MULTI-MODAL REFERENCES IN HUMAN-COMPUTER DIALOGUE*

Jeannette G. Neal, Zuzana Dobes, Keith E. Bettinger, Jong S. Byoun
Calspan-UB Research Center (CUBRC)
4455 Genesee Street
Buffalo, NY 14225

Abstract

Multi-modal communication is common among humans. People frequently supplement natural language (NL) communication with simultaneous coordinated pointing gestures and drawing on ancillary visual aids. Similar multi-modal communication can facilitate human interaction with modern sophisticated information processing and decision-aiding computer systems. In this paper, we focus on the use of deictic pointing gestures with simultaneous coordinated NL in both user input and system-generated output. Key knowledge sources and methodology for referent resolution are presented. The synergistic mutual disambiguation of simultaneous NL and pointing is discussed as well as a methodology for handling inconsistent NL/pointing expressions and expressions that have an apparent null referent. This work is part of the Intelligent Multi-Media Interface Project [Neal & Shapiro, 1988] which is devoted to the development of intelligent interface technology that integrates speech, NL text, graphics, and pointing gestures for human-computer dialogues.

1 Introduction

Multi-modal communication is common among humans. People frequently supplement natural language (NL) communication with simultaneous coordinated pointing gestures and drawing. Such multi-modal communication can be used very effectively for human-computer dialogue also. Modern information processing and decision-aiding computer systems are complex and require a full range of communication media to facilitate interaction with the human user. As the complexity of the computer systems increase, there must be a comparable increase in the capability of information transfer with the user. The computer/user interface must not only take advantage of multiple media, but must make use of the synergistic properties of these media and minimize the user's cognitive workload. Such multi-media communication is now becoming possible with the development of a variety of I/O devices including fast high-resolution color-graphics displays, pointing devices, touch screens, and speech recognition and production systems.

The Intelligent Multi-Media Interface Project [Neal & Shapiro, 1988] is devoted to the development of intelligent interface technology that integrates speech, natural language text, graphics, and pointing gestures for human-computer dialogues in a flexible, context-sensitive, and highly integrated manner modelled after the manner in which humans converse in simultaneous coordinated multiple modalities. As part of the project, a knowledge-based interface system, called CUBRICON (CUBRc Intelligent CONversationalist) is being developed as a research prototype. The application domain being used to drive the research is that of military tactical air control.

In this paper, we focus on the use of deictic pointing gestures with simultaneous coordinated natural language in both user input and system-generated output. In the CUBRICON system, pointing gestures are part of its unified multi-media language. Pointing gestures are an integral part of the system's grammar and are processed by the multi-media language parser/generator.

Simultaneous coordinated deictic pointing gestures and natural language can be used very effectively in both input and output to reduce cognitive workload for the user. For user input, the combination of deictic pointing gestures with simultaneous natural language is a very efficient means of expressing a definite reference or locative adverbial phrase. For example, a person could simply say "this" or "this SAM" (Surface-To-Air missile) and point to an entity on the display to disambiguate which of several visible SAM systems is meant. Without the pointing gesture, the person would need to compose a NL definite reference that provides sufficient modification to specify the particular SAM intended (e.g, "the SAM system at 12.3 degrees longitude and 50.5 degrees latitude"). The combined natural language and pointing reference is efficient since the cognitive process of generating the dual-media reference would be much shorter than the generation of the NL-only reference. The result is a reduction in cognitive workload for the user.

The combination of pointing with simultaneous natural language can similarly be used by the system in output. In many sophisticated applications systems, CRT displays become complex and fairly cluttered due to the nature of the task domain. Such is the case with tactical air mission planning systems, particularly the geographical maps that provide vital information for the planning process. In such an environment, system-generated spoken natural language accompanied by coordinated pointing (via blinking or highlighting) can be used very effectively to assist the user in locating important entities or information on a busy display. Speech accompanied by coordinated pointing is also very effective in presenting information about activities or events that must occur sequentially in a spatial environment (e.g., planned movements of military forces).

Use of such dual-media references entails certain problems, however. A point by the user can be ambiguous if he points to the area where two or more graphical figures or icons overlap. The user can also inadvertently miss the object at which he intended to point. For output, one of the problems is deciding when pointing is helpful and appropriate, rather than distracting.

