Multimodal References in GEORAL TACTILE

Jacques Siroux
Université de Rennes I, IRISA/IUT Lannion, ENSSAT, 6, rue de Kerampont
BP 447 F-22305 Lannion Cedex France
siroux@enssat.fr

Marc Guyomard
Université de Rennes I, IRISA/ENSSAT 6, rue de Kerampont
BP 447 F-22305 Lannion Cedex France
guyomard@enssat.fr

Franck Multon and Christophe Rémondeau
ENSSAT Lannion France

Abstract

The paper specifically presents how linguistic (oral) and tactile references are dealt with in the GEORAL system which has already been described in other papers. In this system, users can formulate their queries and provide their responses using the oral (linguistic) modality and the tactile modality separately or together. We describe the referential phenomena which occur in such a context and we point out why the oral modality has to be the basis of the processing of the references and why robustness problems have to be dealt with. We then provide details about the three steps of the reference processing (linguistic analysis, processing of the tactile events and merging process) as well as the modeling of the communicative acts used in the system (as planning operators).

Introduction

Adding a new modality in an existing oral dialogue system poses many interesting problems. Among these, those which concern the referential phenomena deserve to be quoted. How do the users designate the referents? If several modalities are used, how can we model the various activities on each modality? How to exploit them? how to deal with the performance problem (robustness)?... All these questions arise, but not all have complete answers despite numerous work done in the domain [2, 12].

In order to eliminate some drawbacks of speech recognition, we have added a touch screen to an oral system [7]. Thus users can formulate their queries and provide their responses using the oral (linguistic) modality and the tactile modality separately or together. This paper presents responses to some of the questions above. First we describe the system the goal of which is the querying of a geographic and tourist database. Then we analyse the referential phenomena both from the linguistic and the tactile point of view. This analysis is originated from an experimental (WOZ) work [11] and the results have been confirmed by way of a first evaluation of the system with naive users. The different types of problems are underlined. The paper continues by presenting the general principles which guide us and the description of the principal processing methods. The principles are concerned with the choice of the modality on which the resolution of references is based, the architecture of the steps and the type of modeling. In the course of the description of the processing methods, we provide details about the format and the content of the data used. In the conclusion, we outline future planned studies using the system. More details about the preliminary experiments, the architecture and the system evaluation can be found in [11, 18, 19].

System description

GEORAL Tactile is a multimodal system which is able to provide information of a touristic nature to naive users. Users can ask for information about the location of places of interest (city, beach, château, church,...) within a region or a subregion, or distance and itinerary between two localities. Users interact with the system using three different modalities: in a visual mode by looking to the map displayed on the screen, in an oral and natural language mode (thanks to a speech recognition board[1]) and in a gesture mode pointing to or drawing on a touch screen. The system itself uses both the oral channel (text-to-speech synthesis) and graphics

[1]The speech input is processed by the recognition board MEDIA50 (licenced by France Telecom CNET).

such as the flashing of sites, routes and zooming in on subsections of the map, so as to best inform the user.

The dialogue model is an adaptation of the LOKI model [22]. It allows to build up a structured dialogue history based on the theme of the queries. Some dialogue functionalities (spelling, repeating, writing) are added in order to take into account specific features of the oral mode. The model also contains co-operative algorithms [9, 10] which avoid producing empty responses and allow to manage the interaction in a directed but friendly manner.

Referential phenomena

In such a context, we implemented a WOZ experiment [11] in order to study how users designate the referents. The main outcomes of this study have been confirmed during a first evaluation of the system [19]. We examine here primarily the deictic problems concerning both the tactile and linguistic users' activities. Most standard anaphora are dealt with in the system (using lexical, syntactic and semantic information). This, however, is not discussed here.

What are the referents?

In this application, the possible referents are sites[2] and localities which are recorded in the database and which may appear on the displayed map. From a user point of view, and depending on the state of the dialogue, a possible referent may be displayed on the screen (for example at the beginning of the dialogue, the main localities are presented on the map; after a suggestive response of the system, a few interesting localities are presented on the map with a flashing effect possibly after zooming in) or may only be evoked when, for example a user would like to know if a certain site not displayed on the screen exists in a zone of the map.

How do users designate?

Various types of tactile activities Two main types of tactile activities are observed:

- the pointing mode: the user points out a point on the screen which may correspond to a referent (site, locality) or to a place for which there is no referent.
- the zoning mode: the user drawns up a zone in which the user want to do a search for referents. The drawing of the zone may be quite complex: it can use some elements of the map (coast, river,...), the surface can be open or closed... In the current implementation, only the closed zones are dealt with.

