REFERRING TO WORLD OBJECTS WITH TEXT AND PICTURES

Elisabeth André, Thomas Rist

German Research Center for Artificial Intelligence (DFKI)
D-66123 Saarbrücken, Germany, e-mail: {andre, rist}@dfki.uni-sb.de

ABSTRACT: It often makes sense to employ both text and pictures when referring to world objects. In this paper, we present a model for referring which is based on the assumption that concepts may be activated not only by text, but also by pictures and text-picture combinations. By means of a case study, we demonstrate that failure and success of referring acts can be explained by the user's ability to infer certain links between mental representations and object descriptions. Finally, we show how the model has been incorporated into a plan-based multimedia presentation system by defining operators for concept activation.

1 INTRODUCTION

From a speech act theoretical point of view, referring is a planned action to achieve certain goals (Appelt and Kronfeld, 1987). Although natural language may be the most conventional vehicle for referring, it has been widely accepted that pictures can be used as well. For example, Goodmann (1969) points out that pictures can be employed to refer to both an individual object and the type of which an object is an exemplary of. Moreover, there are good reasons to include pictures in referring acts. Pictures effectively convey discriminating object properties such as surface attributes and shape. If an object can only be discriminated against alternatives through its location, a picture may provide the spatial context of the object. Since depictions are explicit material representations of the world objects to which they correspond, new attributes of the type 'being depicted as ...' are introduced which, in turn, provide an additional source for object discrimination (e.g., the knob which is represented by the black circle ...). Last but not least, several graphical focusing techniques can be applied to effectively constrain the set of alternatives (e.g., arrows, blinking). Unfortunately, there is also a dark side of the picture. An obvious drawback is that pictures do not provide for syntactical devices to distinguish between a reference-specifying and a predication-specifying part since objects and their properties are hardly separable once depicted. Another difficulty is that pictures lack the means to distinguish definite from indefinite descriptions. Thus, it may remain unclear whether a particular object or whether an arbitrary exemplary of a class is depicted. The conclusion we can draw from these considerations is that it often makes sense to employ both text and pictures when referring to domain objects. Pictures may be used in order to simplify verbal reference expressions. On the other hand, ambiguities of pictures can be resolved by providing additional information through text. When analyzing illustrated documents such as assembly manuals and instructions for use, different kinds of referring expression can be found:

Multimedia referring expressions refer to world objects via a combination of at least two media. Each medium conveys some discriminating attributes which in sum allow for a proper identification of the intended object. Examples are NL expressions that are accompanied by pointing gestures and text-picture combinations where the picture provides information about the appearance of an object and the text restricts the visual search space as in "the switch on the frontside".

Anaphoric referring expressions refer to world objects in an abbreviated form (Hirst, 1981) presuming that they are already explicitly or implicitly introduced in the discourse. The presentation part to which an anaphoric expression refers back is called the antecedent of the referring expression. In a multimedia discourse, we have not only to handle linguistic anaphora with linguistic antecedents, but also linguistic anaphora with pictorial antecedents, and pictorial anaphora with linguistic or pictorial antecedents. Examples, such as "the hatched switch," show that the boundary between multimedia referring expressions and anaphora is indistinct. Here, we have to consider whether the user is intended to employ all parts of a presentation for object disambiguation or whether one wants him to infer anaphoric relations between them.

Cross-media referring expressions do not refer to world objects, but to document parts in other presentation media (Wahlster et al., 1991). Examples of cross-media referring expressions are "the upper left corner of the picture" or "Fig. x". In most cases, cross-media referring expressions are part of a complex multimedia referring expresssion where they serve to direct the reader's attention to parts of a document that has also to be employed in order to find the intended referent.