* This research is supported by the Defense Advanced Research Projects Agency and monitored by the Rome Air Development Center under Contract No. F30603-87-C-0136.

CUBRICON includes methodology to handle these problems.

Some systems use default techniques to handle ambiguous pointing. These techniques include: (1) a point returns the entity represented by the "foremost" icon where the system uses a mechanism to remember the order in which icons are "painted" on the display (i.e., which are further in the background and which are foremost in the foreground); (2) the icons or entities are assigned weights representing importance and the icon with the largest weight is selected as the interpretation of an ambiguous point; or (3) the icon whose "center" is closest to the location pointed at is selected. Combinations of the such techniques can also be used. A serious disadvantage of the above listed point-interpretation techniques is that it is difficult, if not impossible, for certain icons to be selected via a point reference. Such default techniques have deliberately not been used in the CUBRICON system. CUBRICON's acceptance of NL accompanying a point gesture overcomes the limitations of the above weak default techniques and provides a more flexible referencing capability.

CUBRICON also includes the ability to infer the intended referent when the referring dual-media expression is inconsistent or produces no apparent candidate referents. A dual-media expression is inconsistent when the natural language part of the expression and the accompanying point cannot be interpreted as referring to the same object(s). For example, the user might say "this SAM" and point to an airbase. A dual-media expression has no apparent referent when the user's point touches no icons (i.e., he points to an "empty" area).

The referent resolution problem has been addressed for systems that accept natural language only [Grosz, 1981, 1986; Sidner, 1983]. The problem of ambiguity, including referent ambiguity, is well recognized in natural language understanding [Hirst, 1988]. The problem of correcting reference identification failures during the natural language understanding process has been addressed using a relaxation technique [Goodman, 1985]. Generation of natural language references is addressed by Sondheimer et al. [1986] and McDonald [1986]. In contrast to these efforts, the work discussed in this paper addresses the problem of referent identification and reference generation for language consisting of combined natural language and deictic pointing gestures. Related work includes the development of TEMPLAR [Press, 1986] at TRW and XTRA [Kobsa et al., 1986] at the University of Saarbrucken. The TEMPLAR system seems to provide only for a pointing gesture to substitute for a natural language definite reference within a natural language sentence during input, rather than allow a pointing gesture to also be used simultaneously with a NL reference during both input and output. In the TEMPLAR system, the natural language phrase for the object selected by the point is inserted in the input string to allow the NL parser to complete its processing. Our work is closer to that of Kobsa and colleagues with the XTRA system. XTRA accepts input of simultaneous NL and pointing gestures. Our approach provides for a more diverse set of referent types and resolution knowledge sources.

The next section presents a brief overview of the CUBRICON system. Subsequent sections discuss the knowledge sources used to process these dual-media expressions in input and output, the syntax and interpretation of such expressions used in input, and the process of generating such combined media output.

2 System Overview

The CUBRICON system is intended to imitate, to a certain extent, the ability of humans to simultaneously accept input from different sensory devices (such as eyes and ears), and to simultaneously produce output in different media (such as voice, pointing motions, and drawings). The design provides for input to be accepted from three input devices: speech input device, keyboard, and mouse device pointing to objects on a graphics display. Output is produced for three output devices: color-graphics display, monochrome display, and speech output device. The CUBRICON software is implemented on a Symbolics Lisp Machine using the SNePS semantic network processing system [Shapiro, 1979, 1986], an ATN parser/generator [Shapiro, 1982] and Common Lisp. Speech recognition is handled by a Dragon Systems VoiceScribe 1000.

Subsequent sections of this paper present example sentences that include simultaneous coordinated pointing gestures to objects on a graphics display. Figure 1 shows one of the geographical displays that was generated by the CUBRICON system. The example sentences in this paper are expressed with simultaneous pointing to objects on such a display. CUBRICON also generates other types of displays including other visual illustrations, tables, and forms.