[2]In our system, site is a place of interest (eg. a church), locality is a place name and zone is a region delimited by the system or the user.

Tactile and Linguistic joint activities The presence of the tactile screen modifies the linguistic behaviour of the user: some particular deictic terms (around 10 words and expressions are dealt with) and particular syntactic structures appear. Three possible relationships between oral utterances and tactile gestures have been identified (the two main ones follow):

- bound relationship in which one deictic item appearing in the oral utterance and a touch activity are used together to designate a referent (or a set of referents) on the map[3]:

  (1) U: Are there any beaches in this locality? + a touch on a locality.
  (2) U: Are there any châteaux here? + a touch on a locality.
  (3) U: Are there any churches in this zone? + drawing a zone on the map.
  (4) U: Here + a touch on a locality.
  (5) U: This one + a touch on a site.

- confirmative relationship for which the oral syntagm is sufficient enough to comprehend but is however accompanied by a tactile designation, which is redundant with the linguistic reference:

  (6) U: Are there any beaches in Lannion? + a touch on Lannion.
  (7) U: in Lannion + a touch on Lannion.

Some difficulties

These designation modes seem to be very straightforward and easy to deal with but some problems which depend on the user behaviour arise and make processing more complicated.
The relationship between the deictic term and the tactile designation is not systematic. For example, one can find an utterance (2) occuring together with a drawing of a zone (in French pointing would better correspond to the item *here*), or second example, utterance (3) can also merely come with a pointing which is not suggested by the syntagm *in this zone*. From another point of view, the user tactile activities may be imprecise, for example the pointing together with utterance (1) could designate a place where there is no locality. These problems suggest to design mechanisms which will have to deal with these imprecisions in a static and dynamic way.

Experiments have also displayed differences between users concerning the use of the tactile screen. The degree of use of the tactile is highly dependant of the subjects: on average 36% of the initial requests are composed of a tactile activity but the highest rate for any

[3]These examples are litteral translation of the examples actually used in the system.

given user is around 95% and the lowest 2%. These facts mean that oral modality has to keep a dominating role as far the reference processing is concerned.

Processing the references

Main principles

Some main principles can characterize our approach of the problem; they reflect two major concerns: the first one is the robustness in order to deal with the imprecision and the second one is the flexibility in order to tune out the behaviour of the system and to make easier its evolution. These principles are as follows:

- The global processing is divided up in three sequential steps: a linguistic analysis, a tactile analysis and a merging process.
- The processing is mainly based on the oral modality, i.e. on the analysis (syntactic and thematic) of the oral utterance. This fact not only allows us to take into account the different uses of the system (with or without the tactile screen), but also compels us to do so because the tactile mode does not provide sufficient information in most of the cases. This choice presents some consequences and drawbacks: the speech recognition becomes very important and the discrepancies between oral and tactile activities will be dealt with by the tactile analysis.
- The algorithms are based on contextual information. The dialogue history (context) provides the necessary elements (referents) to progressively solve the references, and the oral utterance (co-text) provides predictions about foreseen referents and type of tactile activity. Nevertheless, we chose an approach where the results of the linguistic and tactile activities are merged instead of an approach where the activities are interpreted using the tactile context [20].
- Finally, the modelling of the different activities is based on planning operators which allow us to easily build up communicative acts that represent the joint activities of the user.

Main Processes

Linguistic Analysis This analysis is made up of two modules: the syntactical and the thematic analysis which are triggered after the speech recognition. The syntactic analysis produces a complete syntactic tree using the difference list method [6]. The deictic and anaphoric syntagms are only spotted in the utterance and coded inside the tree. The thematic analysis has two principal roles as regards the tactile function. It determines the possible types (style) of tactile touch (point, zone) as well as the theme of the question (type of object in question). It also produces an intermediary structure, a so-called dialogue act, of which the modeling of propositional content is inspired by [1]. For example, the user's utterance:

are there any beaches here ?

will be transformed as :

ASK(U, S, informref(S, U, beach, Q(beach, locality(deicphore(pointing)))))

where *deicphore(pointing)* is generated by the presence of the keyword *here* and indicates a user tactile activity to point out the place where the system will have to explore. The transmission of the theme to the tactile processor is accompanied by the relevant objects of the database.