When viewing referring as a planned action, we have to specify which goals underly the use of different types of referring expressions. Appelt and Kronfeld (1987) distinguish between the *literal goal* and the *discourse purpose* of a reference act. Whereas the literal goal is to establish mutual belief between a speaker and a hearer that a particular object is being talked about, the discourse purpose is to make the hearer recognize what kind of identification is appropriate and to have him identify the referent accordingly. When addressing illustrated documents, the question arises of what identification means when domain objects are referred to via pictures (and text). As with language this varies from discourse to discourse. For example, if the user is confronted with a picture showing how to insert the filter of a coffee machine, he has to recognize whether

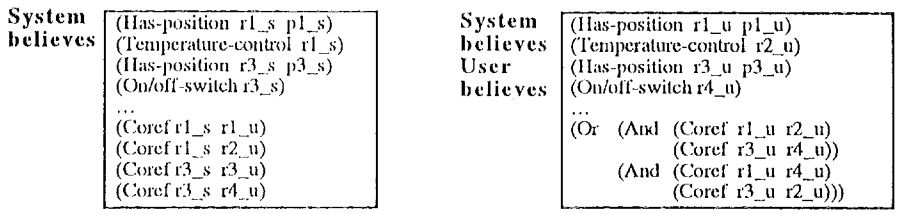

Figure 1: Modelling Example: Different Knowledge Concerning the Identity of Objects

any object with the feature 'being a filter' can be inserted or whether a particular object is meant. In the first case, he has to identify the picture object as an exemplary of a certain class whereas, in the second case, he has to look for something in the world which fits the graphical depiction. In other situations, identification involves establishing a kind of cohesive link between document parts. If the user is confronted with a sequence of pictures showing an object from different angles, he has to recognize that in all pictures the same object is depicted (*pictorial anaphor with pictorial antecedent*). When reading an utterance, such as "the resistor in the figure above," he has to recognize an anaphoric relationship between the textual description and the graphical depiction (*linguistic anaphor with pictorial antecedent*).

Previous work on the generation of referring expressions in a multimedia environment has mainly concentrated on single reference phenomena, such as references to pictorial material via natural language and pointing gestures (Allgayer et al., 1989; Claassen, 1992; Stock et al., 1993) and the generation of cross-media references from text to graphics (McKeown et al., 1992; Wahlster et al., 1993). The aim of this paper is, however, to provide a more general model that explains which kinds of coreferential link between referring expressions, objects of the world and objects of the multimedia presentation have to be established to ensure the comprehensibility of a referring expression.

2 A MODEL FOR REFERRING WITH TEXT AND PICTURES

When referring to domain objects a presentation system has to find intelligible object descriptions which will activate appropriate representations. We assume that representations can be activated in the sense of picking them out of a set of representations which are already available or which have to be built up (e.g., by localizing an object in a user's visual field). Representations can be activated by textual descriptions, by graphical descriptions or by mixed descriptions. Whereas the order in which representations are activated by a text is influenced by the discourse structure, it is less than clear in which order a picture activates representations. If several objects are depicted, the corresponding representations may be activated simultaneously.

2.1 Representations of World Objects

To ensure the transferability of our approach, we don't presuppose a certain knowledge representation language. However, an essential part of the model concerns the distinction between the system's beliefs about the world and the system's beliefs about the user's beliefs. We represent these beliefs in different models. For example, the system may classify a certain object as an espresso machine while it assumes the user regards the object as a coffee machine. Furthermore, we have to consider that the user's and the system's beliefs about the identity of objects may differ. The system may believe that the user has different representations for one and the same object without knowing how they are related to each other. Conversely, it may happen that the user is assumed to have only one representation for objects which the system considers as distinct entities. As a consequence, our models can contain different representations for one and the same world object. We use the predicate

(Coref rep1 rep2)

to express that rep1 and rep2 are representations of the same world object.

