKE information is shown in Figure 2. The structure of the knowledge base is shown in Figure 3.

Sentence
"Kaigi(100) no(200) moushikomi(300) youshi(400) wo(500) okutte(600) kudasai(700)."
(“Please send me the registration form for the conference.”)

Here, the numbers are the ID numbers of the morphologies.

KAKARIUKE information
(1) (Object 400 600) : This notation means that one word (ID 400) modifies another word (ID 600) and that semantic relationship is "Object".
(2) (Purpose 100 300) : This notation means that one word (ID 100) modifies another word (ID 300) and that semantic relationship is "Purpose".

Figure 2 An Example of KAKARIUKE

3.1 Nodes
Nodes come hierarchically in the following three types: Event node, Class node and Primitive node. (They are usually seen in that order from top to bottom)
(1) Event node
Event nodes indicate an event, for example "present", "attend", etc. This node corresponds to predicates such as verb and adjective.
(2) Class node
Class nodes group the similar caseframe fillers. The label of class nodes are optional because they are introduced only to group castfillers. We present the label as "event node-label + semantic relationship", for example "present-object", "attend-object", etc.
(3) Primitive node
Primitive nodes correspond to actual words in the language. This node corresponds to words which belong to noun, adjective, verb and adverb. Although both verb and adjective are defined as event nodes, they are also defined as primitive nodes because event nodes and primitive nodes are utilized for different purposes as described later.

3.2 Links

There are two types of links: semantic link and isa-link.
(1) Semantic link
Semantic links indicate the semantic relationship between words. The semantic relationship can be divided into two categories: those between a primitive node and an event node and those between primitive nodes. The former corresponds to CASE. The latter includes Part of Whole (the leg of a dog), Object of Attribute (the water temperature), and so on.

A specific semantic relationship "predicate" between an event node and a primitive node which corresponds to an actual "predicate word" is defined in the knowledge base in order to handle the predicate word in the same way as castfillers.
(2) Isa link
Ias are links used to connect a class node and primitive nodes.

3.3 Weights
Both semantic-links and isa-links have weights. The weight indicates the strength of a semantic connection between two words and is determined according to the frequency of the co-occurrence in the linguistic database. As shown below, the weights of outward links, except the semantic relationship "predicate", from one node (ei) are normalized so that their total equals 1. The weight of the semantic link "predicate" and the isa-link below it is set to 1.

\[ \sum_{j} w_{ij} = 1 \]

4. INFERENCE METHOD

The construction of this knowledge base is based on the actual co-occurrence between words. Therefore, as the vocabulary becomes larger, the network becomes more complex. For this reason, the marker passing method is adopted as a parallel memory search mechanism. This mechanism considers related nodes only in a large and complex network. In addition, this approach can keep the context as the activating state in the memory established after understanding previous sentences. Related works are shown in Hirst, G.and et al. [8][9], Charniak, E. [10] and Tomabechi H.[11]. Our work is different from theirs in the way that weighted links are used. As regards weighted links and parallel searches, this inference method resembles the neural network associative mechanism except that inhibit links are not used.

4.1 Node Activation

Input sentence: "Moushikomiyoushibi shikyuu okuraseteitadakimasu." ("We will send you the registration form immediately")

Output from the speech recognition system:

moushikomiyoushibi 0.78 shikyuu 0.37 kiraseteitadakimasu 0.71
moushikomiyoushibi 0.69 shikyuu 0.06 kiraseteitadakimasu 0.12
moushikomiyoushibi 0.07 shikyuu 0.05 okuraseteitadakimasu 0.10
moushikomiyoushibi 0.06 shikyuu 0.02 okuraseteitadakimasu 0.07

Here, the number shows the speech recognition score which is directly proportional to the value.

Figure 1 An Example of Output from the Speech Recognition System
4.1.1 Principle

There are two types of inference process, activation and prediction. Activation corresponds to exciting the cell in the neural net. Prediction is introduced to check the semantic connection between two related words and to guide the search path to a more likely word as described in Section 4.2. For example, if the word ei modifies the word ej, the connection is shown in Figure 4. In this illustration, ei and ej correspond to primitive nodes in the knowledge base and wij is the weight of the link.

![Figure 4](image-url)

**Figure 4** The Image of the Node Connection

**Activation of ei**

If the word ei is recognized by the speech recognition system accompanied by the recognition score si (0 ≤ si ≤ 1), activation of the node ei is performed in equation (1).

\[
ai = F(si)
\]  

(1)

Here, ai is the activation score of the word ei and F(x) is introduced to transform the prediction score into a appropriate range. At present, F(x) = x is adopted on experiment.

