Knowledge and Data Flow Architecture for Reference Processing in Multimodal Dialog Systems

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

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

ABSTRACT

This paper is concerned with the part of the system dedicated to the processing of the user's designation activities for multimodal search of information. We highlight the necessity of using specific knowledge for multimodal input processing. We propose and describe knowledge modeling as well as the associated processing architecture. Knowledge modeling is concerned with the natural language and the visual context; it is adapted to the kind of application and allows several types of filtering of the inputs. Part of this knowledge is dynamically updated to take into account the interaction history. In the proposed architecture, each input modality is processed first by using the modeled knowledge, producing intermediate structures. Next a fusion of these structures allows the determination of the referent aimed at by using dynamic knowledge. The steps of this last process take into account the possible combinations of modalities as well as the clues carried by each modality (linguistic clues, gesture type). The development of this part of our system is mainly complete and tested.

Categories and Subject Descriptors

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

General Terms

Algorithms, Languages

Keywords

Multimodal human-computer communication, Multimodal fusion, reference, natural language, gesture

1. INTRODUCTION

We are concerned with inputs of a multimodal dialog system for geographical information searches. During a communication turn, the user performs an oral utterance and/or a gesture on the touch screen taking into account her goal, the state of the interaction and her perception of the map displayed on the screen.

These activities convey a communicative act (CA) whose components include the user's intention and references to necessary objects for the search. The purpose of this paper is to show how we process the multimodal inputs from the user in order to identify the referents aimed at. Obviously this processing has to be based on both knowledge and algorithms in order to deal with oral utterances (syntax, semantics, pragmatics) as well as the gestures and the user visual perception. Moreover, it is also necessary to take into account some errors during the interaction due to system components (i.e. speech recognition errors) [5] as well as from the user (performance problems).

In the following sections, we describe more closely the problem addressed. We illustrate the different cases of designation activities, the necessary knowledge and the difficulties of multimodal reference resolution. Next, we show an overview of the proposed architecture.We then detail the fusion and resolution processes to resolve referential expressions whether or not accompanied by gestures. Finally, we propose a knowledge modeling concerned with the natural language and the visual context.

2. CURRENT SYSTEM

The framework we use to analyze and implement our approach is the Georal tactile system [6]. It is a multimodal system principally used to provide information of a touristic and geographical nature. Users can ask for information about the location of sites of interest (beach, campsite, château, etc.) by specifying a place, a zone (a particular geographical object: river, road, city, etc.) (figure 1). Georal offers the user the following modes and modalities: a) Oral input to as well as oral output from the system. Users can formulate their requests and responses to the system by voice and in natural language (NL) in a spontaneous manner (no particular instructions of elocution). Some system outputs use speech synthesis. b) The visual mode by displaying a map of a region on the screen; this map contains a representation of the usual geographical and touristic information: cities, roads, rivers, etc. c) The gesture mode by the intermediary of a touch screen: the user can designate objects displayed on the screen by various types of gesture (point, zone, line, etc.).

3. MULTIMODAL REFERENCES

Our principal aim in this paper is to refine the processing of the user's designation activities. The user has to designate the place in question by her information search. This

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

designation is performed by means of NL (referential expression *RE*) and/or gesture based on the displayed objects on the map. The system's role is to determine the place aimed at (referent) by analyzing and interpreting the contents of the user's activities.

The following examples illustrate different cases of designation activities, the necessary knowledge and the difficulties of multimodal reference resolution.

Example 1 *give me hotels here* ; accompanied by a pointing gesture to a city displayed on the map.

Example 2 *give me hotels along this line* ; accompanied by the gesture outlined in bold black on figure 1.

Example 3 *give me hotels along this river* ; accompanied by the gesture outlined in bold black on figure 1.

Example 4 *give me hotels in the suburbs of this city* ; without gesture.

Example 5 *give me hotels in the surroundings of the city of Pleumeur-Bodou* ; without gesture.

