Reference Resolution Mechanisms in Dialogue Management

Petra Gieselmann
Interactive System Labs
Universität Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany
petra@ira.uka.de

Abstract

Humanoid robots which are able to walk and behave human-like became very popular in the last few years. Now it is high time that they are able to use more natural communication means so that the human-robot interaction resembles more and more to human-human communication. Therefore, in this paper, we evaluate different reference resolution mechanisms within a dialogue management system for human-robot communication in a household environment. User studies showed that most of the pronouns can be resolved by a pragmatic, simplified approach.

1 Introduction

Dialogue management systems as well as mechanisms for reference resolution are well known research areas. Nevertheless, they have mostly been analyzed from different points of view until now. In this paper, we want to combine both by using well known pronoun resolution mechanisms within a dialogue management system for human robot communication in a household environment. In this context which is specifically tailored for unexperienced users, it is important that the user can talk to the robot in the same way as to a human servant for example. Therefore, the communication has to be as natural as possible which also includes pronoun resolution and multimodal communication mechanisms.

This paper deals with reference resolution of personal and deictic pronouns. Natural human robot interaction in a household environment is also explored. Section two gives an overview of related work on anaphora resolution in general and on special reference resolution mechanisms used in dialogue management systems in particular. In section three, our dialogue manager is explained. Section four deals with context management and our mechanisms for reference resolution. Section five gives a conclusion and outlook.

2 Related Work

Pronoun resolution is a well examined field in computational linguistics. Different theoretical articles have been written on this topic and methods from the field of Artificial Intelligence, such as inference mechanisms and world knowledge, have been explored in detail. Here, we want to have a look at the problem from a more pragmatic point of view. Therefore, we want to concentrate on deictic pronouns which can be resolved by means of gesture recognition and personal pronouns which are resolved by our pronoun resolution mechanisms. Other resolution mechanisms are the topic of future research.

2.1 Reference Resolution in General

Since there are so many researchers dealing with reference resolution from different point of views, such as philosophy, psychology, linguistics, computer science, etc., we want to take into account here only a small part of them which is relevant for our research.

One of the oldest algorithms for resolving pronouns is Hobb's naive algorithm (Hobbs, 1977). It simply traverses the surface parse trees of the sentences in a text looking for noun phrases of the correct number and gender as antecedents for pronouns. Although this algorithm is quite simple, it works fine and about 90% of the pronouns can be resolved (Hobbs, 1977).

The theory of discourse structure and centering invented and further developed by Grosz et al. (Grosz and Sidner, 1986; Brennan et al., 1987; Grosz et al., 1995; Walker, 1998) serves for tracking discourse context and binding pronouns. First, a set of all the cospecification relationships is created. Then it is filtered, classified and finally ranked by some rules. These rules rely on the relationship between antecedent and pronoun, such as parallelism of grammatical function, recency, etc. Furthermore, continuing with the same entity in the discourse center is preferred over retaining it which is preferred over shifting the discourse entities completely. Although the algorithm is much more complicated than Hobb's naive one, the results are similar (Tetreault and Allen, 2003).

As an extension of the centering model, Strube uses a list of salient discourse entities which is called S-list (Strube, 1998). This list is ranked based on information status. Therefore, it uses the distinction between new and old information in the discourse and incorporates also preferences for inter- and intrasentential anaphora which is not included in the original centering model.

CogNIAC (Baldwin, 1995) is a pronoun resolution engine which defines a set of rules for finding the correct antecedent in a list. These rules are somewhat simple, such as ”If there is only one possible antecedent in the preceding input sentence, use this”; world knowledge is not used for pronoun resolution. Nevertheless, these rules seem to be quite efficient given the fact that he reported about 92% precision.

All of these mechanisms have been developed by means of written texts. They can be also used for spoken communication to a certain extent, but have to be adapted to its special needs, especially covering spontaneous effects. Therefore, the next chapter deals with reference resolution mechanisms used in spoken natural language dialogues.

2.2 Reference Resolution in Multimodal Dialogue Management

Until now, there are only very few dialogue systems which use a reference resolution module because most of them have been specifically tailored for communication via phone, such as flight and train timetable information systems (McTear, 2002; Allen et al., 2000; Stallard, 2000), call-routing systems (Gorin et al., 2002), weather information systems, (Zue et al., 2000) etc. and do therefore not need reference resolution. But now since the number of systems for direct human machine communication from face to face, such as human robot interaction, increases, we need to take into account the situated and context-dependent communication, the changing environment, the multimodal interaction, etc. Therefore, we want to have a look at the resolution mechanisms necessary in situated and context-dependent communication.

