Statistical Method of Recognizing Local Cohesion in Spoken Dialogues

Naoto Katoh and Tsuyoshi Morimoto
ATR Interpreting Telecommunications Research Laboratories
Seika-cho Soraku-gun Kyoto 619-02 Japan
{katonao, morimoto}@itl.atr.co.jp

Abstract

This paper presents a method for automatically recognizing local cohesion between utterances, which is one of the discourse structures in task-oriented spoken dialogues. More specifically we can automatically acquire discourse knowledge from an annotated corpus with local cohesion. In this paper we focus on speech act type-based local cohesion. The presented method consists of two steps 1) identifying the speech act expressions in an utterance and 2) calculating the plausibility of local cohesion between the speech act expressions by using the dialogue corpus annotated with local cohesion. We present two methods of interpolating the plausibility of local cohesion based on surface information on utterances. The presented method has obtained a 93% accuracy for closed data and a 78% accuracy for open data in recognizing a pair of utterances with local cohesion.

1 Introduction

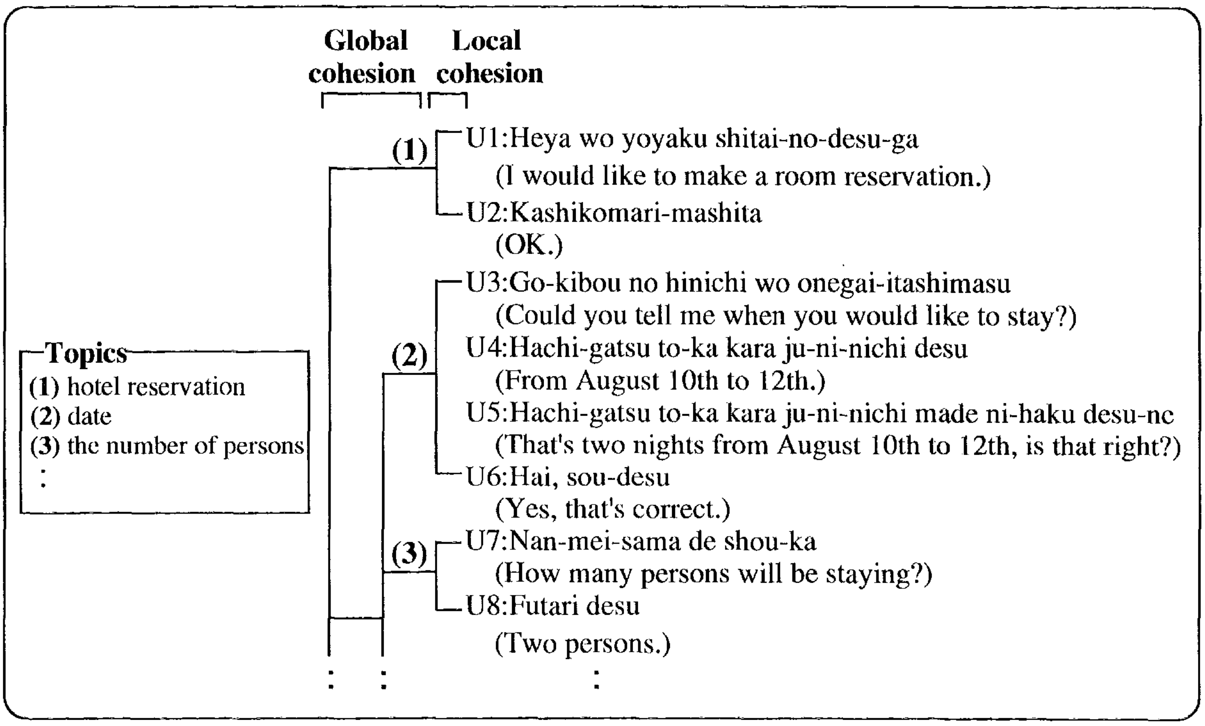
For ambiguity resolution, processing of a discourse structure is one of the important processes in Natural Language Processing (NLP). Indeed, discourse structures play a useful role in speech recognition, which is an application of NLP. In the case of Japanese, it is very difficult to recognize the end in utterances by using current speech recognition techniques because the sound power of an ending tends to be small. For example, "*desu*", which represents the speech act type "response", is often misrecognized as "*desu-ka* (question)" or "*desu-ne* (confirmation)". On the other hand, Japanese can easily select the adequate expression "*desu*", when the intention of the previous utterance is concerned with a question. This is because they use the coherence relation (local cohesion) between the two utterances, question-response.