Fig. 1 gives an example of how to use the concepts introduced above. Let's start from the following situation taken from an espresso machine domain: The system knows that there are two switches (the temperature control and the on/off switch) and also knows where they are located. Let r1_s and r3_s correspond to the system's internal representations of the switches. The user is assumed to look at the espresso machine and to see two switches. Let r1_u and r3_u correspond to internal representations of the switches which the user builds up when looking at the machine. We assume that the user also knows of the existence of the on/off switch and the temperature control, but is not able to localize them. Let r2_u and r4_u be the user's representations for the temperature control and the on/off switch. The fact that he only knows that one of the switches he sees must be the temperature control and the other the on/off switch can be expressed by means of a disjunction. Either a coref relation holds between r1_u and r2_u and between r3_u and r4_u or conversely, between r1_u and r4_u and between r3_u and r2_u. The connection between the system's representations r1_s and r3_s to the representations the user is assumed to have is also expressed by coreference relations.

2.2 Representation of Descriptions

As mentioned in section 1, descriptions can be composed

of text, graphics and further presentation media. To cope with such descriptions, we associate with each syntactical unit (depictions, noun phrases, etc.) the set of object representations which will be activated by that particular part. The referent of the whole description is then considered as a member of the intersection of all sets resulting from partial descriptions.

An important prerequisite of our approach is that the system explicitly represents how it has encoded information in a presentation. Inspired by (Mackinlay, 1986), we use a relation tuple of the form:

(Encodes means information context-space)

to specify the semantic relationship between a textual or graphical means, and the information the means is to convey in a certain context space. In our approach, the third argument refers to the context space to which the encoding relation corresponds to and not to a graphical language as in Mackinlay's approach. This enables us to use one and the same presentation means differently in different context spaces. For example, a depiction of an espresso machine may refer to an individual machine in one context space, but may serve as a prototypical representative of an espresso machine in another. In addition, we not only specify encoding relations between individual objects, but also specify encoding relations on a generic level (e.g., that the property of being red in a picture encodes the property of being defect in the world).

While it can be assumed that a user reads a text in sequential order, it is often not clear at which times a user looks at a picture. Therefore, it makes not always sense to further distinguish between an anaphor and its antecedent. Fortunately, our approach does not require identifying parts of a presentation as anaphora and antecedents. It suffices to recognize which parts of a description are intended to encode a uniquely determined object. To express such cohesive relationships between presentation parts p1 and p2, we define the predicate:

(EncodesSame p1 p2 c) :=
(Exists w (And (Encodes p1 w c) (Encodes p2 w c)
(Forall v (Implies (Or (Encodes p1 v c) (Encodes p2 v c))
(Coref w v)))))

The first part of this definition expresses that there exists an object w that p1 and p2 encode in the context space c while the second part means that this object w is uniquely determined.

2.3 Links between Representations and Descriptions

In understanding a referring expression, the user has to recognize certain links between activated mental representations, between descriptions and mental representations, and between textual and graphical parts of descriptions. Which links are present in a description and which have to be inferred varies from situation to situation. To illustrate this, let's have a look at a case study carried out in our espresso machine domain where text-picture combinations are used to explain how to operate an espresso machine. We assume that the user is requested to turn the temperature control of an espresso machine. In this case, identification means activating a representation the user builds up when localizing the referent in his visual field. Furthermore, we presume the user knowledge of the espresso machine as in Section 2.1; i.e., the user knows of the existence of the on/off- and the temperature control, has visual access to the two switches in the world but is not able to tell them apart. In the diagrams below, we use the abbreviations ES, C and E for the relations EncodesSame, Coref and Encodes respectively.

In the document fragment shown in Fig. 2, the textual reference expression uniquely determines a referent, but activates a representation (r2_u) which doesn't contain any information to localize the referent. Conversely, the representations activated by the picture contain locative information, but here we have the problem that several object representations are activated to the same extent. Since only the property of being a switch, but not the property of being a temperature control is conveyed by the picture, both switch depictions become possible as antecedents of the textual referring expression.

