A Model for Multimodal Representation and Processing for Reference Resolution

Ali Choumane
IRISA/Cordial, University of Rennes 1
6 rue de Kerampont, BP80518
22305 Lannion, France
ali.choumane@irisa.fr

Jacques Siroux
IRISA/Cordial, University of Rennes 1
6 rue de Kerampont, BP80518
22305 Lannion, France
jacques.siroux@univ-rennes1.fr

ABSTRACT

We present a model for dealing with designation activities of a user in multimodal systems. This model associates both a well defined language to each modality (NL, gesture, visual) and a mediator one. It takes into account several semantic features of modalities. Functions link objects from each modality to another, and allow reasoning and referent identification. Processing algorithms related to each language are developed.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Multimodal Communication

Keywords

Multimodal human-computer communication, dialog, reference, speech input, gesture

1. INTRODUCTION

We situate the present work[1] within the general framework of human-machine multimodal dialog systems [8]. The aim of such systems is to allow users to access services. To express their goals to the system, users designate objects by acting on available modes and modalities: oral, language, gesture, etc. We name these designations "referential activities". An important role of a system is to recognize and understand these referential activities.

The system is confronted to many difficulties. The designation activity of the user is not reliable. Indeed, ambiguities, errors, and hesitations lead to "noise" or misunderstandings which are likely to be worsened by the hardware devices and the programs of the system.

In this paper, we are interested in input to human-machine multimodal communication systems. A system, in order to understand the user's goal, must correctly merge inputs which come in from different modes. A critical point of this merge is the resolution of referential expressions (REs). We propose an elegant model for representing and processing multimodal inputs. This model encodes each modality in a separate formal language. Transition functions between languages allow reasoning and referent identification. Appropriate algorithms are developed depending on the reference types and the language concerned. Semantic information is used in and between languages. Histories are helpful for the processing of some designations.

After a section in which we more precisely present framework we are interested in and the problem addressed, we explain the main elements of the proposed solution. We show how multimodal inputs can be represented in our model and processed using the transition functions and algorithms. Finally, we discuss the interest of such a model for multimodal systems.

[1]This work is partially financed by the grant 2116B2-9/ARED 1800 of the regional council of Brittany, France.

MISI'07, November 15, 2007, Nagoya Japan

2. CURRENT SYSTEM AND APPLICATION

2.1 Georal Framework

The framework Georal tactile system [10] has been implemented on the multiagent platform DORIS [6]. Georal is a multimodal system principally used to provide information of a touristic and geographical nature. Users can ask for information about the location of places of interest (beach, campsite, château, church, etc.) by specifying a place, a zone (particular geographical or cartographical element: river, road, city, etc.) (figure 1).
Georal offers the user the following modes and modalities:

- Oral input as well as output to the system. Users formulate their requests and responses to the system by voice and in natural language (NL) in a spontaneous manner (no particular instructions of elocution). System output is also given by speech synthesis to the user.
- Visual mode: the system displays a map of a region on the screen. This map contains the usual geographical and touristic information: cities, roads, rivers, etc. Zooming effects, highlighting, and flashing allow the system to focus the user's attention.
- Gestural mode by a touch screen: the user can designate elements displayed on the screen by various types

of gesture (point, zone, line, etc.).

Figure 1: *A part of the map displayed on the Georal screen*

2.2 Georal Dialog Model

Our dialog model is inspired by the model of Bilange [1]. A dialog with Georal consists of one or several transactions. The transaction level is used to reflect the structuring of the application in the dialog. One transaction, made up of exchanges, is concerned by one theme. The exchanges are used to carry out either a transfer of information, or a clarification or a meta-discursive function. An exchange consists of communication turns for the user and the system. The user's communication turn consists of an oral utterance (contains a dialog act DA) and/or a gesture input. The system's communication turn consists of an oral output (by voice synthesis) and a display on the screen.

3. DEALING WITH REFERENCE RESOLUTION

We deal with inputs of a multimodal system. In a user's communication turn, the problem for the system is to resolve REs, i.e. to find the referent of a symbol in a modality using information present either in the same or in other modalities.