In the conventional approach (i.e., rule-based approach) to processing the discourse structure [Hauptmann 88][Kudo 90][Yamaoka 91][Young 91], NLP engineers built discourse knowledge by hand-coding. However, the rule-based approach has a bottleneck in that it is a hard job to add discourse knowledge when the employed NLP system deals with a larger domain and more vocabulary.

Recently, statistical approaches have been attracting attention for their ability to acquire linguistic knowledge from a corpus. Compared with the above rule-based approach, a statistical approach is easy to apply to larger domains since the linguistic knowledge can be automatically extracted from the corpora concerned with the domain. However, little research has been reported in discourse processing [Nagata 94][Reithinger 95], while in the areas of morphological analysis and syntactic analysis, many research studies have been proposed in recent years.

This paper presents a method for automatically recognizing local cohesion between utterances, which is one of the discourse structures in task-oriented spoken dialogues. We can automatically acquire discourse knowledge from an annotated corpus with local cohesion. In this paper we focus on speech act type-based local cohesion. The presented method consists of two steps 1) identifying the speech act expressions in an utterance and 2) calculating the plausibility of local cohesion between the speech act expressions by using the dialogue corpus annotated with local cohesion. We present two methods of interpolating the plausibility of local cohesion based on surface information on utterances. Our method has obtained a 93% accuracy for closed data and a 78% accuracy for open data in recognizing a pair of utterances with local cohesion.

In Section 2, local cohesion in task-oriented dialogues is described. In Section 3, our statistical method is presented. In Section 4, the results of a series of experiments using our method are described.



U1:Heya wo yoyaku shitai-no-desu-ga
(I would like to make a room reservation.)
U2:Kashikomari-mashita
(OK.)
U3:Go-kibou no hinichi wo onegai-itashimasu
(Could you tell me when you would like to stay?)
U4:Hachi-gatsu to-ka kara ju-ni-nichi desu
(From August 10th to 12th.)
U5:Hachi-gatsu to-ka kara ju-ni-nichi made ni-haku desu-nc
(That's two nights from August 10th to 12th, is that right?)
U6:Hai, sou-desu
(Yes, that's correct.)
U7:Nan-mei-sama de shou-ka
(How many persons will be staying?)
U8:Futari desu
(Two persons.)

Figure 1. An example of a task-oriented dialogue in Japanese

2 Local Cohesion between Utterances

The discourse structure in task-oriented dialogues has two types of cohesion: global cohesion and local cohesion. Global cohesion is a top-down structured context and is based on a hierarchy of topics led by domain (e.g., hotel reservation or flight cancellation). Using this cohesion, a task-oriented dialogue is segmented into several subdialogues according to the topic. On the other hand, local cohesion is a bottom-up structured context and a coherence relation between utterances, such as question-response or response-confirmation. Different from global cohesion, local cohesion does not have a hierarchy. This paper focuses on local cohesion.

Figure 1 shows a Japanese conversation between a person and a hotel staff member, which is an example of a task-oriented dialogue; The person is making a hotel reservation. The first column represents global cohesion and the second column represents local cohesion. For example, the pair of U3 and U4 has local cohesion, because it has a coherence relation for each word in the utterances as follows:

c1) speech act pattern between "*onegai-itashimasu* (requirement)"in U3 and "*desu* (response)" in U4

c2) semantic coherence between nouns, "*hinichi* (date)"in U3, and "*hachi-gatsu to-ka* (August 10th)" and "*ju-ni-nichi* (12th)"in U4

In the same way, (U4, U5) and (U5, U6) have local cohesion. Thus, U3 to U6 are built up as one structure and form a subdialogue with the topic "date".

As observed from this example, whether two utterances have local cohesion with one another or not is determined by coherence relations between the speech act types in the utterances, coherence relations between the verbs in them and coherence relations between the nouns in them. In recognizing local cohesion, our method uses these three coherence relations.

3 Our Approach

3.1 Utterance Model with Local Cohesion

In this paper, we approximate an utterance in a dialogue to a three-tuple;

$$U = (SPEECH\_ACT, VERB, NOUNS) \quad (1)$$

where $SPEECH\_ACT$ is the speech act type, $VERB$ is the main verb in the utterance, and $NOUNS$ is a set of nouns in the utterance (e.g., a subject noun and an object noun for the main verb). Figure 2 shows a dialogue with our utterance model.