The CUBRICON system includes several critical knowledge types that are used during language understanding and generation: (1) task domain knowledge, (2) dual-media language knowledge, (3) sentential syn-

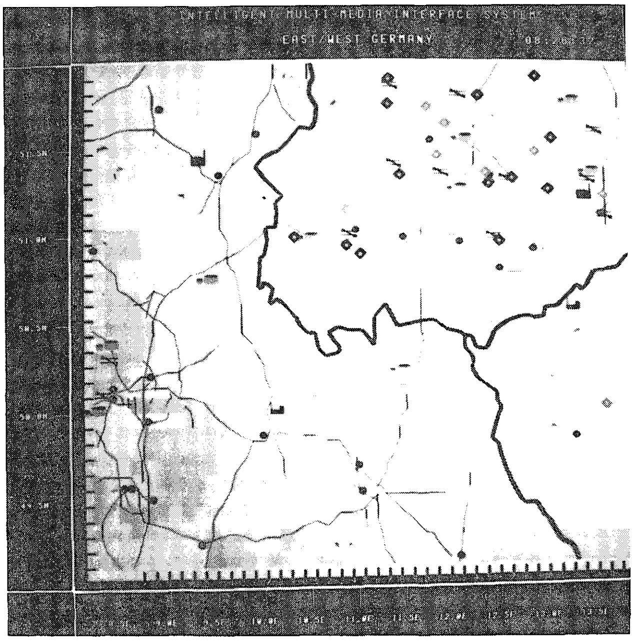

Figure 1 Example CUBRICON Display

tax and semantics, and (4) the discourse context. These knowledge types are discussed in the next section.

3 Resources for Referent Determination and Reference Generation

3.1 Domain Knowledge

Task domain entities with information about them are represented declaratively in a semantic network knowledge base. The associated information includes information relevant to mission planning as well as information that is relevant for communication purposes. Task domain entities include airbases, surface-to-air missile (SAM) systems, fuel storage facilities, and targets. The knowledge base is structured as an AKO ("a kind of") hierarchy. The hierarchy is a tree structure with each node representing an entity type or class. Associated with each entity type are attributes and possibly subparts. Examples of attributes are an entity's disposition (friendly vs. enemy) and its geographical location, when appropriate. Relations between concepts are also represented in the knowledge base, for example, the relation between an aircraft type and the type of munitions that it carries.

3.2 Dual-Media Language Knowledge

The CUBRICON grammar defines the language accepted as input and used for output. According to this grammar, pointing gestures are allowed with (or in place of) a noun phrase or a locative adverbial phrase. Case frames associated with the verbs of the lexicon are used for language understanding. Constraints on the slots of the case frames are used during referent resolution for the dual-media noun phrases and locative adverbial phrases.

Another form of semantic language knowledge is the association of words and graphics forms/icons with domain entities in the knowledge base. Each entity type and instance represented in the knowledge base can have one or more names as attributes. Such names would include "fighter base" for a certain type of airbase and "SA-2" for a certain type of SAM. If appropriate, each entity in the knowledge base can have a graphical form or icon for expressing the entity visually.

3.3 Sentential Context

When analyzing user input, the immediate linguistic context (syntax and semantics) of the sentence processed thus far is used in the process of interpreting the remainder of the sentence, including dual-media references. The semantic structures that are particularly useful are:

1. the case frame associated with the main verb of the sentence.
2. a type or category named in a noun phrase or locative adverbial phrase.
3. a property or attribute named in a noun phrase.
4. a relation expressed in a noun phrase.

The use of these semantic structures is discussed in Section 6.

3.4 The Discourse Model

The attentional discourse focus space [Grosz, 1978, 1986; Sidner, 1983; Grosz and Sidner, 1985] is a key knowledge structure that supports continuity and relevance in dialogue. The CUBRICON system tracks the attentional discourse focus space of the dialogue carried out in multi-media language and maintains a representation of the focus space in two structures: (1) a main focus list and (2) a set of ancillary focus lists called virtual displays. The main focus list includes those entities and propositions that have been explicitly expressed (by the user or by CUBRICON) via natural language, pointing, highlighting, or blinking. A virtual display is a list of all the objects that are "in focus" because they are visible in a given window on one of the displays. CUBRICON maintains one virtual display per window.

The dialogue focus space representation is used in both understanding user input and generating system output. When processing user input, the attentional focus space representation is used for determining the interpretation of anaphoric references [Sidner, 1983] and definite descriptive references [Grosz, 1981]. In the case of a definite reference, if an appropriate referent is not found in the main focus list, then CUBRICON consults the virtual displays.

