Cue Phrase Selection in Instruction Dialogue Using Machine Learning

Yukiko I. Nakano
NTT Information and Communication Systems Laboratories
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa 239-0847 Japan
yukiko@nttnly.isl.ntt.co.jp

Tsuneaki Kato
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
kato@cslab.kecl.ntt.co.jp

Abstract

The purpose of this paper is to identify effective factors for selecting discourse organization cue phrases in instruction dialogue that signal changes in discourse structure such as topic shifts and attentional state changes. By using a machine learning technique, a variety of features concerning discourse structure, task structure, and dialogue context are examined in terms of their effectiveness and the best set of learning features is identified. Our result reveals that, in addition to discourse structure, already identified in previous studies, task structure and dialogue context play an important role. Moreover, an evaluation using a large dialogue corpus shows the utility of applying machine learning techniques to cue phrase selection.

1 Introduction

Cue phrases are words and phrases, such as "first", "and", "now", that connect discourse spans and add structure to the discourse both in text and dialogue. They signal topic shifts and changes in attentional state (Grosz and Sidner, 1986) as well as expressing the relation between the individual units of discourse (Moore, 1995; Rösner and Stede, 1992). In this study, we focus on the former kind of cue phrases, organization cue phrases that signal the structural organization of discourse.

In instruction dialogue, the organization cue phrases play a crucial role in controlling dialogue and making the material easy to understand. Moreover, in dialogue systems, the user cannot comprehend the structural organization of the dialogue unless the appropriate cue phrases are included in the system's utterances. Therefore, for dialogue generation, we must identify the determining factors of organization cue phrases and select the cue phrases appropriately.

In previous studies that have investigated the relationship between cue phrases and the types of structural change (e.g. pop, push), the taxonomies of cue phrases have been presented (Grosz and Sidner, 1986; Cohen, 1984; Schiffrin, 1987). These taxonomies are, however, not sufficient for generation because the correspondence between cue phrase and structural change is many-to-many quite often. For example, "now","and", and "next" are all classified as the category signaling push in attentional state. Therefore, the indication of structural shifts in discourse is not sufficient to fully constrain cue phrase selection.

In this study, we reveal what factors affect organization cue phrase selection, and establish more precise selection rules for generating instruction dialogues. As factors for cue phrase selection, we examine a variety of features concerning discourse structure, task structure, and dialogue context. The reason that we examine these three factors is as follows. First, discourse structure is indispensable for selecting cue phrase as claimed in previous studies (Grosz and Sidner, 1986; Cohen, 1984; Eugenio et al., 1997). We examine some features concerning this factor such as the global structure of discourse and structural shifts in discourse. Second, while the discourse structure provides information about the preceding discourse, Cawsey (1993) claimed that information about the succeeding discourse (e.g., length and complexity) is also necessary in order to select cue phrases dynamically in dialogue systems. From this point of view, task structure is expected to be effective because discourse structure strongly reflects task structure in task oriented dialogue (Grosz, 1977; Guindon, 1986). Finally, in contrast to these structural aspects of dialogue, we think it important to consider sequential contexts of dialogue such as the types of dialogue exchange (Stenström, 1994) immediately preceding to the cue phrase.

In this paper, using a machine learning technique, C4.5 (Quinlan, 1993), we examine these features in terms of their effectiveness in selecting organization cue phrase and identify the most effective set of learning features. In addition, we evaluate the accuracy of decision trees obtained using a large corpus.

Our result reveals that, in addition to discourse structure whose effectiveness has already revealed in previous studies, task structure and dialogue context play important roles. Especially important are the place of the segment in the global structure of the dialogue and the type of the immediately preceding dialogue exchange.

The organization of this paper is as follows. Section 2 discusses related work. Section 3 mentions the annotation of our dialogue corpus while section 4 details the learning experiment and its results are discussed. Section 5 refers to further work and concludes this paper.