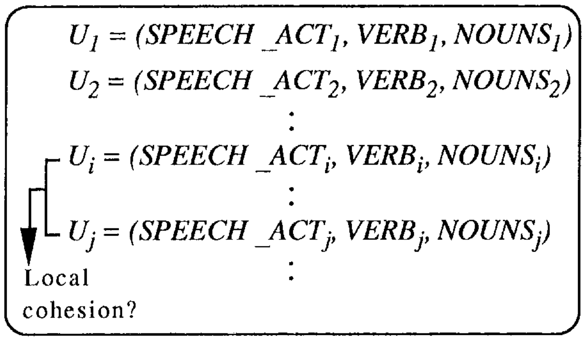

$U_1 = (SPEECH\_ACT_1, VERB_1, NOUNS_1)$
$U_2 = (SPEECH\_ACT_2, VERB_2, NOUNS_2)$
$U_i = (SPEECH\_ACT_i, VERB_i, NOUNS_i)$
$U_j = (SPEECH\_ACT_j, VERB_j, NOUNS_j)$
Local cohesion?

Figure 2. A dialogue with our utterance model

As mentioned in Section 2, when the *i*th utterance $U_i (1 \leq i \leq j-1)$ and the *j*th utterance U_j in a dialogue have local cohesion with one another, $(SPEECH\_ACT_i, SPEECH\_ACT_j)$, $(VERB_i, VERB_j)$ and $(NOUNS_i, NOUNS_j)$ have coherence relations. Therefore, the plausibility of local cohesion between U_i and U_j can be formally defined as:

$$\begin{aligned} &cohesion\_local(U_i, U_j) \\ &= \lambda_1 cohesion\_speechact(SPEECH\_ACT_i, SPEECH\_ACT_j) \\ &+ \lambda_2 cohesion\_verb(VERB_i, VERB_j) \\ &+ \lambda_3 cohesion\_noun(NOUNS_i, NOUNS_j) \qquad (2) \end{aligned}$$

where λ_1, λ_2 and λ_3 $(\lambda_1 + \lambda_2 + \lambda_3 = 1)$ are nonnegative weights contributing to local cohesion, and *cohesion_speechact*, *cohesion_verb* and *cohesion_noun* are functions giving the plausibility of coherence relations between speech act types, verbs and nouns respectively. The problem of deciding an utterance that has local cohesion with U_j can then be formally defined as finding a utterance with the highest plausibility of local cohesion for U_j, which is the result of the following function:

$$U_{opt}{}^{j} = \arg\max_{0 \leq i \leq j-1} cohesion(U_i, U_j) \qquad (3)$$

As the first step, this paper uses only speech act types in the calculation (i.e. $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = 0$). This is because the speech act types are more powerful in finding local cohesion than the verbs or the nouns as follows:

r1) The speech act types are independent of domain.

r2) The speech act types are stable, while the nouns and the verbs are sometimes omitted in utterances in spoken dialogues.

Thus, Equation (2) is reduced to:

$$\begin{aligned} &cohesion\_local(U_i, U_j) \\ &= cohesion\_speechact(SPEECH\_ACT_i, SPEECH\_ACT_j) \qquad (4) \end{aligned}$$

In order to calculate Equation (4), two kinds of information to answer the following questions are required as discourse knowledge:

q1) What expressions in an utterance indicates a speech act type?

q2) What speech act patterns have local cohesion?

We automatically acquire these discourse knowledge from a annotated corpus with local cohesion. According to the information, our method is composed of two processes, 1) identifying the expressions which indicate a speech act type (called speech act expressions) in an utterance and 2) calculating the plausibility of the speech act patterns by using the dialogue corpus annotated with local cohesion.

3.2 Identifying Speech Act Expressions in an Utterance

The first process in our method identifies the speech act expression in each of the utterances by matching the longest pattern with the words in a set of speech act expressions. The words can be collected by automatically extracting fixed expressions from the parts at the end of utterances, because the speech act expressions in Japanese have two features as follows:

f1) The speech act expression forms fixed patterns.

f2) The speech act expressions lie on the parts at the end of the utterance. For example, "*desu*" in "*Futari desu*"(U8) in Figure 1 represents a speech act type "response". We call these expressions ENDEXPR expressions.