4 Multi-Media Language Understanding

CUBRICON accepts coordinated simultaneous natural language and pointing (via a mouse device). The user can input natural language (NL) via the speech device and/or the keyboard. Input from the NL and pointing devices is accepted and fused into a compound stream maintaining the information as to which point gesture(s) occurred with (or between) which word(s) of the sentence.

The CUBRICON Parser/Interpreter is an ATN that accepts the compound stream produced by the Input Coordinator and produces an interpretation of the compound stream. Each noun phrase or locative adverbial phrase can consist of zero or more words of text along with zero or more pointing references to icons on the display (there must be at least one point or one word). The pointing input that is a component of a noun phrase or locative adverbial phrase can occur anywhere within the phrase.

From observing people using mouse points, it seems that mouse points commonly

(a) substitute for an entire noun phrase: "What is the status of <point>? "

(b) substitute for the head noun: "What type of SAMs are these <point>$_1$ <point>$_2$ <point>$_3$?"

(c) are used in conjunction with a complete NL noun phrase: "Display the status of this <point> airbase."

The objects that can be referenced via pointing in the CUBRICON system are of four types:

1. a geometric point represented by a pair of coordinates on a map or graph;
2. an entity represented graphically;
3. a table entry;
4. a window on a display.

Pointing gestures alone can be categorized according to the following: (1) mouse click on intended icon(s) only, (2) mouse click on the region where the extents of two or more icons overlap and not all were intended to be selected, or (3) mouse click misses the intended icon(s) altogether. The problem is in determining the intended referent(s) of a combined natural language and pointing reference. In the second case listed above, the mouse point alone is ambiguous. In the third case, the point gesture has no immediate referent. When pointing gestures are used in the context of natural language dialogue, several knowledge sources can be applied to the problem of identifying the intended referent. The CUBRICON methodology for referent resolution is discussed in the next section.

5 Referent Resolution Methodology

CUBRICON uses the several knowledge sources discussed in Section 3 when determining the referent of a combined natural language and pointing reference. For "ill-formed" expressions in which the interpretation of the NL is inconsistent with the object(s) touched by the point and those expressions which apparently have a null reference (e.g., the user points at an empty area), CUBRICON infers the intended referent according to the methodology discussed at the end of this section. The following examples illustrate the CUBRICON methodology.

The first example depends primarily on the use of the task domain knowledge represented in the knowledge base as well as the ancillary graphical discourse focus list (refer to Section 3.3).

USER: "What is the name of this <point> airbase?"

When the phrase "this <point> airbase" is parsed, the system uses the point coordinates to determine which icons are touched by the point. The virtual display is then searched in order to retrieve the semantic network nodes representing the objects which were graphically displayed by the "touched" icons. Within the knowledge base, the system has a representation of the category to which each object belongs as well as a representation of the airbase concept. From the hierarchy of the knowledge base, the system determines which of the objects selected by the point gesture are airbases and discards the others. If the user has pointed at a minimum of one airbase, then the system uses this (these) airbase instance(s) as the referent of the dual-media noun phrase. Discussion of the situation in which the user has pointed at no airbases is deferred to the end of the section.

The second example entails the use of the syntax and semantics of the sentence processed thus far, along with the knowledge base, to determine the referent of the phrase "this <point>". Here the concept of "mobility" is the critical item of information .

USER: "What is the mobility of this <point> ?"

From the virtual display, the system retrieves the objects represented by the icons which were touched by the point gesture. From the syntax of the noun phrase "the mobility of this <point>" and the semantics of the word "mobility" as represented in the knowledge base, the system deduces that mobility is a property (as opposed to a subpart or some other possible relation that could exist between the concepts mentioned) of the object mentioned in the prepositional phrase. The system then determines which of the objects selected by the point gesture have a property called mobility by consulting the knowledge base. The other objects selected by the point are discarded. The resulting set is used as the referent of the phrase "this <point>".

In the next example sentence, the case frame plays an important role in referent determination.

USER: "Are these battalions <point>$_1$ <point>$_2$ <point>$_3$ based here <point> ?"