2 Related work

While cue phrases can appear in different places in instruction dialogues, we focus on the organization cue phrases that occur at the beginning of discourse segments referring to goals or direct actions. This is because such kind of cue phrases have the important function of describing the basic structure of the dialogue. In a procedural instruction dialogue, the sequence of actions for the procedure is directed step by step. In terms of Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), it is considered that the basic structure of such kind of discourse is constructed by connecting segments that refer to goals or primitive actions with "sequence" relation (Rösner and Stede, 1992; Kosseim and Lapalme, 1994). Therefore, the cue phrases which occur at the beginning of segments that are connected with "sequence" relation and refer to goals or direct actions play important roles in signaling the basic structure of the dialogue. Moreover, such kind of cue phrases are observed very frequently in instruction dialogues. In their empirical study on the characteristics of task oriented dialogues, Oviatt and Cohen (1990) reported that, in instruction dialogues on assembling a pump, cue phrases such as "Okay", "now" and "next" occur at the beginning of 98.6% of the new segments that instruct assembly actions in telephone dialogues. Based on the above, we think it important for dialogue generation to select and set appropriate cue phrases at the beginning of discourse segments that refer to goals or direct actions.

Moser and Moore (1995a) and Moser and Moore (1995b) investigated the relationship between cue placement and selection. They showed that the cue phrases are selected and distinguished depending on their placement. Somewhat differently, we tackle the problem of selecting cue phrases that occur at the same place in the segment (at the beginning of the segment).

As indicated in (Eugenio et al., 1997), in terms of natural language generation, cue usage consists of three problems, *occurrence*: whether or not a cue should be included, *placement*: where the cue should be placed, and *selection*: what cue should be used. We tackle the third problem, the selection of cue phrases. Our final goal is to establish a strategy for selecting organization cue phrases and apply it in the generation of instruction dialogues. While the empirical approach of this study is close to that of (Eugenio et al., 1997), they apply a machine learning technique to predicating cue occurrence and placement, not cue phrase selection.

3 Annotation of dialogue corpus

In this section, we mention the way of the annotation in our corpus. Then, the inter-coder agreement for the annotations is discussed.

3.1 Class of cue phrases

The domain of our dialogue corpus in Japanese is to instruct the initial setting of an answering machine. The corpus consists of nine dialogues with 5,855 utterances. There are, 1,117 cue phrases in 96 distinct

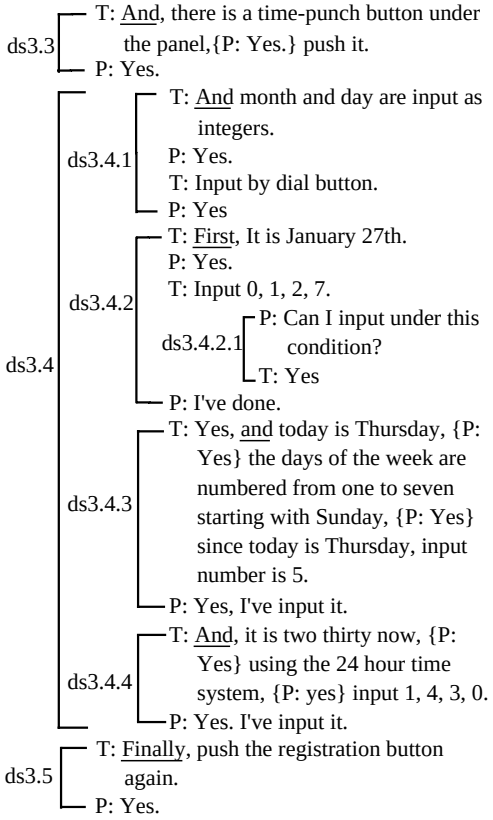

Figure 1: An example of annotated dialogue

cues[1]. There are 31 cue phrases that occur more than five times.

As the result of classifying these 31 cue phrases based on the classification of Japanese connectives (Ichikawa, 1978; Moriyama, 1997) and cue phrase classification in English (Grosz and Sidner, 1986; Cohen, 1984; Knott and Dale, 1994; Moser and Moore, 1995b), 20 cue phrases, which occurred total of 848 times, were classified into three classes: *changeover*, such as *soredeha, deha* ("now", "now then" in English), *conjunctive*, such as *sorede, de* ("and", "and then"), and *ordinal* , such as *mazu, tsugini* ("first", "next"). Besides these simple cue phrases, there are composite cue phrases such as *soredeha-tsugini* ("now first"). Note that meaning and the usage of each of these Japanese cue phrases does not completely correspond to those of the English words and phrases in parentheses. For example, the meaning of the Japanese cue phrase *soredeha* is close to the English word *now* in its discourse sense. However, *soredeha* does not have a sentential sense though *now* does.