For the automatic extraction, we use a slight modification of the cost criteria-based method [Kita 94], which uses the product of frequency and length of expressions, because this is easy when dealing with languages that do not use delimiters in their writings such as Japanese. Kita et al. extract expressions in the order of larger cost criteria values. We do so in the order of longer fixed-expressions with cost-criteria values above a certain threshold. For the more details of our extraction method, see [Katoh 95].

When the ENDEXPR expressions, which are listed in the set of the speech act expressions (represented as Set-ENDEXPR), are defined as a symbol *ENDEXPR*, we can approximate the speech act types as *SPEECH_ACT=ENDEXPR*. Thus, Equation (4) is transformed to:

$$\begin{aligned} &cohesion\_local(U_i, U_j) \\ &= cohesion\_speechact(SPEECH\_ACT_i, SPEECH\_ACT_j) \\ &= cohesion\_endexpr(ENDEXPR_i, ENDEXPR_j) \qquad (5) \end{aligned}$$

where *cohesion_endexpr* is a function giving the plausibility of coherence relations between the ENDEXPR expressions.

3.3 Calculating the Plausibility of Local Cohesion

The second process is to calculate the plausibility of local cohesion between utterances from the dialogue corpus using a statistical method. In this paper, we define the plausibility of local cohesion, i.e., Equation (5), as follows:

$$cohesion\_local(U_i, U_j)$$
$$= cohesion\_endexpr(ENDEXPR_i, ENDEXPR_j)$$
$$= f(ENDEXPR_i, ENDEXPR_j) - \bar{f}(ENDEXPR_i, ENDEXPR_j) \quad (6)$$

where

$$f(ENDEXPR_i, ENDEXPR_j) = P(ENDEXPR_i \cap ENDEXPR_j) \times \log \frac{P(ENDEXPR_i \cap ENDEXPR_j)}{P(ENDEXPR_i) \times P(ENDEXPR_j)}$$

$$\bar{f}(ENDEXPR_i, ENDEXPR_j) = \bar{P}(ENDEXPR_i \cap ENDEXPR_j) \times \log \frac{\bar{P}(ENDEXPR_i \cap ENDEXPR_j)}{P(ENDEXPR_i) \times P(ENDEXPR_j)}$$

The f and $\bar{f}$ are modified functions of mutual information in Information Theory. We call them pseudo-mutual information. $P(\cdot)$ is the relative frequency of two utterances with local cohesion and $\bar{P}(\cdot)$ is that of two utterances without local cohesion. $ENDEXPR_i \cap ENDEXPR_j$ means that $ENDEXPR_j$ appears next to $ENDEXPR_i$. We call a series of two ENDEXPR expressions (i.e., $ENDEXPR_i \cap ENDEXPR_{i+1}$) an ENDEXPR bigram.

The larger the value of Equation (6) in two utterances gets, the more plausible the local cohesion between them becomes. For example, in applications in speech recognition, the optimal result has the largest plausibility value among several candidates obtained from a module of speech pattern recognition.

3.4 Smoothing Methods

Although the statistical approach is easy to implement, it has a major problem, i.e., sparse data problem. Indeed Equation (6) gives very small values in some cases. In order to overcome this problem (to interpolate the plausibility), we propose two smoothing techniques.

[Smoothing Method 1]

Interpolate the plausibility by using partial fixed expressions in Set-ENDEXPR. For example, in Japanese, "*itadake-masu-ka*" can be segmented into the smaller morpheme "*masu-ka*" or "*ka*", and the original morpheme "*itadake-masu-ka*" is interpolated by these two morphemes as follows:

$$cohesion\_endexpr(\text{"itadake-masu-ka"}, \text{"desu"})$$
$$= \mu_0 cohesion\_endexpr(\text{"itadake-masu-ka"}, \text{"desu"})$$
$$+ \mu_{21} cohesion\_endexpr(\text{"masu-ka"}, \text{"desu"})$$
$$+ \mu_{31} cohesion\_endexpr(\text{"ka"}, \text{"desu"})$$

where μ_0, μ_{21} and μ_{31}($\mu_0 + \mu_{21} + \mu_{31} = 1$) are nonnegative parameters.