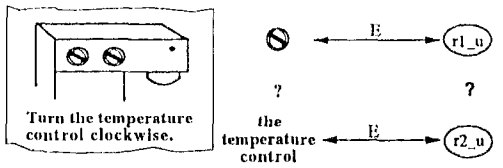

Figure 2: Missing Cohesive Link between Text and Picture

In Fig. 3, the verbal description discriminates the referent from its alternatives by attributes of the world object, namely 'being a switch', and 'being depicted in the figure' and an attribute of the depiction, namely 'being dark'. But, in contrast to the previous example, only one of the representations activated by the picture fits the verbal description. Thus, the user should be able to discover the anaphoric link between the verbal description and the graphical depiction and activate an appropriate representation.

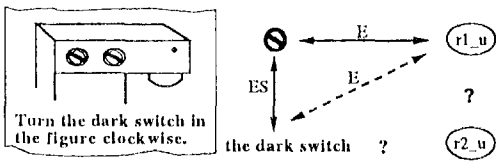

Figure 3: Establishing a Cohesive Link by Incorporating Picture Attributes in Verbal Descriptions

In the previous example, an anaphoric link between text and picture has been established by including pictorial attributes in the verbal description. An alternative is to apply graphical focusing techniques as in Fig. 4. Here, it's very likely that the user will be able to draw a link between text and picture because he will assume that the pictorial and the textual focus coincide. This example also illustrates how the user's knowledge of the identity of objects can be enriched by means of a referring act. The verbal

description without the graphics and the graphical depiction without the text activate different representations of the switch. When considering both text and graphics, the user will conclude that they refer to the same object. Thus, he is not only able to identify the switch as required, he is also able to combine the different representations of the switch into one. Note that this phenomenon can also be explained in terms of centering theory (Grosz et al., 1983). In the example, the preferred center of the picture would coincide with the backward looking center of the text.

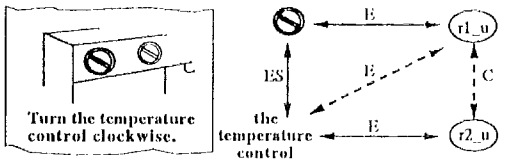

Figure 4: Establishing a Cohesive Link by Correlating Textual and Pictorial Focus

The example shown in Fig. 5 differs from the previous ones in that no correspondency link between picture objects and real world objects can be established. Although the user is able to draw an anaphoric link between the verbal and the pictorial description, he is not able to visually identify the intended referent.

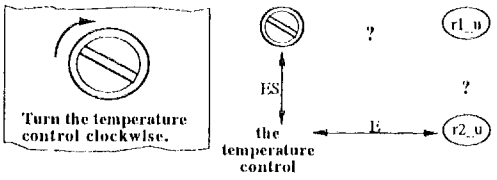

Figure 5: Missing Correspondency between Picture and World

Summing up, it can be said that a referring act is only successful when the description provides an access path to an appropriate representation. The user has to infer such a path from encoding relationships and cohesive links between the parts of a description. As the examples show, the following cases occur: a) If the user does not recognize which picture parts correspond to which world object, the referring act either fails (cf. Fig. 5) or the picture contributes nothing to its success. b) If the relationship between pictorial depictions and verbal descriptions is unclear, the referent can either not be found (cf. Fig. 2) or one of the media has no influence on referent identification. c) If a graphical depiction and a verbal description activate different representations of one and the same object and the user recognizes not only these links, but also a link between the two presentation parts, he is not only able to find the referent, but also able to combine the different representations into one (cf. Fig. 4).

