EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

LEARNING DIALOGUE STRUCTURES FROM A CORPUS

Jan Alexandersson and Norbert Reithinger *

DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
E-Mail: {alexandersson,reithinger}@dfki.de

ABSTRACT

This paper demonstrates some aspects of a plan processor which is a subcomponent of the dialogue module of VERBMOBIL. We describe how we transfer results from the research area of grammar extraction for the semi-automatic acquisition of plan operators for turn classes. We exploit statistical knowledge acquired during learning the grammar and incorporate top down predictions to enhance the correct analysis of turn classes described. A first evaluation shows a relative recognition rate of around 70% on unseen data.

1. INTRODUCTION

Due to many factors e.g. topic, number of participants and task, the structures of spoken dialogues varies very strongly. Different researchers have tried to define dialogue structures analytically. As one example, in [12] heuristics where developed for describing e.g. control in both task-oriented as well as advice-giving mixed initiative dialogues.

For the speech translation system VERBMOBIL [7, 4] we also need a description of the dialogue structure (e.g. for translation purposes) which identifies sections of a dialogue like for instance opening or closing. We have developed a plan processor that builds up such a description for time scheduling negotiation dialogues. As we started to code, enabling the plan processor to process *turns*, we soon recognized this as a cumbersome task – in the VERBMOBIL dialogues the number of ways people convey the same goal makes the input to the plan processor very hard to foresee.

This paper elaborates on some ideas reported in [2]. It describes how one can, by viewing plan recognition as parsing, transfer results from the field of grammar extraction to (semi-) automatically extract the plan operators from an annotated corpus. To model the turns, i.e. the contribution of one dialogue participant that consists of one or more utterances, we follow the approach taken by a number of researchers, namely to describe utterances by means of dialogue acts which describe the underlying intention and to model the characteristics of turns by the concept of initiative and response (see below). Using this approach the problem arises that a particular sequence of dialogue acts can belong to more than one turn class. To partly solve these problems, e.g. to find the correct turn class, we incorporate top down predictions which enhance the recognition rate significantly.

*This work was funded funded by the German Federal Ministry for Education, Research, Science and Technology (BMBF) in the framework of the VERBMOBIL project under grant 01IV101K/1. The responsibility for the contents of this study lies with the authors. Special thanks to our students Michael Kipp and Ralf Engel for unvaluable help with implementation and experiment issues.

2. THE INTENTIONAL STRUCTURE

The component responsible for the construction of the intentional structure is the plan processor [1, 3]. It uses a plan hierarchy describing the negotiation dialogues in a declarative way. The plan hierarchy is compiled off-line into a context-free grammar [11]. Due to the mediating scenario of VERBMOBIL it is used not to plan actively, but to recognize plans. The intentional structure is a tree-like structure mirroring different abstraction levels of the dialogue (c.f. dialogue phase, turn). It divides into four levels (see fig 1):

The Dialogue Act Level implements, with some minor extensions, the dialogue act hierarchy [6]

The Turn Level connects the utterances inside a turn.

The Phase Level distinguishes the three dialogue phases greeting, negotiation, and closing.

The Dialogue Level spans over the whole dialogue, eventually distinguishing negotiations of more than one appointment.

Plan operators for processing three of these levels have been hand coded. The turn level, however, turned out to be very hard to code due to the conversational setting – the dialogues contain a lot of phenomena typical for spontaneous speech.

Central for modeling turns is the concept of the forward/backward looking aspects of a turn. By backward looking aspect we mean that (a part of) the turn contains a direct reaction to something (e.g. proposal) introduced earlier in the dialogue. The forward looking aspect