Formally, if we assume the partial fixed expressions $ENDEXPR_i^p \in$ Set-ENDEXPR and $ENDEXPR_j^q \in$ Set-ENDEXPR, and represent the "smaller" as a symbol "<" (e.g., "*ka*"<"*masu-ka*"<"*itadake-masu-ka*"), we can interpolate the original plausibility by using these smaller morphemes:

$$cohesion\_endexpr(ENDEXPR_i, ENDEXPR_j)$$
$$= \mu_0 cohesion\_endexpr(ENDEXPR_i, ENDEXPR_j)$$
$$+ \sum_{\substack{1 \le p \le m \\ 1 \le q \le n}} \mu_{pq} cohesion\_endexpr(ENDEXPR_i, ENDEXPR_j)$$

where

$$ENDEXPR_i^1 < ENDEXPR_i^2 < \ldots < ENDEXPR_i^m$$
$$ENDEXPR_j^1 < ENDEXPR_j^2 < \ldots < ENDEXPR_j^n$$

[Smoothing Method 2]

Interpolate the plausibility by using the speech act types themselves. For example, "*itadake-masu-ka* (requirement)" and "*desu* (response)" are interpolated by the relation of the speech act types (not the speech act expressions), i.e., "requirement-response", as follows:

$$cohesion\_endexpr(\text{"itadake-masu-ka"}, \text{"desu"})$$
$$= \mu_0 cohesion\_endexpr(\text{"itadake-masu-ka"}, \text{"desu"})$$
$$+ \mu_1 cohesion\_speechact\_type(requirement, response)$$

where μ_0 and μ_1 ($\mu_0 + \mu_1 = 1$) are nonnegative parameters.

These speech act types are automatically constructed by clustering *ENDEXPR*s based on *ENDEXPR* bigrams, and then the type bigrams are re-calculated from the *ENDEXPR* bigrams.

Formally, when the speech act types are denoted by *SACT_TYPE*, we can interpolate an original plausibility by using these type patterns:

$$
\begin{aligned}
&\mathit{cohesion\_endexpr}(\mathit{ENDEXPR}_i, \mathit{ENDEXPR}_j)\\
&= \mu_0 \mathit{cohesion\_endexpr}(\mathit{ENDEXPR}_i, \mathit{ENDEXPR}_j)\\
&+ \mu_1 \mathit{cohesion\_speechact\_type}(\mathit{SACT\_TYPE}_a,\\
&\qquad\qquad \mathit{SACT\_TYPE}_b)
\end{aligned}
$$

where

$$\mathit{ENDEXPR}_i \in \mathit{SACT\_TYPE}_a,$$

$$\mathit{ENDEXPR}_j \in \mathit{SACT\_TYPE}_b$$

cohesion_speechact_type is a function giving the plausibility of coherence relations between the speech act types.

The former method must use the $n \times m$ parameters (i.e., μ_{pq}) and the latter one must produce the speech act types. We chose the former method for our first experiments, because it was easier to implement.

4 Experiments

A series of experiments was done to evaluate our method. The experiments were carried out to decide whether one utterance and the next one have local cohesion with each other or not. The results of the experiment were able to segment a dialogue into several subdialogues.

First, Set-ENDEXPR was constructed automatically by our extraction method for fixed expressions from the ATR speech and language database [Morimoto 94]. The database includes about 600 task-oriented dialogues concerned with several domains, such as hotel reservation, flight cancellation and so on and each of the dialogues includes about 50 utterances on average. We have extracted about one hundred fixed-expressions from the database by the extraction method. Table 1 shows examples of the fixed expressions.

Table 1. Examples of results by the extraction method

the end of expressions (speech act type)
arigatou-gozai-mashita (thank-you)
itadake-masu-ka (requirement)
onegai-shimasu (requirement)
de-shou-ka (question)
masu-ka (question)
desu-ka (question)
su-ka (question)
desu (response)
ka (question)

Secondly, we chose 60 dialogues from the ATR database and annotated them with local cohesion by hand-code such as that shown in Figure 1 in Section 2. Then, six dialogues were taken from the 60 dialogues to use in testing for the open data, and the rest of the dialogues (54 dialogues) were used to calculate *ENDEXPR* bigrams. Moreover, six dialogues were taken from the 54 dialogues for the closed data. Using the 54 dialogues, the *ENDEXPR* bigrams were produced in the following four parts:

An *ENDEXPR* bigram
part A : with local cohesion + turn-taking.
part B : with local cohesion + no turn-taking.
part C : without local cohesion + turn-taking.
part D : without local cohesion + no turn-taking.

where "turn-taking" means that an utterance and the next one are produced by different persons and "no turn-taking" means that they are done by the same person.

Table 2 shows examples of *ENDEXPR* bigrams with local cohesion.