In order to determine the interpretation of the phrase "here <point>", the use of the case frame for the verb phrase "is based" is necessary. If we consider the phrase "here <point>" alone, the interpretation is unclear. Should it be a location represented by a pair of coordinates, or should it be some institution at the location indicated by the deictic reference? The case frame of the verb phrase "is based" provides the necessary information. This case frame requires an agent and an object. The object must be an institution with which the agent is (or can be) officially affiliated. The knowledge base provides information concerning what types of entities are based at what types of facilities or institutions. Thus the phrase "here <point>" of the example sentence is interpreted as the institution at the location specified by the <point> due to the constraints of the verb's case frame. If the user's point gesture touches no graphic icon, then the system infers the intended referent, as discussed in the next paragraph.

In the above paragraphs, we deferred discussion of the event in which the interpretation of natural language together with the point reference results in an apparent null referent. This event can occur in two ways: (1) the user's point touches at least one icon, but it (they) is (are) inconsistent with the natural language expression (e.g., the user says "airbase" but points to a SAM) or (2) the user points at a location on the display which contains no objects. CUBRICON includes methodology to infer the intended referent in both of these situations. CUBRICON uses the information from the sentence parsed and interpreted thus far as filtering criteria for candidate objects. The system searches in the vicinity of the location of the user's point to find the closest object(s) that satisfy the filtering criteria. If one is found, then the system responds to the user's input (e.g., command for action, request for information), but also indicates to the user that the object to which he pointed was not consistent with the natural language phrase that he used and states the inferred referent. In the event that no qualified object is found in the vicinity of the user's point, then an appropriate response is made to the user with a request for him to restate his input.

The methodology described in this section provides CUBRICON with the ability to determine the referent of expressions that consist of natural language and pointing gestures. This methodology handles both well-formed expressions as well as expressions in which the user's point is inconsistent with the accompanying natural language.

6 Multi-Media Reference Generation

CUBRICON has the ability to intelligently use combined pointing and natural language references when responding. The system currently points at an object displayed on a CRT by blinking the object. We are considering other pointing techniques such as displaying a blinking arrow that points to the desired object and displaying a blinking circle around the desired object. The algorithm that CUBRICON uses to generate an expression for a given entity is as follows:

1. if the entity to be expressed is the most salient one of its gender and number according to the dis-

course focus list, then express the entity with a pronoun of the appropriate gender, number and case.

2. else if the entity to be expressed is currently displayed on one of the CRTs (determined by consulting the virtual display), then express the entity by the natural language phrase "this XXXX" with simultaneous "pointing" to the entity on the display. The name for the entity represented by "XXXX" is selected from the name of the basic level category [Peters & Shapiro, 1987] to which the entity belongs.
3. else if the entity to be expressed is the most salient one of its kind according to the discourse focus list, then express the entity with the definite determiner "the" followed by the name of the class.
4. else generate the most specific reference possible for the entity.

7 Current Status and Future Direction

The work discussed in this paper has been implemented in the CUBRICON system. Our current discrete speech recognition system will be replaced by a continuous speech recognition system in the near future. When this change occurs, we anticipate that we may need a more sophisticated method of coordinating the timing of the individual words of an input sentence with the user's mouse-pointing gestures. We anticipate two possible problems: (a) accounting for the "speech recognition delay" (the delay between the time a word is spoken and the time it is available to the processor) since mouse-point gestures entail no delay and (b) the occurrence of pointing gestures that are not expressed by the user in coordination with their corresponding natural language phrase (if it exists).

Additional work needs to be done on the question of when the system should generate a reference in combined natural language and pointing. Generation of such references should depend on a variety of factors such as: the modality of the user's input, the level of complexity or clutter on the graphics display, the level of sophistication of the user, and attributes of the discourse context such as the number of times a given entity has recently been referenced.

8 Summary

Multi-modal communication is common among humans. People frequently supplement natural language communication with simultaneous coordinated pointing gestures and drawing on ancillary visual aids. Such multi-modal communication can be used very effectively for huma-computer dialogue also. The Intelligent Multi-Media Interface Project [Neal & Shapiro, 1988] is devoted to the development of intelligent interface technology that integrates speech, natural language text, graphics, and pointing gestures for human-computer dialogues. This paper focused on the use of deictic pointing gestures with simultaneous coordinated natural language in both user input and system-generated output. We discussed several critical knowledge types that are used during multimedia language understanding and generation: (1) task domain knowledge, (2) dual-media language knowledge, (3) sentential syntax and semantics, and (4) the discourse context. A referent resolution methodology for processing dual-media input references was discussed. This methodology handles the synergistic mutual disambiguation of simultaneous natural language and pointing as well as inferring the referent(s) of inconsistent NL/pointing expressions and expressions that have an apparent null referent. We also presented a methodology that supports context-sensitive generation of deictic dual-media references based on the above knowledge sources. The work discussed in this paper has been implemented in the CUBRICON system.