The purpose of this study is to decide which of these three classes of simple cue phrases should be selected as the cue phrase at the beginning of a dis-

[1] Cue phrases which occur in the middle of the segment and in the segment other than action direction such as clarification segment are included.

course segment. We do not deal with composite types of cue phrases.

3.2 Annotation of discourse structure

As the basis for examining the relationship between cue phrase and dialogue structure, discourse segment boundary and the level of embedding of the segments were annotated in each dialogue. We define discourse segment (or simply segment) as chunks of utterances that have a coherent goal (Grosz and Sidner, 1986; Nakatani et al., 1995; Passonneau and Litman, 1997). The annotation of hierarchical relations among segments was based on (Nakatani et al., 1995).

Figure 1 shows an example from the annotated dialogue corpus. This dialogue was translated from the original Japanese. This example provides instruction on setting the calendar and clock of the answering machine. The purpose of ds3.4 is to input numbers by dial buttons and each input action is directed in ds3.4.2, ds3.4.3, and ds3.4.4, for inputting the date, the day of the week, and the time, respectively. Subdialogues such as confirmation and pupil initiative clarification are treated as one segment as in ds3.4.2.1. The organization cue phrases are underlined in the sample dialogue. For example, the cue phrase for ds3.3 is "And", and that for ds3.5 is "Finally" [2].

3.3 Annotation of discourse purpose and pre-exchange

As the information about task structure and dialogue context, we annotated the discourse purpose of each segment and the dialogue exchange at the end of the immediately preceding segment.

In annotating the discourse purpose, the coders selected the purpose of each segment from a topic list. The topic list consists of 127 topics. It has a hierarchical structure and represents the task structure of the domain of our corpus. When the discourse purpose cannot be selected from the topic list, the segment was annotated as "others". In such segments, the information about task structure cannot be obtained.

The pre-exchange is annotated as a kind of dialogue context and used as one of the learning features itself. The coders annotated the kind of pre-exchange by selecting one of nine categories of exchanges which are defined in section 4.1 in detail.

3.4 Inter-coder agreement for the annotation

As mentioned in the previous sections, we annotated our corpus with regard to the following characteristics: the class of cue phrases (ordinal, changeover, conjunctive), segment boundary, and hierarchical structure of the segment, the purpose of the segment, and the dialogue exchange at the end of the immediately preceding segment.

The extent of inter-coder agreement between two coders in these annotation are calculated by using Cohen's Kappa κ (Bakeman and Gottman, 1986; Carletta, 1996). The inter-coder agreement (κ) about the class of cue phrase is 0.68, about the purpose of the segment is 0.79, and about the type of pre-exchange is 0.67. The extent of agreement about the segment boundary and the hierarchical structure is calculated using modified Cohen's Kappa presented by (Flammia and Zue, 1995). This Cohen's Kappa is 0.66.

Fleiss et al. (1981) characterizes kappas of .40 to .60 as fair, .60 to .75 as good, and over .75 as excellent. According to this categorization of levels of inter-coder agreement, the inter-coder agreement for cue phrase, pre-exchange, and discourse boundary and structure is good. The agreement on segment purpose is excellent. Thus, these results indicate that our corpus coding is adequately reliable and objective.

When the two coders' analyses did not agree, the third coder judged this point; only those parts whose analysis is output by more than two coders was used as learning data.

4 Learning experiment

4.1 Learning features

This section describes a learning experiment using C4.5 (Quinlan, 1993). First, we define 10 learning features concerned with three factors.

(1)Discourse structure: Structural information about the preceding dialogue.

- **Embedding** The depth of embedding from the top level.
- **Place** The number of elder sister segments.
- **Place2** The number of elder sister segments except pupil initiative segments.
- **Recent elder sister's cue (Res-cue)** The cue phrase that occurs at the beginning of the most recent elder sister segment. They are classified into three kinds of simple cue phrases: ord (ordinal), ch (changeover), con (conjunctive) or a kind of composite cue phrase such as ch+ord (changeover + ordinal).
- **Res-cue2** The cue phrase that occurs at the beginning of the most recent elder sister segment except pupil initiative segments.
- **Discourse transition (D-trans)** Types of change in attentional state accompanied by topic change[3] such as push and pop. Pop from the pupil initiative subdialogue is categorized as "ui-pop".