For example, Kumar et al. uses an approach based on cognitive grammar which assumes conceptual semantics (Kumar et al., 2003). Reference domains identify representations for subsets of contextual entities to which can be referred, such as individual objects and also collection of objects. The important feature of a reference domain are its partitions which define in conjunction with focus and salience the criteria for reference resolution. Underspecified reference domains are composed with the existent context structure by means of grouping and assimilation. In this way, references can be resolved by finding the corresponding node within a context structure. Since the same mechanism is used for linguistic expressions and for gestures, different kinds of references, such as deixis and pronouns, can be resolved.

Other researchers (Landragin and Romary, 2003) propose a classification of referring modes which describes referring actions, and disambiguation principles to define the correct referent. References can be resolved by means of unification with the information available in context so that the one with the best unification result is kept. In this way, also deictic pronouns and pointing gestures can be resolved.

For the galaxy system, a whole context resolution server has been developed (Filisko and Seneff, 2003) which includes repairing mechanisms, anaphora and ellipsis resolution, history functions, etc. Pronouns are resolved by means of a discourse entity list which is searched for a possible antecedent.

Out of these approaches, we created a reference resolution model which uses similar methods, such as a list of possible antecedents and rules for the agreement between the antecedent and the pronoun. It also works for personal and deictic pronouns and is specifically tailored for human robot communication by including for example some knowledge about the actual situation of the robot. Therefore, it is not as theoretically complex as some of the mentioned approaches, but works efficiently in our scenario.

3 Dialogue Management

Our dialogue manager is based on the approaches of the language and domain independent dialogue manager ARIADNE (Denecke, 2002) which is specifically tailored for rapid prototyping because general concepts are already available and can be reused. Only the domain and language dependent components have to be implemented for new applications, such as: An ontology, a specification of the dialogue goals, a data base, a context-free grammar and generation templates.

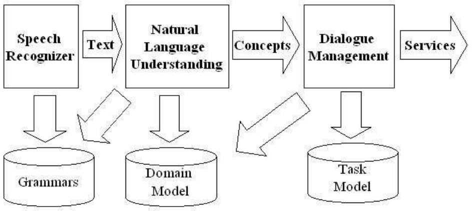

Figure 1: The Dialogue Management Workflow with Its Resources

The dialogue manager uses typed feature structures to represent semantic discourse information (Carpenter, 1992). In figure 1, the whole dialogue management workflow can be seen: First of all, the user utterance is parsed by means of a context-free grammar which is enhanced by information from the ontology defining all the objects, tasks and properties about which the user can talk. In figure 2, you can see a part of the ontology we defined for our robot dialogue system. It consists of different objects available in the kitchen, actions the robot can accomplish for the user and properties of the objects resp. the actions.

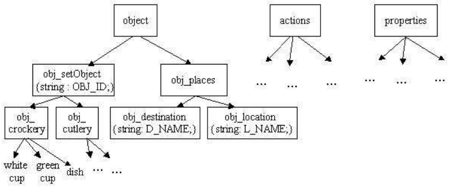

Figure 2: Part of the Ontology

An example of the semantic representation which is created during parsing can be found in figure 3. This semantic representation is compared against the dialogue goals. If all the necessary information to accomplish a goal is available, the dialogue system calls the corresponding service. But if some information is still missing, the dialogue manager uses clarification questions to get this information from the user. The spoken output is created by means of generation templates.

```
[ act_put  OBJ
  [ obj_puttable
    [ generic:NAME [ "it" ] ]
  ]
  [ DESTINATION
    [ DEST [ "table" ] ]
  ]
```

Figure 3: Semantic Representation of the Sentence ”put it on the table”

The database serves as a context model which includes different world knowledge sources and is used for the resolution of references as described below. Therefore, you can find there information on the position of the objects in the world as well as information on possible antecedents.

Also the ontology plays an important role in reference resolution because it is used to define the semantic agreement between the reference and its antecedent: If both of them belong to the same category or to a subcategory in the ontology, then there is a semantic agreement between them. In the example in figure 3, you can see that ”it” refers to an object which is puttable because of the verb

”put” which expects a puttable object. This means that other possible antecedents are semantically excluded. If the user said in the previous sentence for example ”get the cup from the board”, the ”board” cannot be an antecedent for the pronoun ”it” because it belongs to another category in the ontology. In this way, we assure that only semantically useful antecedents are taken into account by our algorithm.