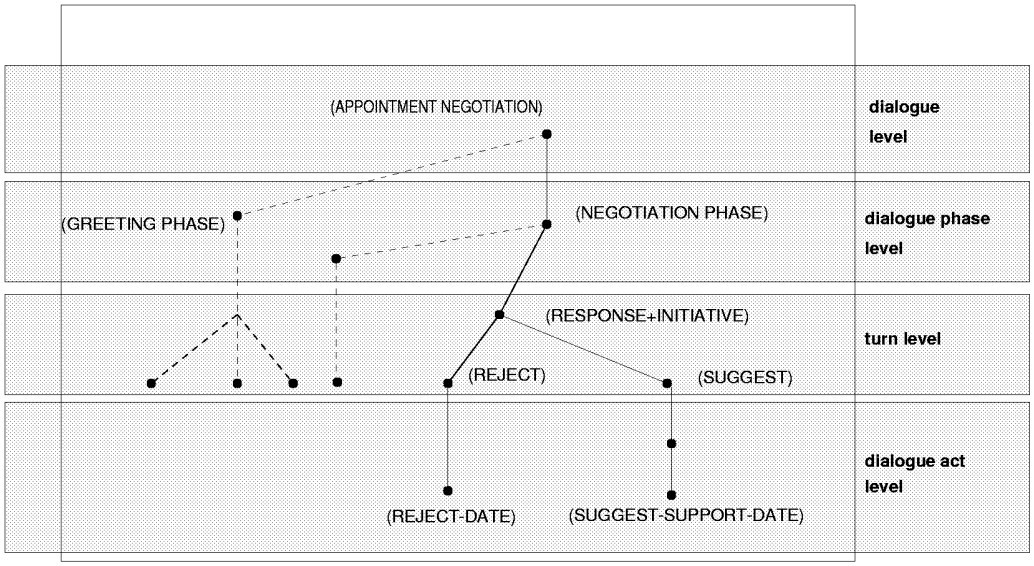

Figure 1: The different layers of the plan tree

roughly covers the cases where a new proposal or topic was introduced, which opens up a new discourse segment.

The main turn classes for negotiation dialogues are:

Initiative A (part of a) turn is annotated with the label *Initiative* when the turn has a forward looking aspect and (*i*) when something is suggested and the dialogue contains no open topics, or (*ii*) when a suggestion refines a previous proposal that has been accepted explicitly, or (*iii*) when a direct counter proposal is made.

Response A (part of a) turn is annotated with the label *Response* when the turn has a backward looking aspect. This occurs (*i*) when some earlier proposal is rejected or accepted, or (*ii*) when a declarative suggestion with an implicit acceptance contains a refinement of an earlier proposal.

Transfer-Initiative A (part of a) turn is annotated with the label *Transfer-Initiative* when the turn has a forward looking aspect and (*i*) when a topic is introduced without the locutor making a suggestion, or (*ii*) when a suggestion is explicitly requested.

Confirm A (part of a) turn is annotated with the label *confirm* when the turn has a backward looking aspect and (*i*) when a preceding acceptance is confirmed, or (*ii*) when a summarization of the agreement achieved so far is accepted.

Additionally we have used the labels *greet*, *bye*, *clarify-query*, and *clarify-answer* to mark the start and end of the dialogue, and to characterize turns which contain (parts of) clarification sub-dialogues. Moreover, a turn can be labeled with not just one of these labels, but with the concatenation of two of the above, indicated with a + (c.f. *greet+initiative*). In these classes the first contains a backward looking aspect and the second a forward looking aspect. This means that the turn carries the two functions as indicated by the classes used, and it also indicates that the structure of the turn contains two parts. Finally we introduced the classes *unknown* for cases where it is impossible to classify the turn given the above set, and *garbage* when the turn contains irrelevant contributions or – garbage. Currently we use the 25 turn-classes[1] shown in figure 2.

3. LEARNING THE STRUCTURE OF TURNS

[11] showed how plan recognition can be viewed as the parsing of a context free grammar. This viewpoint has many advantages since the research in parsing technology has produced a lot of both efficient and robust techniques. Our plan processor [1] is designed so that the plan hierarchy is compiled into a grammar which is, during runtime, processed using a simple top down left to right parsing technique, not consuming words or part-of-speech tags, but dialogue acts.

The research in the field of grammar induction has recently produced a lot of interesting results (c.f. [9, 5]). For the turn level we derive (stochastic) context free grammars for each turn class using the BOOGIE [10] system. It is a workbench for deriving structures enriched with statistical information (e.g. Hidden Markov Models and Stochastic Context Free Grammars) based on Bayesian model merging. It allows for a wide range of parameterization, which can make the learning algorithm, for instance, to generate a grammar that generalizes over the training set. The feature of being stochastic is important to us, since this information can be used to enhance the recognition rate (see below).

[1]It is theoretically possible to derive classes like *greet+bye* which would result in a total of 72 classes, but since some of them either do not appear in our corpus or are not likely to occur, we have not introduced them.