The user, in her communication turn encodes on the different modalities a communication act (CA). It conveys the user's intention and references to necessary objects. Both the user intention and designated objects are required to realize the user's goal. Therefore, the system must reconstitute this CA by identifying in a precise and non-ambiguous manner the referents designed by the user. Typically, this reconstitution must take into account the possibility of incorrect results in speech recognition and understanding [9] as well as in gesture interpretation. It has to use linguistic methods [7] and specific algorithms for gesture. Choumane et al. have shown in [3, 2] that this is not sufficient. Thus, we have presented elements to take this into account to produce a complete CA like the visual perception (degree of salience [5]) and semantic relationships.

Within the framework presented, Choumane et al. [3] have proposed a definition of a general model for RE resolution. This model is based on two main principles. On the one hand, it associates both a well defined language to each modality (NL, gesture, visual) and a mediator one. On the other hand, functions link objects from each modality to another allowing reasoning and referent identification. The languages allow us to represent, for each communication turn, entities or objects resulting from modalities. The current context and the interaction history are memorized using those languages. Treatments are associated to each modality (e.g. anaphora treatment for NL). Specific treatments are setup to determine the referents named on several modalities.

The purpose of this paper is to develop the proposed model in-depth and to set up the use of semantic information in processing steps to insure the reference resolution.

4. THE PROPOSED MODEL

4.1 Knowledge Representation

To make explicit the representation of different parts of CA, we have proposed to represent each modality in a separate language and to maintain relations between them. In the following, we present a brief definition of languages used and we show how multimodal input can be stored in the model.

4.1.1 L Language

The L language is a representation of the NL in which we have added elements required for other processes. This language (L) allows us to represent objects described in the NL part of the multimodal input. The lexicon and syntax of L are those of the natural language in the specific application of Georal system. General knowledge is required in the interpretation of maps like "Lannion is a city", etc. L contains a logic representation of the user utterance (it shares elements like referential expressions resolved or not with their types, user's goal, etc.). Moreover we associate to L the syntactic analysis tree. Semantic information is used in L processing to filter linguistic expressions and to take into account "imprecise" user input (cf. section 4.2).

4.1.2 T Language

The T language contains information about gestures. The basic lexicon of the T consists of dots, lines, and curves. The syntax of T is obtained by concatenation of elements from the basic lexicon, for example, two basic lines may constitute a polyline (or line) gesture. We assume that there are three kinds of gesture primitives (in Georal system): dot, polyline and zone. Indeed, these three primitives can be represented using the primitives of T. These primitives are the result of rebuilding of the recognized list of points (figure 2), (for more details see section 4.4). For representing more complex gesture type, we should enrich the T lexicon.

Notice that polyline gestures may be segmented into significant partitions, depending on important changes in the gestural trajectory. Therefore T represents these different partitions.

4.1.3 P Language

P corresponds to the maps displayed on the computer screen. It is the internal representation of the common vi-

sual context (definition below) between the user and the system. The lexicon of P consists of graphical objects that allow us to represent all Georal screen objects like dot, icons, lines, curves, etc. The relationships between the basic categories of P constitute its syntax. We notice that there is a common part of lexicon types between P and T. This is due to the way that an object is designated. Indeed, in a system which shares objects like dots, polylines, etc. with the user, we can obviously look at dots, polylines, etc. gesture types.

Common Visual Context
The common visual context between user and system is important for understanding user input. We model it in three layers:

1. The first layer is the map displayed on the Georal screen.
2. The second one is the coding of objects displayed on the screen. This internal representation associates a vector to every object which contains: the display name on the screen (it can be different to the database name), color, form, size, and salience. These characteristics are used when designating an object by speech and/or by gesture (more details in [2, 4]). The salience of a graphical object consists of its visual and contextual weight. To determine these weights, we use a salience distribution algorithm to objects in the common visual context [2, 4] which distinguishes two moments of salience uses: saliences initialization at the beginning of each dialog and saliences modification during interaction. An object with its vector is an element of P.
3. The third one is the relationship between objects represented in the second layer. There are syntactic rules of P.

Notice that during the interaction with Georal, the display screen may be modified: objects may be added, characteristics of existing objects may be updated, etc.