4 Context Modeling for Reference Resolution

As you can also see in the example below (see figure 4), the two different types of references we want to resolve are personal pronouns and deictic pronouns. The reference resolution for both of them takes place during the creation of the semantic representation. Therefore, the input is the parsed user utterance transformed into a semantic representation as you can see in figure 3 and the output is the semantic representation enhanced with reference resolution information.

4.1 Our Context Model

The context model contains information on the environment: For example, all the available objects are stored there with their three-dimensional position in the room. This information can also be updated during the actual dialogue processing, if an object is moved by the user or by the robot itself.

In addition, possible antecedents are stored in the context model in a list similar to Strube’s S-list. Since we only found nominal antecedents for the pronouns in our user studies, we decided to resolve only these pronouns in a first step. In addition, some expletive pronouns are already covered by the grammar by means of expressions such as ”it is too dark in here”; others cannot be resolved at the moment.

We implemented our context model in such a way that it works similar to the human brain and therefore ”forget” old antecedents after a certain period of time (Clark, 1978). Whenever a new user utterance comes in, the context model is updated with the corresponding possible antecedents.

4.2 Mechanisms for Pronoun Resolution

For reference resolution, the context model is used and linguistic expressions, such as personal pronouns, as well as pointing gestures and deictic pronouns are both resolved - in multimodal parsing or in pronoun resolution.

4.2.1 Deictic Pronouns

We made a user study with our household robot where the users interacted with the robot via speech and gestures. They were told that they can use pointing gestures and we found in about 10% of the sentences pointing gestures coupled with deictic pronouns (see table 1).

Total Number of Turns	1151
Turns with Deictic Pronouns	125
Deictic Pronoun Rate (in %)	10.86

Table 1: Number of turns with deictic pronouns in an experiment with our household robot

For resolving deictic pronouns, we assume that a referring pointing gesture is available at the same time, as you can see in the second example of figure 4. We use a gesture recognizer and multimodal parsing of speech and gestures so that the information from both input modalities is merged on a semantic base by means of time stamps (Gieselmann and Denecke, 2003).

Therefore, gesture input is resolved by means of the context model which consists of different objects in the kitchen, such as cups, dishes, forks, knifes, spoons and lamps. An n-best list with all the pointing gestures matching a possible target object from the context model is created. The disambiguation is then performed by merging speech and gesture in a multimodal parsing process (Stiefelhagen et al., 2004). Deictic pronouns without a referring gesture cannot be resolved at the moment.

4.2.2 Personal Pronouns

In another small user study, where the users had to make the robot set the table, we found in about 6% of the sentences personal pronouns (see table 2).

By means of the context model, personal pronouns can be resolved, as you can see in the first

```
User: Robbi, get the blue cup from the board.
Robbi: Going to take the blue cup from the board.
User: Bring it to me.
Robbi: Going to bring you the blue cup.

User: Switch on that light. + pointing gesture to the big lamp
Robbi: Switching on the big lamp.
```

Figure 4: Example Dialogue taken from our user studies with a household robot

Total Number of Turns	572
Turns with Personal Pronouns	37
Personal Pronoun Rate (in %)	6.47

Table 2: Number of turns with personal pronouns in an experiment with our household robot

example in figure 4 in two different ways:

- out of the dialogue context taking into account information from the previous sentences
- out of the situation. This means that some kind of simple world knowledge is used. For example, if the robot has a cup in its possession, and the user tells it ”Put *it* there”, then it can be assumed that ”it” refers to this cup.

Therefore, there are two different ways how pronouns can be resolved. On one hand, the information on what can be found in the robot's possession is in the context model and can therefore be used for the resolution. In this way, pronouns can be simply resolved by replacing the pronoun by the object in the robot's possession.

On the other hand, we use our list of possible antecedents in the context model and look there whether there is a possible antecedent. Similar to the pronoun resolution mechanisms mentioned above, we also use some rules, such as that the pronoun and the antecedent have to agree in their syntactic and semantic features. This means that they have to have the same number and gender as far as syntax is concerned and both of them have to belong to the same category or a subcategory in the ontology, as far as semantic is concerned. Since the antecedents are ranked by their appearance and also deleted, if they are too old, we can use the first possible antecedent which is found, and put its semantic representation in the discourse.

Both methods are not very complex, but work efficiently in our scenario so that about 90% of the pronouns can be resolved. In our user study even all the pronouns can be resolved just out of the situation by means of the world knowledge in the context model. Therefore, we do not even need the more complex mechanism with all the possible antecedents in the context model. But since this might also be due to the fact that the scenario is quite simple at the moment, we will test this with an enhanced version in a more complex scenario.