Processing the tactile activities The aim of this processing is, starting from the elementary events which correspond to tactile activity of the user, to provide the possible objects of the database which correspond to the designated or desired referents. The process is based on a prediction about the style transmitted by the thematic analyser as well as on objects (potential referents) pre-selected from the database. It produces a possible empty list (tactile acts) of designated objects.

At this level the robustness problems are two-fold: (1) the gestural activity may consist of several gestures (drawing multiple zones, touching multiple locations), (2) the gestural activity may not be consistent with the linguistic activity. As far as the first type is concerned, we only take into account the two final points or the last drawn zone. As for the second type, we have designed a small set of rules (easily modifiable), which allows us to modify the gestural activity observed according to the expected style. Up to now, this modification does not take into account neither the geographical context nor the dialogic one.

Producing Communicative Acts A communicative act (CA) expresses the user's intention, which in turn is simultaneously conveyed by the speech activity (the dialogue act) and the tactile activity. In order to determine a CA one must merge the two types of activity, whilst using to a maximum advantage the redundant information conveyed by the two media [4, 20]. This merging must also take into account possible incoherence problems between media, as a result either of the system (speech recognition and understanding) or of the user. According to the dialogue context, (recognized type of dialogue act) the merging is carried out in two ways. A set of rules, taking into account the different situations on the media, is used for dealing with phatic dialogue acts, as it is for a closed question. They enable a decision to be taken. The following rule (simplified):

identification(L2):-speech (L1),tactile(L2).

specifies that the location designated by the user will be L2 because it was designated in a tactile fashion despite the fact that speech recognition provided L1. The expertise currently coded in the rules is the outcome of the WOZ corpus; it has a tendency to favor tactile

NAME:	Request_completive mode
HEADER:	REQUEST(U, S, INFORMEREF(S,U, ?x, Q(?x, ?P'(?o'))))
BODY:	ASK(U,S, informerf(S,U, ?x, Q(?x, ?P(?D))))
	Désigner(U, S, ?o)
CONSTRAINT:	
PRECONDITION:	Déictique(?P(?D)), ?o, ?P'(?o'))

Figure 1: A model for the Request Communicative Act

media. The specification by means of rules will make future modifications easier.

The second mode of merging concerns the dialogue acts labelled as requests. It is based on a representation of CA as plan operators [5, 15, 16, 21]. Dialogue acts as well as the tactile acts are considered as low level plan operators. Whilst merging, we solve tactile and linguistic co-references [3, 8].

For example, Figure 1 shows a model of the CA REQUEST for the completive mode where a tactile event and a dialogue act have to be merged. This case is specified by the presence of the tactile act "Désigner" in the Body part of the model.

The predicate Déictique in the precondition part checks the consistency beetween the deictic term (?P(?D)) within the dialogue act ASK and the referent (?o) provided by the tactile act. It produces the referent (?P'(?o')) to be placed in the communicative act. For example, the deictic term locality (deicphore(pointing)) in the dialogue act:
ASK(U, S, informref(S, U, beach, Q(beach, locality(deicphore(pointing)))))
and the referent (Lannion, pointing, 1, X, Y) in the tactile act
Désigner(U, S, (Lannion, pointing, 1, X, Y))
will be recognized as compatible and allow to produce the referent locality(Lannion) for the CA REQUEST:
Request(U, S, informref(S, U, beach, Q(beach, locality(Lannion))))
Further details are provided in [17]. We can also observe the considerable flexibility brought about by the use of rules in the verification of preconditions.

Discussion and Future Plans

We described the methods and models we designed in order to take into account deictic references within a specific framework. These solutions are mainly based on contextual information. They allow to deal with some robustness problems and to provide a certain flexibility for specifying the operations in order to take into account future new semantic or pragmatic information. Clearly, the solutions we have presented in this framework are strongly dependent on the kind of application targeted. For example, an application in which users handle objects using the system presents other characteristics and so will need other solutions.
We plan to extend this work in several directions:

- to take into account other styles of designations and tactile activities. For example, to allow activities such as:
  U: Are there any beaches along this coast? + U follows with his finger a coast
  or
  U: Are there any châteaux above this river? + U touches on a river.
  These possibilities need to add a lot of new item to the speech recognition vocabulary but in addition require new knowledge about the cartographic context. It will be also necessary to pay closer attention to user behaviour and perhaps even contemplate user modelling.
- to allow the user to ask about typical features of certain referents (for example: opening hours of a site). In this case, clearly the nature of the referential phenomena will change and will need some more sophisticated processing [14].

Acknowledgements. This project was partially funded by CNET (France Telecom), contract 92 7B.