References

[Goodman, 1985] B.A. Goodman, "Repairing Reference Identification Failures by Relaxation," *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, Chicago, Ill., pp. 204-217, 1985.

[Grosz, 1978] B. J. Grosz, "Discourse Analysis," in D. Walker (ed.), *Understanding Spoken Language*, Elsevier North-Holland, New York, pp. 229-345, 1978.

[Grosz, 1981] B.J. Grosz, "Focusing and Description in Natural Language Dialogues," in *Elements of Discourse Understanding*, A.Joshi, B. Webber, & I.Sag (eds.), Cambridge Univ. Press, pp. 84-105, 1981.

[Grosz and Sidner, 1985] B.J. Grosz and C.L. Sidner, "Discourse Structure and the Proper Treatment of Interruptions," *Proc. of IJCAI-85*, pp. 832-839, 1985.

[Grosz, 1986] B.J. Grosz, "The Representation and Use of Focus in a System for Understanding Dialogs," in *Readings in Natural Language Processing*, B.J. Grosz, K.S.Jones, B.L.Webber (eds.), Morgan Kaufmann Pulishers, pp. 353-362, 1986.

[Hirst, 1988] G. Hirst, "Semantic Interpretation and Ambiguity," *Artificial Intelligence*, Vol. 34,No.2, pp.131-177, 1988.

[Kobsa et al., 1986] A. Kobsa, J. Allgayer, C. Reddig, N. Reithinger, D. Schmauks, K. Harbusch, W. Wahlster, "Combining Deictic Gestures and Natural Language for Referent Identification," *Proceedings of the 11th International Conference on Computational Linguisticsm*, Bonn, FR Germany, 1986.

[McDonald, 1986] D. McDonald, "Description Directed Control: Its Implications for Natural Language Generation," in *Readings in Natural Language Processing*, B.J. Grosz, K.S. Jones, B.L. Webber (eds.), Morgan Kaufmann Publ., pp. 519-537, 1986.

[Neal and Shapiro, 1988]J.G. Neal and S.C. Shapiro, "Intelligent Multi-Media Interface Technology," *Proceedings of the Workshop on Architectures for Intelligent Interfaces: Elements and Prototype*, Lockheed AI Center, Monterey, CA. pp. 69-91, 1988.

[Peters and Shapiro,1987] S.L. Peters and S.C. Shapiro, "A Representation for Natural Category Systems," *Proc. of IJ-CAI-87*, Milan, Italy, pp.140-145, 1987.

[Press, 1986] B. Press, "The U.S. Air Force TEMPLAR Project Status and Outlook," *Western Conf. on Knowledge-Based Engineering and Expert Systems*, Anaheim, CA, pp. 42-48, 1986.

[Shapiro, 1979] S.C. Shapiro, "The SNePS Semantic Network Processing System". In N. Findler, ed. *Associativ Networks - The Representation and Use of Knowledg by Computers*, Academic Press, New York, pp. 179-203, 1979.

[Shapiro, 1982] S.C. Shapiro, "Generalized Augmented Transition Network Grammars for Generation from Semantic Networks," *AJCL*, Vol. 8, No. 1, pp. 12-25, 1982.

[Shapiro and Rapaport, 1986] S.C. Shapiro and W. Rapaport, "SNePS Considered as a Fully Intensional Propositional Semantic Network," *Proc. of AAAI-86*, pp. 278-283; in G. McCalla & N. Cercone (eds.) *Knowledge Representation*, Springer-Verlag Pub, 1986.

[Sidner, 1983] C.L. Sidner, "Focusing in the Comprehension of Definite Anaphora," in *Computational Models of Discourse*, M. Brady & R.C. Berwick (eds.), The MIT Press, pp. 267-330, 1983.

[Sondheimer and Nebel, B, 1986] N.K. Sondheimer and B. Nebel, "A Logical-Form and Knowledge-Base Design for Natural Language Generation," *Proc. of AAAI-86*, pp. 612-618, 1986.