Also a combination of both methods sounds promising. Namely, there are situations where the method based only on the previous sentences will fail because the previously mentioned correct antecedent is too many sentences away and cannot be found therefore. On the other hand, also the method of just using the information what is in the robot's possession can fail easily, if the robot has something else than the user is referring to. Therefore, we want to do further experiments with a combination of both methods to see whether we can resolve even more pronouns by this combination.

5 Conclusion and Outlook

In this paper, we developed some methods for reference resolution in human robot communication. We focused our attention on the pragmatic aspects of the resolution and started with personal and deictic pronouns. Both of them are resolved by means of the context model.

In our user studies, we found out that it was possible to resolve the personal pronouns just by taking into account the current situation without using any knowledge of the previous sentences. For the future, we want to evaluate whether this is also feasible in more complex situations which would facilitate reference resolution a lot.

Furthermore, we also want to evaluate whether a combination of the two mentioned methods leads to better results and how these methods can be efficiently combined to take advantage from both of them while avoiding their disadvantages.

Acknowledgments

This work was supported in part by the German Research Foundation (DFG) as part of the SFB 588 and within the FAME project by the European Union as IST-2000-28323.

References

James Allen, George Ferguson, Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski Zollo. 2000. Dialogue systems: From theory to practice in trains-96. *Robert Dale, Hermann Moisl, and Harold Somers, eds.: Handbook of Natural Language Processing*, pages 347–376.

Breck Baldwin. 1995. *CogNIAC: A High Precision Pronoun Resolution Engine*. University of Pennsylvania Department of Computer and Information Sciences Ph.D. Thesis, Pennsylvania, US.

E. Brennan, Marilyn Walker Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. *Proceedings of the 25th Annual Meeting of the Association of Computational Linguistics*, pages 155–162.

Bob Carpenter. 1992. The logic of typed feature structures.

H. H. Clark. 1978. On inferring what is meant. *W. J. M. Levelt and G. B. Flores d'Arcais (Eds.). Studies in the Perception of Language*, pages 295–322.

Matthias Denecke. 2002. Rapid prototyping for spoken dialogue systems. *Proceedings of the 19th International Conference on Computational Linguistics*.

Edward Filisko and Stephanie Seneff. 2003. A context resolution server for the galaxy conversational systems. *Proceedings of the Eurospeech*.

Petra Gieselmann and Matthias Denecke. 2003. Towards multimodal interaction with an intelligent room. *Proceedings of the Eurospeech*.

A. L. Gorin, A. Abella, T. Alonso, G. Riccardi, and J. H. Wright. 2002. Automated natural spoken dialog. *IEEE Computer Magazine*, 35(4):51–56.

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

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

Jerry R. Hobbs. 1977. Resolving pronoun references. *Lingua*, 44:311–338.

Ashwani Kumar, Susanne Salmon-Alt, and Laurent Romary. 2003. Reference resolution as a facilitating process towards robust multimodal dialogue management: A cognitive grammar approach. *International Symposium on Reference Resolution and Its Application to Question Answering and Summarization*.

F. Landragin and L. Romary. 2003. Referring to objects through sub-contexts in multimodal human-computer interaction. *Seventh Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck'03)*, pages 67–74.

Michael F. McTear. 2002. Spoken dialogue technology: Enabling the conversational interface. *ACM Computing Surveys*, 34(1):90–169.

David Stallard. 2000. Talk'n'travel: A conversational system for air travel planning. *Proceedings of the Association for Computational Linguistics 6th Applied Natural Language Processing Conference (ANLP 2000)*, pages 68–75.

R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. 2004. Natural human-robot interaction using speech, gaze and gestures. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*.

Michael Strube. 1998. Never look back: An alternative to centering. *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1251–1257.

Joel Tetreault and James Allen. 2003. An empirical evaluation of pronoun resolution and clausal structure. *Proceedings of the 2003 International Symposium on Reference Resolution and its Applications to Question Answering and Summarization*, pages 1–8.

Marilyn A. Walker. 1998. Centering, anaphora resolution, and discourse structure. *Marilyn A. Walker, Aravind K. Joshi and Ellen F. Prince: Centering in Discourse*.

Victor Zue, Stephanie Seneff, James Glass, Joseph Polifroni, Christine Pao, Timothy J. Hazen, and Lee Hetherington. 2000. Jupiter: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):100–112.