(2)Task structure: Information that estimates the complexity of succeeding dialogue.

[2] When a cue phrase follows acknowledgement (Yes) or a stammer, these speech fragments that do not have propositional content are ignored and the cue phrases after the fragments are annotated as the beginning of the segment.

[3] Clark (1997) presents a term "discourse topic" as concept equivalent to focus space in (Grosz and Sidner, 1986), and call their transition "discourse transition". For example, "push" is defied as the transition to the sub topic, and "next" is defined as the transition to the same level proceeding topic.

Table 1: The learning features

factor	feature name	values
Discourse structure	Embedding	integer
	Place	integer
	Place2	integer
	Res-cue	nil, ord, ch, con, ch+ord, con+ord, con+ch, other
	Res-cue2	nil, ord, ch, con, ch+ord, con+ord, con+ch, other
	D-trans	pop, push, next, ui-pop, NA
Task structure	T-hierarchy	integer
	Subgoal	integer
Dialogue structure	Pre-exchange	conf, req, inf, quest, ui-conf, ui-req, ui-inf, ui-quest, NA
	Ps-cue	nil, ord, ch, con, ch+ord, con+ord, con+ch, other

Task-hierarchy (T-hierarchy) The number of goal-subgoal relations from the current goal to primitive actions. This estimates the depth of embedding in the succeeding dialogue.

Subgoal The number of direct subgoals of the current goal. If zero, then it is a primitive action.

(3)Dialogue context Information about the preceding segment.

Pre-exchange Type of exchange that occurs at the end of the immediately preceding segment, or type of exchange immediately preceding the cue phrase. There are four categories, conf (confirmation-answer), req (request-answer), inf (information-reply). ques (question-answer). They are also distinguished by the initiator of the exchange; explainer initiative or pupil initiative. When the category of the exchange is not clear, it is classified as not applicable (NA). Therefore, there are nine values for this feature.

Preceding segment's cue (Ps-cue) The cue phrase that occurs at the beginning of the immediately preceding segment.

The values of these features are shown in Table 1. Among the above learning features, *Embedding, Place, Place2, Res-cue, Res-cue2, Ps-cue*, and *D-trans* are derived automatically from the information about segment boundary and the segment hierarchy annotated in the corpus (an example is shown in Figure 1). The depth of task hierarchy (*T-hierarchy*) and the number of direct subgoals (*Subgoal*) are determined by finding the annotated segment purpose in the given task structure.

4.2 Learning algorithm

In this study, C4.5 (Quinlan, 1993) is used as learning program. This program takes two inputs, (1)the definition of classes that should be learned, and the names and the values of a set of features, and (2) the data which is a set of instances whose class and feature values are specified. As a result of machine learning, the program outputs a decision tree for judgement.

We use cross-validation for estimating the accuracy of the model because this method avoids the disadvantages common with small data sets whose number of cases is less than 1000. In this study, 10-fold cross-validation is applied, so that in each run 90% of the cases are used for training and the remaining 10% are used for testing. The C4.5 program also has an option that causes the values of discrete attribute to be grouped. We selected this option because there are many values in some features and the decision tree becomes very complex if each value has one branch.

4.3 Results and discussion

Decision trees for distinguishing the usage of three kinds of cue phrases (changeover, ordinal, and conjunctive) were computed by the machine learning algorithm C4.5. As learning features, the 10 features mentioned in section 4.1 are used. From nine dialogues, 545 instances were derived as training data. In 545 instances, 300 were conjunctive, 168 were changeover, and 77 were ordinal. The most frequent category, conjunctive, accounts for 55% of all cases. Thus, the baseline error rate is 45%. This means that one would be wrong 45% of the time if this category was always chosen.