4.1.4 G Language

G is used as a mediator language between L, T and P. This language allows to construct graphical objects, to confront and to unify, if needed, elements coming from the different languages. The lexicon and syntax of G consist of the partial ones of L, T, and P languages. G contains graphical objects like dots, polylines, intersection, inside, etc. The G language is necessary to bridge L and T elements to P ones. On the one hand, G represents gesture graphical objects that can't be represented immediately in P, for instance, elements designated by a composite gesture, unspecified map element, etc. On the other hand, G represents some linguistic objects, for instance, *along* (in the NL CA part: *along this river*) would correspond to a G representation that must be founded in P. We claim that linguistic objects in G, as along, inside, between, right etc. denote spatial relations between referents in question. They will be confronted with the gesture objects for computing the geometry of physical search space. The case of the incoherence problem between modalities will be resolved in G depending on information encoded on the different modalities.

4.2 Semantic information use

We use semantic information for L processing and for the transition from L into G (cf. section 4.4) basing on three axis: syntactico-semantic filter, semantic network, and geometrical projection.

We use syntactico-semantic filter to detect incoherence problem in user utterance [11]. For example, the spatial preposition *along* should precede a word that refers to an object of line or polyligne type. This filter allows us to take into account speech recognition errors.

The semantic network contains information about synonym, hyponym, etc. of a word. For example, *river* is an hyponym of *stream*. Thus, when the user asks about something near to a *stream*, then we take into account all screen objects of *river* type too (we use EuroWordNet).

We project some geometrical objects into application objects using hierarchical relations between objects analogically to linguistic synonym and hyponym. This allows us to address inputs like *give me campsite along this line* in which *river, road* are hyponyms of *line*. Thus, the system take into account all *river* and *road* objects when resolving the RE *this line*.

4.3 Histories

The history is a centralized, structured, and synchronized representation of multimodal dialog. This representation of histories allows us to know at every moment "who said and acted when, how, and in which context". We have represented these histories using XML in which its grammar follows the Georal dialog grammar. NL history is used for anaphora resolution, visual history is used for graphical anaphora resolution, etc.

4.4 Data Flow Processing

We describe in this section how a multimodal input can be represented and processed in our model (figure 2). We illustrate this process by the following example (in the Georal system):

Example 1 *I would like campsites on the east of this stream*, uttered jointly with a designation polyline gesture on the screen.

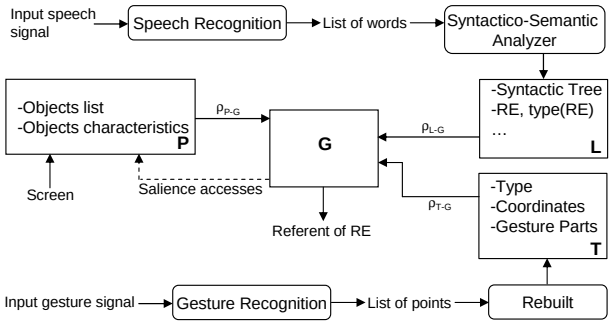

Figure 2: *Steps to process multimodal input*

After the speech recognition and the syntactico-semantic analysis (it includes the filter to detect incoherence in the RE *east of this stream*) steps, the following elements are stored in the L language: the syntactic tree of the utterance, the RE *east of this river*, its type is probably *deictic*, and the DA predicate *request(campsite, east(river (X)))* representing the utterance. X means that the location is not yet

recognized, but that it is of *river* type. Notice that, the above RE consists of the sub-RE *this river* and the spatial preposition *east of*.

In parallel to the first step, we start the recognition and the addressing of the input gesture to provide the following elements to the T language: the gesture type *polyligne* and coordinates *((x1,y1), (x2,y2), ...)* (the gesture is not segmented in the case of a simple trajectory without important changing in the curvature). The screen is encoded in the P language: objects and their characteristics are available.

At this moment, some of the encoded elements in L, T, and P are to be translated into G. This merges input modalities and resolves RE in G. We present the transition functions basing on the example shown above. Some transition functions are resolved using different algorithms depending of parameters type and input category (with or without gesture):

- We use the algorithm proposed in [2] to resolve oral designation, without gesture, of objects in the visual context. This algorithm uses elements from two of the languages in the model: L and P.
- We use also the algorithm proposed in [4] to resolve designation by oral jointly to gesture (example 1). This algorithm uses elements from three of languages of the model: L, T, and P.

ρ_{L-G}: transitions from L into G.