References

[1] Allen J. Natural Language Understanding. The Benjamin/Cummings Publishing Company, Inc., 1987.

[2] Proceedings of CMC95 International Conference on Cooperative Multimodal Communication, H. Bunt, R.-J. Beun and T. Borghuis (eds), Eindhoven, The Netherlands, 24-26 May 1995.

[3] Cohen, P. R. The Pragmatics of Referring and the Modality of Communication. *Computational Linguistics*, Vol. 10, Num. 2, April-June 1984, p. 97-146.

[4] Cohen, P. R. The role of Natural Language in a Multimodal Interface. *Proceedings of 2nd FRIEND21, International symposium on next generation human interface technology*, Tokyo, Japan, Nov. 1991.

[5] Feiner S.K. and McKeown K.R. Coordinating Text and Graphics in Explanation Generation. *Proceedings of the AAAI-90*, July 30-August 3, 1990.

[6] Gal A, Lapalme G. and Saint-Dizier P. Prolog pour l'analyse automatique du langage naturel. Éditions Eyrolles, Paris, 1988.

[7] Gavignet F., Guyomard M., Siroux J. Implementing an oral and geographic multimodal application: the Géoral project. *Pre-proceedings of the Second Venaco Workshop on the Structure of Multimodal Dialogue, NATO*, Acquafredda di Maratea, Italy, September, 16-20, 1991.

[8] Green G. M. Pragmatics and Natural Language Understanding. Lawrence Erlbaum associates, 1987.

[9] M. Guyomard and J. Siroux. Suggestive and Corrective Answers : a Single Mechanism, in *The structure of multimodal dialogue*, Taylor M. M., Néel F. and Bouwhuis D.G. Éditeurs, North Holland, 1989, p. 361-374. (Workshop NATO The structure of multimodal dialogue, Vénaco, 1986.)

[10] Guyomard M., Siroux J. and Cozannet A. The Role of Dialogue in Speech Recognition. The Case of The Yellow Pages System. *Proceedings EUROSPEECH 91*, Genova, Italy, 1991, p. 1051-1054.

[11] Guyomard M., Le Meur D., Poignonnec S. and Siroux J. Experimental work for the dual usage of voice and touch screen for a cartographic application. *Proceedings of ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, may 30-June 2, Vigsø, Denmark, 1995.

[12] Proceedings of First International Workshop On Intelligence and Multimodality in Multimedia Interfaces: Research and Applications. J. Lee (ed.), Edinburg, Scotland, UK, 13-14 July 1995.

[13] Litman D. J., Allen J. A Plan Recognition Model for Subdialogue in Conversations. *Cognitive Science* 11, p. 163-200, 1987.

[14] Mathieu F.-A. Prise en compte de contraintes pragmatiques pour guider un système de reconnaissance de la parole: le système COMPTA. *PHD thesis*, université Henry Poincaré, Nancy, Mars 1997.

[15] Maybury M.T. Planning Multimedia Explanations Using Communicative Acts. *Proceedings of the Ninth National Conference on Artificial Intelligence, AAAI 91*, Anaheim, CA, July, 14-19, 1991.

[16] Maybury M.T. Communicative Acts for Explanation Generation. *International Journal of Man-Machine Studies*, 37(2), 135-172.

[17] Multon F. GEORAL tactile un système multimodal. Rapport de DEA, IFSIC, université de Rennes 1, 1994.

[18] Siroux J., Guyomard M., Multon F. and Rémondeau C. Modeling and Processing of the Oral and Tactile Activities in the GÉORAL Tactile System. *Proceedings of CMC95 International Conference on Cooperative Multimodal Communication*, Eindhoven, The Netherlands, 24-26 May 1995.

[19] Siroux J., Guyomard M., Multon F. and Rémondeau C. Speech and Tactile Based GEORAL System. *Proceedings of EUROSPEECH95*, 18-21 September, Madrid, 1995.

[20] Tyler S.W., Schlossberg J.L. and Cook L.K. CHORIS : An Intelligent Interface Architecture for Multimodal Interaction. *Proceedings of the AAAI91 Workshop on Intelligent Multimedia Interfaces*, Anaheim, CA, July, 14-19, 1991.

[21] Wahlster W., André E., Graf W. and Rist T. Designing Illustrated Texts : How Language Production is influenced by Graphics Generation. *Proceedings of EACL 91*, Berlin, April 1991.

[22] Wachtel T. Discourse structure -LOKI.*NL1-1.1*, Research Unit for Information Science and Artificial Intelligence, University of Hamburg, 1985.