**Prediction and Activation of ej**

After the word ei is activated, all words modified by the word ei are predicted and given the prediction score. The prediction score (pj) of the word ej is calculated as follows.

\[
pj = wij \times ai
\]  

(2)

After prediction, if the word ej is actually recognized, the activation score of the word ej is calculated in equation (3). In addition, the word ei is activated again from the word ej and the new score is calculated in equation (4).

\[
a^i = pj + sj \]  

(3)

\[
a^i = ai + wij \times aj
\]  

(4)

As a result, the activation of ei and ej becomes si + wij × sj and aj + wij × si respectively. This means that the activation of semantically related words are mutually made more active when they are recognized.

However, if the word ej is not recognized, the word ej is not activated even if the prediction score is high.

The prediction score must be canceled at the appropriate timing, for example at the end of a sentence or at a change of topic. At present, the prediction score is canceled after one sentence is treated.

4.1.2 Event node utilization

In the Japanese word order, the predicate word comes generally last, such as casefiller-casefiller-predicate, moreover the predicate word is sometimes omitted. As described in Section 4.2, when the beam search is achieved from left to right, the event node should be activated so that one casefiller can be predicted by previously recognized casefiller even if the predicate word of the event node is not recognized.

Given the following connection, ej is the event node and ei, ek and el are casefillers for different cases. In this illustration, class nodes are not displayed because they are easily understood.

![Figure 3](image-url)

**Figure 3** The Structure of the Knowledge Base

If the casefiller ei is recognized, the activation of event node ej is calculated in equation (5) and ek and el are predicted in equation (6).

\[
j_e = wij \times ai
\]  

(5)

\[
ph = wij \times aj (h = k, l)
\]  

(6)
If the word ek is actually recognized after ei is recognized, the activation score of ek and ej is calculated in equation (7) and (8) respectively. In addition, the word ei is predicted again in equation (9) and the word el is activated again in equation (10).

\[
\begin{align*}
    a_k' &= p_k + a_k \\
    a_j' &= s_j + w_{ij} \times a_k \\
    pl' &= w_{ij} \times a_j' \\
    a_i' &= p_i + w_{ij} \times a_k'.
\end{align*}
\]

(7) (8) (9) (10)

In this way, activation and prediction are repeated.

4.2 Beam search

The number of sentence hypotheses increases exponentially as the sentence becomes longer. Therefore, an effective procedure is required. Beam search from left to right is adopted because low score hypotheses can be discarded without backtracking.

Given the following output from the speech recognition system, the total number of search trees becomes n^m. Here, m and n indicate the number of phrases and candidates respectively.

\[
\begin{array}{ccc}
    e_1 & e_2 & e_m \\
    e_{11} & e_{21} & e_{m1} \\
    e_{12} & e_{22} & \cdots & e_{m2} \\
    e_{1n} & e_{2n} & \cdots & e_{mn}
\end{array}
\]

As for the above output, activation and prediction are performed repeatedly.

\[
\text{activation}(e_1) \rightarrow \text{prediction} \rightarrow \text{activation}(e_2) \rightarrow \cdots
\]

The score of a partial path is calculated as follows and the path with a low score is pruned. Here, ai j is the activation score of word ej.

\[
P(l,i) = \Sigma a_{ij} \quad (11)
\]

4.3 Control steps of this knowledge base

Two types of inference process, activation and prediction are performed by using A-marker and P-marker respectively. The inference in the knowledge base is achieved according to the following steps.

[step 1] When words are acquired from the speech recognition system, A-marker is put on primitive nodes.

[step 2] If the primitive nodes which get the A-marker have semantic links, they send the P-marker to connected nodes by semantic links. If they have no semantic links, no operation is performed.

[step 3] Primitive nodes which get the A-marker at [step 1] send the A-marker to upper class nodes in the network. This step continues before the A-marker reaches event nodes.

[step 4] When event nodes receive the A-marker, they send the P-marker to all class nodes below except those nodes from which they have already received the A-marker. Further, the A-marker is sent to those nodes from which have already received the A-marker.

[step 5] When class nodes receive the P-marker or the A-marker, they send it to primitive nodes below connected through isa-links.

Steps 1 - 5 are repeated whenever a candidate of the phrase is input to the knowledge base.

5. CONCLUSION

In this paper, we proposed a knowledge base which describes both the semantic relationship and its strength according to the frequency of the co-occurrence between words. Moreover, as the knowledge base is created from ATR's linguistic database, it is easy to extend to a large vocabulary and calculate the weight between words. The knowledge base is experimentally constructed on Symbolics 3600 and contains 833 primitive nodes, 625 class nodes and 183 event nodes.

However, syntactic rules between phrases are not utilized here. It is expected the combination of semantic and syntactic knowledge will make the speech understanding more accurate. We are currently working on finding a good way to combine this knowledge base and syntactic rules between phrases.

ACKNOWLEDGMENTS

The authors are deeply grateful to Dr. Akira Kurematsu, President of ATR Interpreting Telephony Research Laboratories and all the members of the Knowledge & Data Base Department for their continuous support and encouragement.

REFERENCES