For each syntactic category of L there is a corresponding type in G, these transitions use access to database which contains the representation of touristic objects (campsites, beaches, hotels, ...):

$\rho_{L-G}(campsite) = campsite(icon1, black, ...)$

$\rho_{L-G}(stream) = \begin{cases} river(polyline, blue, ...) \\ stream(polyline, blue, ...) \end{cases}$

$\rho_{L-G}(stream)$ uses the semantic network of *stream*.

$\rho_{L-G}(east) = right$

$\rho_{L-G}(this\ stream) = unknown$

$\rho_{L-G}(east(this\ stream)) = right(\rho_{L-G}(this\ stream))$

ρ_{T-G}: transitions from T into G.

For each set of points we "match" an object (referent) on the map.

$\rho_{T-G}(((x1, y1), (x2, y2), ...)) = object(riverName, blue, ...)$

ρ_{T-G} is resolved using the algorithm concerned by designation with gesture. It takes into account P, the object types founded by $\rho_{L-G}(stream)$, and designation probabilities (more details in [4]).

By inference:

$\rho_{L-G}(this\ stream) = object(riverName, blue, ...)$ (thanks to the gesture coupled with the expression *this stream*)

Now, we determine the referent of *east of this stream* for computing the physical search space. This problem consists in resolving $right(object(riverName, blue, ...))$ (we obtain this predicate by unification in $right(\rho_{L-G}(this\ stream))$.

This step is a spatial reasoning problem to determine a region on the *right* of an object (*stream* in this case).

5. CONCLUSION AND FUTURE WORKS

We have proposed a model for the representation and processing of multimodal inputs. This model allows us to deal with designation activities of a user in multimodal system. It associates both a well defined language to each modality (NL, gesture, visual) and a mediator one. Moreover, functions link objects from each modality to another allowing reasoning and referent identification. We note two main roles of the proposed model:

- It allows a formal representation of the multimodal inputs. It selects and adds required information during the data flow processing for each language.
- It allows inferences based on the encoded information thanks to the transition functions. These inferences ensure a good merge of input modalities. Before performing the transitions into the G language, local processing in each language is setup. After the transitions, global processing which concern several languages is installed in G.

Semantic information is used in some processing like syntactico-semantic filter, semantic network, and geometrical projection. These allow us to take into account some speech recognition errors and imprecise user input. An algorithm for oral designation resolution has been developed in the same way as an algorithm for resolving designation by gesture. Both algorithms use a salience distribution algorithm to object in the visual context depending on defined characteristics and on interaction. The proposed transition functions and inferences are under development. We are developing a spatial reasoning logic theory, basing on the mereotopology, which aims to compute the physical search space.

6. REFERENCES

[1] E. Bilange. *Dialogue personne-machine. Modélisation et réalisation informatique.* Hermes Paris, 1992.

[2] A. Choumane. Traitement de désignations orales dans un contexte visuel. In *Récital07*, volume 1, pages 479–488, 2007.

[3] A. Choumane and J. Siroux. Toward a generic model including knowledge and treatments for multimodal reference resolution. In *Inscit2006*, volume 2, pages 298 – 302, Spain, 2006.

[4] A. Choumane and J. Siroux. Interpretation of multimodal designation with imprecise gesture. In *IE07*, Germany, September 2007.

[5] F. Landragin. Referring to objects with spoken and haptic modalities. In *ICMI*, page 99, USA, 2002.

[6] J. L'Hour, O. Boëffard, J. Siroux, L. Miclet, F. Charpentier, and T. Moudenc. Doris, a multiagent/ip platform for multimodal dialogue applications. In *ICSLP*, pages 3049–3052, Korea, 2004.

[7] R. Mitkov. *Anaphora Resolution.* Pearson Education, 2002. isbn: 0-582-32505-6.

[8] S. Oviatt. Ten myths of multimodal interaction. *Commun. ACM*, 42(11):74–81, 1999.

[9] S. Qu and J. Y. Chai. Salience modeling based on non-verbal modalities for spoken language understanding. In *ICMI*, pages 193–200, USA, 2006.

[10] J. Siroux, M. Guyomard, F. Multon, and C. Rémondeau. Multimodal references in georal tactile. In *35th Meeting Of the ACL*, Spain, 1997.

[11] C. Vandeloise. *L'espace en français.* Editions du Seuil, Paris, 1986.