Table 2a. Examples of *ENDEXPR* bigrams with local cohesion + turn-taking (part A)

relative frequency	*ENDPXPR* bigrams
.01307	*wo-onegai -shimasu kashikomari-mashita*
.01226	*deshou-ka desu*
.01144	*de-onegai-shimasu kashikomari-mashita*
.00735	*desu-ka desu*
.00735	*desu de-gozai-masu-ne*

Table 2b. Examples of *ENDEXPR* bigrams with local cohesion + no turn-taking (part B)

relative frequency	*ENDPXPR* bigrams
.03063	*desu desu*
.01685	*imasu desu*
.01225	*shoushou-omachi-kudasai omatase-itashimashita*
.01225	*hai desu*
.01072	*desu-ne desu-ne*

Using these *ENDEXPR* bigrams, a series of experiments was done for the closed data. and the open data. In the experiments, we defined the two utterances with local cohesion, if they had the plausibility above a certain threshold, and we chose Smoothing Method 1 under the condition that μ_0=0.7, $\sum_{\substack{1\le p\le m\\ 1\le q\le n}} \mu_{pq}$=0.7 and $\mu_{pq}=\frac{1}{m \times n}$.

Table 3 shows the accuracy of recognizing local cohesion in three cases: the turn-taking case, the no turn-taking case and the total case.

Table 3. The accuracy of recognizing local cohesion

data	method	turn-taking	no turn-taking	total
closed data	Our method without smoothing	90.6%	93.9%	92.3%
	Our method with smoothing	93.6%	93.9%	93.8%
	Default method	74.8%	54.2%	64.7%
open data	Our method without smoothing	77.4%	70.9%	73.9%
	Our method with smoothing	81.8%	75.5%	78.4%
	Default method	76.6%	50.4%	65.3%

In Table 3, the "default method" assumed that all of the pairs of utterances in a dialogue has local cohesion, and the accuracy was calculated as:

The accuracy in the "default method"

$$= \frac{\text{The number of the pairs of utterances with local cohesion}}{\text{The total number of the pairs of utterances in a dialogue}}$$

As shown in Table 3, the accuracy of our method was higher than that of the "default method". Using the smoothing method, we obtained a 93.8% accuracy for the closed data and about a 78.4% accuracy for the open data.

5 Conclusion

We described our statistical method of recognizing local cohesion between utterances, based on pseudo-mutual information. We focused on speech act expressions in calculating the plausibility of local cohesion between two utterances. The results of the first experiments showed a 93% accuracy for closed data and a 78% accuracy for open data. We conclude that the presented method will be effective for recognizing local cohesion between two utterances.

To improve our method, we will use the coherence relations in the verbs and set of nouns and will use a larger corpus with local cohesion. We plan to apply the method to speech recognition in a speech-to-speech machine translation system.

References

[Hauptmann 88] Hauptmann, A.G. et al.: Using Dialog-Level Knowledge Sources to Improve Speech Recognition, Proceedings of AAAI-88, pp. 729-733, 1988.

[Kudo 90] Kudo, I: Local Cohesive Knowledge for a Dialogue-machine Translation System, Proceedings of COLING-90, Helsinki, Vol. 3, pp. 391-393, 1990.

[Katoh 95] Katoh, N. and Morimoto, T.: Statistical Approach to Discourse Processing, Proceedings of SIG-SLUD-9502-3 of JSAI, pp. 16-23, 1995. (in Japanese)

[Kita 94] Kita, K et al.: Application of Corpora in Second Language Learning - The Problem of Collocational Knowledge Acquisition, Proceedings of WVLC-94, pp. 43-56,1994.

[Nagata 94] Nagata, M. and Morimoto, T.: An Information-Theoretic Model of Discourse for Next Utterance Type Prediction, Trans. of IPSJ, Vol. 35, No. 6, pp. 1050-1061, 1994.

[Reithinger 95] Reithinger N. and Maier E.: Utilizing Statistical Dialogue Act Processing in Verbmobil, ACL-95, pp. 116-121, 1995.

[Yamaoka 91] Yamaoka, T. and Iida, H.: Dialogue Interpretation Model and its Application to Next Utterance Prediction for Spoken Language Processing, EUROSPEECH-91, pp. 849-852, 1991.

[Young 91] Young, S. and Matessa, Y.: Using Pragmatic and Semantic Knowledge to Correct Parsing of Spoken Language Utterances, Proceedings of EUROSPEECH-91, pp. 223-227, 1991.