3 USING THE MODEL TO GENERATE REFERRING EXPRESSIONS

In the following, we will sketch how we have integrated the approach into the multimedia presentation system WIP (Wahlster et al., 1993). At the heart of the WIP system is a presentation planner that is responsible for determining the contents and selecting an appropriate medium combination. The presentation planner receives as input a presentation goal (e.g., the user should know where a certain switch is located). It then tries to find a presentation strategy which matches this goal and generates a refinement-style plan in the form of a directed acyclic graph (DAG). This DAG reflects reflects the propositional contents of the potential document parts, the intentional goals behind the parts as well as the rhetorical relationships between them, for details see (André and Rist, 1993). While the top of the presentation plan is a more or less complex presentation goal (e.g., instructing the user in switching on a device), the lowest level is formed by specifications of elementary presentation tasks (e.g., formulating a request or depicting an object). These elementary tasks are directly forwarded to the medium-specific generators, currently for text (Kilger, 1994) and graphics (Rist, and André, 1992).

The content of referring expressions is determined by the presentation planner that also decides which representations should be activated and which medium should be chosen for this. To be able to perform these steps, we need presentation strategies for linking propositional acts with activation acts. An example of such a strategy is [1].

[1] **Header:** (Request S U (Action ?action) Text)
Effect: (BMB S U (Goal S (Done U ?action)))
Applicability Conditions:
(And (Goal S (Done U ?action))
(Bel S (Complex-Operating-Action ?action))
(Bel S (Agent ?agent ?action))
(Bel S (Object ?object ?action)))
Main Acts:
(S-Request S U
(?action-spec (Agent ?agent-spec) (Object ?object-spec)))
Subsidiary Acts:
(Activate S U (Action ?action) ?action-spec Text)
(Activate S U (Agent ?agent) ?agent-spec Text)
(Activate S U (Object ?object) ?object-spec Text)

This strategy can be used to request the user to perform an action. In this strategy, two kinds of act occur: an elementary speech act S(urface)-Request and three activation acts for specifying the action and the semantic case roles associated with the action (Activate). The strategy prescribes text for the subsidiary acts because the resulting referring expressions (?action-spec, ?agent-spec and ?object-spec) are obligatory case roles of an S-Request speech act which will be conveyed by text. For optional case roles any medium can be taken. In addition to strategies for linking propositional and activation acts, we need strategies for different kinds of activation and for establishing Coref- and EncodesSame-relationships. For example, strategy [2] can be used to activate a representation ?r-1 by text and to simultaneously enrich the user's knowledge

about the identity of objects. The strategy only applies if there exists already an image ?pic-obj which encodes ?r-1, the system believes that ?r-1 and ?r-2 are representations of the same world object and if the system's model of the user's beliefs contains ?r-2. If the strategy is applied, the system a) provides a unique description ?d for ?r-2 (main act) and b) ensures that the user recognizes that this description and the corresponding image specify the same object (subsidiary act).

[2] **Header:** (Activate S U (?case-role ?r-1) ?d Text)
Effect: (BMB S U (Coref ?r-1 ?r-2))
Applicability Conditions:
(And (BMB S U (Encodes ?pic-obj ?r-1 ?c))
(Bel S (Coref ?r-1 ?r-2))
(Bel S (Bel U (Thing ?r-2))))
Main Acts:
(Provide-Unique-Description S U ?r-2 ?d Text)
Subsidiary Acts:
(Achieve S
(BMB S U (EncodesSame ?d ?pic-ob ?c)) ?medium)

For a), we use a discrimination algorithm similar to the algorithm presented in (Reiter and Dale, 1992). However, we have investigated additional possibilities for distinguishing objects from their alternatives. We can refer not only to features of an object in a scene, but also to features of the graphical model, their interpretation and to the position of picture objects within the picture, see also (Wazinski, 1992). A detailed description of our discrimination algorithm can be found in (Schneiderlöchner, 1994). Task b) can be accomplished by correlating the visual and the textual focus, by redundantly encoding object attributes, or by explicitly informing the user about a Coref-relationship. Such a Coref-relationship can be established by strategies for the generation of cross-media referring expressions (as in "The left switch in the figure is the temperature control") or by strategies for annotating objects in a figure.