Negotiation phase	*initiative* *transfer-initiative* *initiative+response* *confirm* *confirm+transfer-initiative*	*response* *response+initiative* *response+transfer-initiative* *confirm+initiative*
Opening phase	*greet* *greet+transfer-initiative*	*greet+initiative*
Closing phase	*response+bye* *bye*	*confirm+bye*
Turns containing clarification dialogues	*clarify-query* *greet+clarify-query* *response+clarify-query* *confirm+clarify-query*	*clarify-answer* *clarify-answer+response* *clarify-answer+initiative* *clarify-answer+transfer-initiative*
Misc	*unknown*	*garbage*

Figure 2: The turn classes

Statistical disambiguation When we run the plan processor with the acquired operators or rules, we get some problems. The main is that more than one of the classes may be recognized with the same input. This is due to the fact that some dialogue act sequences may correspond to several classes. For instance the simple sequence (SUGGEST_SUPPORT_DATE) is, depending on context, member of the class *response* or *initiative*. Also the grammar induction algorithm generalizes over the training set, so that the resulting grammar can recognize sentences which are not in the training set. Obviously this is necessary, but has the side effect that a turn grammar also recognizes turns as belonging to its class that are better be classified as a different turn class.

If we have competing results from the plan recognition, we need a criterion to decide which class is most probable. Since the grammar is a stochastic one we can use the probability of the parse. The evaluation presented in the next section shows, however, that this is not very reliable. Therefore we combine it with the prediction of turn classes. In [8] it was shown how statistical methods from the field of speech recognition can be used for the prediction of dialogue acts. In this work we apply this method to the prediction of turn-classes.

Formally, let $P(G_i)$ be the probability given by the grammar[2] for class i, $P(C_i)$ and $P(C_i|C_{i-1})$ the uni- and bigram probabilities for the turn classes C_X. Then we use the maximization

$$C' = \arg\max_c q_0 P(G_i) + q_1 P(C_i) + q_2 P(C_i|C_{n-1})$$

($\Sigma q_i = 1$) to compute the most probable class for a particular turn.

[2]The probability of a Stochastic Context Free Grammar is the sum of the probability of all the parses. We do not construct all parse trees for each input. Instead we approximate this probability with the probability of the leftmost top down derivation.

4. EVALUATION

The Corpus As a basis of our experiments we have annotated a corpus consisting of 277 dialogues annotated with dialogue acts, and the 25 classes. For our experiments we have randomly split the data into four partitions where each partition consists of 3 disjunct sets: 70% for training, a validation set (20%) for adjusting the parameters (see below), and 10% for test. We evaluate our approach in four experiments to show on a variety of data how good the approach performs

Predicting the Turn Classes Using linear interpolation with uni– and bigrams we obtain the following results for one experiment:

Pred. Depth	1	2	3	4	5
Corr. Pred. %	38.3	54.6	72.5	77.7	81.1

If we regard just the most probable class predicted, we receive a correctness of 38.3% on the test set, while for the prediction depth 5 over 80% of the next turn class is within the prediction set. By trying the five most probable classes we can, depending on how we want to utilize the predictions, save a lot of computing, since we in over 80% of all cases will get the correct class. The reason for not getting 100% hit rate even for prediction depth 5 is due to irregularities like clarification dialogues. Clarification dialogues can occur anywhere in a dialogue and are therefore very hard (not to say impossible) to predict.

Effect of the predictions To see the effect of incorporating the top down predictions, we show how the recognition rate depends on the different probabilities produced by the stochastic context free grammar and the n-gram statistics. The table below shows different values for the q_is as described above on the first of our four partitions.

q_0	q_1	q_2	%	% (rel)
1.0	0.0	0.0	45.0	47.9
0.0	1.0	0.0	28.0	29.8
0.0	0.0	1.0	35.5	37.7
0.3	0.3	0.4	67.0	71.3

As can be seen, neither the grammar probability alone, nor the probabilities of the uni- or bigrams alone provide for a good classification. Only in the case where the interpolation factors are adapted to the evaluation set using the EM-algorithm, the recognition rate is satisfactory.

Evaluation of four experiments Evaluating the classification method described above on all four data partitions resulted in the numbers as given in figure 3. The first line (Total) shows the total number of turns in the different partitions. The second and third line (Possible) shows in how many a plan structure for the turn was computed at all . "CC" shows the number of correct classifications. The two last lines contain the percentages for correct classification with respect to the "Total" and the "Possible" lines.