First, the prediction power of each learning feature is examined. The results of learning experiments using single features are shown in Table 2. In pruning the initial tree, C4.5 calculates actual and estimated error rates for the pruned tree. The error rate shown in this table is the mean of estimated error rates for the pruned trees under 10-fold cross-validation. The 95% confidence intervals are shown after "$\pm$". Those are calculated using Student's t distribution. The error rate e_1 is significantly better than e_2 if the upper bound of the 95% confidence interval for e_1 is lower than the lower bound of the 95% confidence interval for e_2. As shown in Table 2, the decision tree obtained with the *Pre-exchange* fea-

Table 2: The error rates with each model

Embedding	46.5 ± 0.1	Pre-exchange	41.5 ± 0.5
Place	42.5 ± 0.4	Ps-cue	46.5 ± 0.3
Place2	43.8 ± 0.4	DS model	35.6 ± 0.4
Res-cue	44.9 ± 0.3	Task model	41.8 ± 0.3
Res-cue2	45.1 ± 0.4	DC model	39.1 ± 0.6
D-trans	45.0 ± 0.5	All feature model	29.9 ± 0.4
T-hierarchy	42.4 ± 0.3	Simplest model	30.6 ± 0.3
Subgoal	42.5 ± 0.3		

Table 3: The set of learning features for each model

	Discourse Structure						Task Structure		Dialogue Context	
Model	Embedding	Place	Place2	Res-cue	Res-cue2	D-trans	T-hierarchy	Subgoal	Pre-exchange	Ps-cue
DS	√	√	√	√	√	√				
Task							√	√		
DC									√	√
All feature	√	√	√	√	√	√	√	√	√	√
Simplest	√	√				√		√	√	√

ture performs best, and its error rate is 41.5%. In all experiments, the error rates are more than 40% and none are considerably better than the baseline. These results suggest that using only a single learning feature is not sufficient for selecting cue phrases correctly.

As the single feature models are not sufficient, it is necessary to find the best set of learning features for selecting cue phrases. We call a set of features a model and the best model (the best set of features) is obtained using the following procedure. First, we set some multiple features models and carry out learning experiments using these models in order to find the best performing model and the best error rate. We then eliminate the features from the best performance model in order to make the model simpler. Thus, the best model we try to find is the one that uses the smallest number of learning features but whose performance equals the best error rate.

We construct four multiple feature models. The name of the model and the combination of features in the model are shown in Table 3. The discourse structure model (the *DS model*) used learning features concerned with discourse structure. The *Task model* used those concerned with task structure, and the dialogue context (the *DC model*) used those concerned with dialogue context. The *All feature model* uses all learning features. The best error rate among these models is 29.9% in *All feature model* as shown in Table 2. The error rate is reduced about 15% from the baseline.

Therefore, the best model is the one that uses fewer learning features than the *All feature model* and that equals the performance of that model. In order to reduce the number of features considered, we examined which features have redundant information, and omitted these features from the *All feature model*. The overlapping features were found by examining the correlation between the features. As for numerical features that take number values, the correlation coefficient between *Place* and *Place2*, and between *T-hierarchy* and *Subgoal* are high (ρ=0.694, 0.784, respectively). As for categorical features, agreement between *Res-cue* and *Res-cue2* is 95%. These highly correlated features can be represented by just one of them. As the result of many experiments varying the combination of features used, we determined the *Simplest model* which uses six features: *Embedding, Place, D-trans, Subgoal, Pre-exchange*, and *Ps-cue* as shown at the bottom line in Table 3. The error rate of the *Simplest model* is 30.6% as shown in Table 2. It is very close to that of the *All feature model* though the difference is statistically significant.

In addition to comparing only the overall error rates, in order to compare the performance of these two models in more detail, we calculated the information retrieval metrics for each category, changeover, ordinal, and conjunctive. Figure 2 shows the equations used to calculate the metrics. For example, recall rate is the ratio of the cue phrases correctly predicted by the model as class X to the cue phrases of class X in the corpus. Precision rate is the ratio of cue phrases correctly predicted to be class X to all cue phrases predicted to be class X. In addition, in order to get an intuitive feel of overall performance, we also calculated the sum of the deviation from ideal values in each metric as in (Passonneau and Litman, 1997). The summed deviation is calculated by the following numerical formula:

(1 - Recall) + (1 - Precision) +
Fallout + Error

Table 4 shows the results of these metrics for the two models. Standard deviation is shown in parentheses. The value of each metric is the average of

Table 4: Performance on training set using cross-validation

Model	Cue phrase	Recall	Precision	Fallout	Error	Summed Deviation
All feature model	ordinal	0.50 (0.18)	0.64 (0.10)	0.05 (0.03)	0.11 (0.03)	1.03 (0.23)
	changeover	0.53 (0.14)	0.58 (0.07)	0.17 (0.05)	0.26 (0.04)	1.32 (0.23)
	conjunctive	0.80 (0.06)	0.73 (0.05)	0.38 (0.11)	0.28 (0.04)	1.12 (0.16)
Simplest model	ordinal	0.48 (0.17)	0.66 (0.17)	0.45 (0.03)	0.11 (0.02)	1.01 (0.26)
	changeover	0.50 (0.12)	0.62 (0.08)	0.14 (0.03)	0.25 (0.05)	1.27 (0.24)
	conjunctive	0.85 (0.04)	0.72 (0.04)	0.40 (0.08)	0.26 (0.04)	1.09 (0.17)

		Corpus	
		Class-X	not-Class-X
C4.5 Program	Class-X	a	b
	not-Class-X	c	d

$$\text{Recall} = \frac{a}{(a+c)} \qquad \text{Fallout} = \frac{b}{(b+d)}$$

$$\text{Precision} = \frac{a}{(a+b)} \qquad \text{Error} = \frac{(b+c)}{(a+b+c+d)}$$

Figure 2: Information retrieval metrics

the metrics on the test set in each run of 10-fold cross-validation. Comparing the summed deviation, the performance of the *Simplest model* is better than that of the *All feature model* in all categories of cue phrases. The summed deviations of the *Simplest model*, 1.01 for ordinal, 1.27 for changeover, and 1.09 for conjunctive, are lower than those of the *All feature model*. Thus, as a result of evaluating the models in detail using the information retrieval metrics, it is concluded that the *Simplest model* is the best performing model. In addition, the *Simplest model* is the most elegant model because it uses fewer learning features than the *All feature model*. Just six features, *Embedding, Place, D-trans, Subgoal, Pre-exchange*, and *Ps-cue*, are enough for selecting organization cue phrases.

Classifying the six features in the *Simplest model*, it is found that these features come from all factors, discourse structure, task structure, and dialogue context. *Embedding, Place, D-trans* are the features of discourse structure, *Subgoal* is about task structure, and *Pre-exchange* and *Ps-cue* are about dialogue context. This result indicates that all the factors are necessary to predict cue phrases. The important factors for cue phrase selection are task structure and dialogue context as well as discourse structure, the focus of many earlier studies.

While we identified the six features from the three kinds of factors, by looking at the decision trees created in the learning experiment, we found which features were more important than others in selecting cue phrases. The features appearing near the root node are more important. Figure 3 shows the top part of a decision tree obtained from the *Simplest model*. In all 10 decision trees resulting from the cross-validation experiment in the *Simplest model*, *Place* feature appears at the root node. In 7 of 10 trees, *Embedding* and *Pre-exchange* appeared just below the root node. In these trees, if the *Place* of the segment is the first at that level (i.e. there is no elder sister.), then *Embedding* appears at the next node, otherwise if the segment is not the first one at that level, then *Pre-exchange* appears at the next node. Thus, if there are some elder sister segments, information about dialogue context is used for selecting cue phrases. On the other hand, if there is no elder sister segment, information about discourse structure is used for the judgement. These results suggest that the information about discourse structure, especially place of segments and the depth of embedding, and the dialogue context, especially the kind of immediately preceding dialogue exchange, play important roles in cue phrase selection.

5 Conclusion and Further work

This paper reported the results of using a machine learning algorithm for identifying learning features and obtaining decision trees for selecting cue phrases. It also reported the result of a quantitative evaluation of the decision trees learned. Learning features concerning three factors, discourse structure, task structure, and dialogue context, were examined. By carrying out many experiments in which the combinations of learning features were varied, we found the most simple and effective learning feature set. The accuracy of the best model that uses 6 learning features is about 70%. The error rate is reduced about 25% from the baseline. These results support the claims of previous studies that discourse structure influence cue selection. In addition, it is revealed that task structure and dialogue context are also indispensable factors.