4 CONCLUSION

We have presented a model of referring which is based on the following assumptions: 1) Mental representations of objects may be activated not only by textual, but also by graphical and mixed descriptions. 2) Failure and success of referring acts can be explained by the user's ability to recognize certain links between these mental representations and the corresponding object descriptions. To demonstrate that the model is of practical use for the generation of references, we have defined presentation strategies for concept activation which serve as operators in the plan-based presentation system WIP. WIP is able to generate multimedia, anaphoric and cross-media referring expressions.

ACKNOWLEDGEMENTS: This work is supported by the BMFT under grant ITW8901 8. We would like to thank Doug Appelt for valuable discussions and comments.

REFERENCES

Allgayer, J., Harbusch, K., Kobsa, A., Reddig, C., Reithinger, N. and Schmaucks, D. (1989). *XTRA: A Natural-Language Access System to Expert Systems. Intern. Journal of Man-Machine Studies, 31*, pp. 161-195.

André, E., and Rist, T. (1993). *The Design of Illustrated Documents as a Planning Task.* In M.T. Maybury Ed., *Intelligent Multimedia Interfaces*,The MIT Press, Menlo Park, pp. 94-116.

Appelt, D., and Kronfeld, A. (1987). *A Computational Model of Referring. Proc. of IJCAI-87*, pp. 640-647.

Claassen, W. (1992). *Generating Referring Expressions in a Multimodal Environment.* In R. Dale, E. Hovy, D. Rösner and O. Stock Ed., *Aspects of Automated Natural Language Generation: Proc. of the 6th International Workshop on Natural Language Generation.* Springer, Berlin, pp. 247-262.

Goodman, N. (1969). *Languages of Art.* Oxford University Press, Oxford.

Grosz, B., Joshi, A.K., and Weinstein, S. (1983). *Providing a Unified Account of Definite Noun Phrases in Discourse. Proc. of the 21st ACL*, pp. 44-50.

Hirst, G. (1981). *Anaphora in Natural Language Understanding*. Springer, Berlin.

Kilger, A. (1994). *Using UTAGs for Incremental and Parallel Generation. Computational Intelligence.* to appear.

Mackinlay, J. (1986). *Automating the Design of Graphical Presentations of Relational Information. ACM Transactions on Graphics, 5(2)*, pp. 110-141.

McKeown, K.R., Feiner, S.K., Robin, J., Seligmann, D.D. and Tanenblatt, M. (1992). *Generating Cross-References for Multimedia Explanation. Proc. AAAI-92*, pp. 9-16.

Reiter, E., and Dale, R. (1992). *A Fast Algorithm for the Generation of Referring Expressions. Proc. of COLING-92, 1*, pp. 232-238.

Rist, T., and André (1992). *From Presentation Tasks to Pictures: Towards an Approach to Automatic Graphics Design. Proc. of ECAI-92*, Vienna, Austria, pp. 764-768.

Schneiderlöchner, E. (1994). *Generierung von Referenzausdrücken in einem multimodalen Diskurs.* Diploma Thesis, Universität des Saarlandes, Germany, to appear.

Stock O., and the ALFRESCO Project Team (1993). *ALFRESCO: Enjoying the Combination of Natural Language Processing and Hypermedia for Information Exploration.* In: In M.T. Maybury Ed., *Intelligent Multimedia Interfaces*,The MIT Press, Menlo Park, pp. 197-224.

Wahlster, W., André, E., Graf, W., and Rist, T. (1991). *Designing Illustrated Texts: How Language Production Is Influenced by Graphics Generation. Proc. of EACL-92*, Berlin, pp. 8-14.

Wahlster, W., André, E., Finkler, W., Profitlich, H.J., and Rist, T. (1993). *Plan-Based Integration of Natural Language and Graphics Generation. AI Journal, 63*, pp. 387-427.

Wazinski, P. (1992). *Generating Spatial Description for Cross-modal References. Proc. of ANLP-92*, Trento, Italy, pp. 56-63.