Partition	1	2	3	4	Aver.
Total	288	284	377	317	180.9
Possible	265	257	326	293	163.0
Possible (%)	91.8	90.5	86.2	92.2	90.2
CC	189	173	228	172	190.5
CC Abs (%)	67.0	62.7	58.4	55.7	61.0
CC Rel (%)	71.3	67.3	69.9	58.7	66.8

Figure 3: Correct classifications on the test sets.

5. DISCUSSION

While the numbers for the first and third experiments are quite encouraging, the other experiments do not perform as well, especially the fourth.

One reason is that the criteria for whether a turn should be regarded as, for instance, response or initiative can not been disambiguated by statistical means only. To incorporate knowledge about the focus and how new utterances relates to the current foci is necessary for determining the correct class.

Also, irregularities like embedded clarification dialogues are very hard to include in a grammar-like approach. Partly they can be resolved by allowing a turn to belong to multiple classes. In our corpus clarification dialogues are not always completed, but are left unresolved. These problems however are not within the central scope this work, since we are just focusing on finding a structure describing a turn, and no structures describing, for instance, adjacency pairs.

Whether the drop of correct classifications for the forth partition is just occasionally or not has to be investigated further.

6. CONCLUSION

We demonstrated a new approach to describe dialogue structure using semi-automatically derived plan operators combined by a statistical disambiguation component. It is fully operable and integrated in the VERBMOBIL system

The recognition results so far are encouraging, but we have to analyze more experiments to further improve. One point is to utilize other knowledge sources. The decision has also to use information about the content for the proper selection of the class. Another point is to guide the grammar learning algorithm so that the structures created by the plan processor describe the relations within a turn more clearly. These structures will be used in the future to generate protocols of the dialogue and therefore must be concise.

7. REFERENCES

1. Jan Alexandersson. Plan recognition in VERBMOBIL. In Mathias Bauer, Sandra Carberry, and Diane Litman, editors, *Proceedings of the* IJCAI-*95 Workshop The Next Generation of Plan Recognition Systems: Challenges for and Insight from Related Areas of AI*, pages 2–7, Montreal, August 1995.
2. Jan Alexandersson. Some Ideas for the Automatic Acquisition of Dialogue Structure. In Anton Nijholt, Harry Bunt, Susann LuperFoy, Gert Veldhuijzen van Zanten, and Jan Schaake, editors, *Proceedings of the Eleventh Twente Workshop on Language Technology, TWLT, Dialogue Management in Natural Language Systems*, pages 149–158, Enschede, Netherlands, June 19-21 1996.
3. Jan Alexandersson, Norbert Reithinger, and Elisabeth Maier. Insights into the Dialogue Processing of VERBMOBIL. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLP '97*, pages 33–40, Washington, DC, 1997.
4. Thomas Bub and Johannes Schwinn. Verbmobil: The evolution of a complex large speech-to-speech translation system. In *Proceedings of ICSLP-96*, pages 2371–2374, Philadelphia, PA., 1996.
5. Stanley F. Chen. *Building Probabalistic Models for Natural Language*. PhD thesis, Harward University Cambridge, Massachusetts, 1996.
6. Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J. Joachim Quantz. Dialogue Acts in VERBMOBIL. Verbmobil Report 65, Universität Hamburg, DFKI Saarbrücken, Universität Erlangen, TU Berlin, 1995.
7. Martin Kay, Jean Mark Gawron, and Peter Norvig. *Verbmobil. A Translation System for Face-to-Face Dialog.* Chicago University Press, 1994. CSLI Lecture Notes, Vol. 33.
8. Norbert Reithinger. Some Experiments in Speech Act Prediction. In *AAAI 95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.
9. Andreas Stolcke. *Bayesian Learning of Probabilistic Language Models.* PhD thesis, University of California at Berkeley, 1994.
10. Andreas Stolcke. *How to Boogie: A Manual for Bayesian Object-oriented Grammar Induction and Estimation.* International Computer Sience Institute of Berkley, California, 1994.
11. Marc B. Vilain. Getting Serious about Parsing Plans: a Grammatical Analysis of Plan Recognition. In *Proceedings of AAAI-90*, pages 190–197, 1990.
12. Marlyn Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL'90)*, 1990.