We focus on predicting the cue phrases that occur at the beginning of discourse segments for signaling inter-segment "sequence" relation. Elhadad and McKeown (1990), on the other hand, has presented a model for distinguishing connectives, which link two propositions, using some pragmatic constraints. In (Moser and Moore, 1995a; Moser and Moore, 1995b), the relationship between placement and selection of cue phrases was investigated using the core:contributor relations among units within a segment (Moser and Moore, 1995a). Although we discussed only the "sequence" relation between the

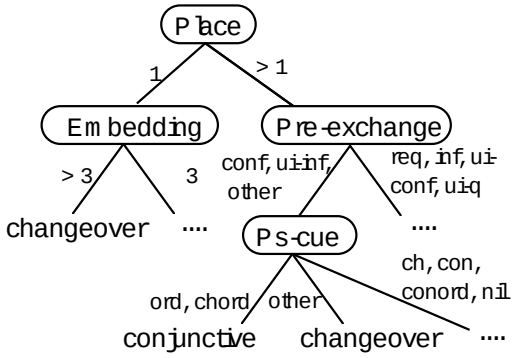

Figure 3: Top part of a decision tree

segments, the methods presented here will be useful in extending our model so as to select other kinds of cue phrases.

References

Roger Bakeman and John M. Gottman. 1986. *Observing Interaction.* Cambridge University Press.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Alison Cawsey. 1993. *Explanation and Interaction: The Computer Generation of Expalanatory Dialogues.* MIT Press.

Herbert H. Clark. 1997. *Using language.* Cambridge University Press.

Robin Cohen. 1984. A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics*, pages 251–258.

Michael Elhadad and Kathleen R. McKeown. 1990. Generating connectives. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 97–101.

Barbara Di Eugenio, Johanna D. Moore, and Massimo Paolucci. 1997. Learning features that predict cue usage. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association of Computational Linguistics*, pages 80–87.

Giovanni Flammia and Victor Zue. 1995. Empirical evaluation of human performance and agreement in parsing discourse constituents in spoken dialogue. In *Eurospeech*, pages 1965–1968.

J. L. Fleiss, J. Cohen, and B. S. Everitt. 1981. *Statistics Methods for Rates and Proportions.* Wiley.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara J. Grosz. 1977. The representation and use of focus in dialogue understanding. Technical Report 151, Artificial Intelligence Center, SRI International.

Raymonde Guindon. 1986. The structure of user-adviser dialogues: Is there method in their madness? In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 224–230.

Takashi Ichikawa. 1978. *Bunshouron gaisetsu (in Japanese).* Kyouiku shuppan.

Alistair Knott and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18:35–62.

Leila Kosseim and Guy Lapalme. 1994. Content and rhetorical status selection in instruction texts. In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 53–60.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, USC/ISI.

Johanna D. Moore. 1995. *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context.* MIT Press.

Takurou Moriyama, 1997. *Speech and Grammar (in Japanese)*, chapter 5. Kuroshio Shuppan.

Megan Moser and Johanna D. Moore. 1995a. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics*, pages 130–135.

Megan Moser and Johanna D. Moore. 1995b. Using discourse analysis and automatic text generation to study discourse cue usage. In *AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*, pages 92–98.

Christine H. Nakatani, Barbara J. Grosz, David D. Ahn, and Julia Hirschberg. 1995. Instructions for annotating discourses. Technical Report TR-21-95, Center for Research in Computing Technology, Harvard University.

Sharon L. Oviatt and Philip R. Cohen. 1990. Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. Technical Report CSLI-90-138, Center for the Study of Language and Information.

Rebecca J. Passonneau and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

John Ross Quinlan. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

Deitmar Rösner and Manfred Stede. 1992. Customizing RST for the automatic production of technical manuals. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Proceedings of the 6th International Workshop on Natural Language Generation*, pages 199–215. Springer-Verlag.

Deborah Schiffrin. 1987. *Discourse markers.* Cambridge University Press.

Anna-Brita Stenström. 1994. *An Introduction to Spoken Interaction.* Longman.