Figure 1: Example of an ambiguous gesture (a small part of the map used by the Georal system)

The processing of the deictic *RE here* of example 1 is reduced to determining only the object designated by the accompanying gesture. Hence, the success of this process depends on whether or not the gesture is ambiguous. If the gesture is ambiguous, information about the objects and their visual impact should be used to choose the most likely object to the user.

The processing of the deictic *RE this line* of example 2 consists in determining the object (of line category) designated by the gesture. Methods which process this *RE* and the joint gesture should access, on the one hand, linguistic and applicative knowledge to identify that a line could be a road, river, etc and, on the other hand, knowledge about the displayed objects to deal only with those of line (or polyline) category. In this example, the accuracy of the gesture is poor: is the targeted object the road (the polyline labeled D788) or the river (the polyline labeled *ruisseau de Gruguil*)? In fact both "candidates" (the road and the river) could refer to the *RE this line*. Gesture disambiguation should take into account additional information about the objects (like the case of ambiguity in example 1) to choose the most likely to the user.

In example 3, the gesture is ambiguous and it could be solved by accessing the NL information about the designated object (the word *river* in the user's utterance).

The *RE this city* in example 4 (produced without gesture) is an anaphoric expression [4]. This RE, for example, could be produced after mentioning a city in a previous communication turn. In this case, the linguistic antecedent representing the referent of *this city* may be the last named city in the history. The referent aimed at by the user is made up of only the *suburbs* of the designated city. This NL information (the word *suburbs*) should be taken into account to satisfy the user's goal by using applicative knowledge.

Example 5 concerns a NL utterance (without gesture) in which the *RE the surroundings of the city of Pleumeur-Bodou* is a complex definite description (DD) used in first mention [7]. This *RE* designates a part of a city displayed on the map. The processing of this *RE* should take into account linguistic and applicative knowledge to identify that the word *surroundings* signifies the *suburbs* of the named city.

These few examples clearly show that the processing of the designation activity should take into account resources depending on the modality like visual perception, semantic relationships, and applicative knowledge.

In this paper we suggest a knowledge modeling (cf. section 5) concerned with natural language and the visual context adapted to the kind of application we are dealing with. This knowledge allows us to improve complex *RE* and gesture processing. We also propose an architecture for input processing (cf. section 4.1) which is based on the proposed knowledge modeling.

4. DATA FLOW PROCESSING

4.1 Architecture Overview

In our architecture (figure 2), we deal with synchronous inputs (a prompt is displayed to allow the user to produce her inputs). Each kind of input is processed separately using the modeled knowledge to produce intermediate structures. The speech recognition grammar used (cf. section 5.1) incorporates syntactic, semantic, pragmatic and applicative knowledge.

The intermediate structures are, on the one hand, the syntactic and semantic representations of the recognized speech input and, on the other hand, an improved list of points representing the gesture as well as its type (point, line, etc). The semantic representation is produced using linguistic and applicative knowledge (cf. section 5). This semantic representation is used during the fusion and resolution processes (cf. section 4.2) and the completion of the CA. At this stage, histories (cf. section 5) are updated to include the current gesture and NL inputs and their intermediate structures.

Next, using dynamic knowledge (histories and the current visual context (CVC)), a fusion of the intermediate structures and the triggering of resolution methods (cf. section 4.2) allow the determination of the referent aimed at. The referent thus found is sent to the dialog manager which completes the CA, by using the semantic representation of the NL input. The dialog manager sends a request to the database to search for the information aimed at by the user (touristic information about each object and place are stored in the database (DB)). The CA of the example 5 is $Request(hotels, suburbs(Pleumeur - Bodou))$.

If the reference resolution process fails, the dialog manager builds up clarification messages for the user depending on the reason for this failure (remaining ambiguity, speech recognition failure, semantic analysis failure).

4.2 Fusion and Resolution Processes

The steps of these processes (figure 3) take into account the possible combinations of modalities as well as clues belonging to each modality (linguistic clues, gesture type).

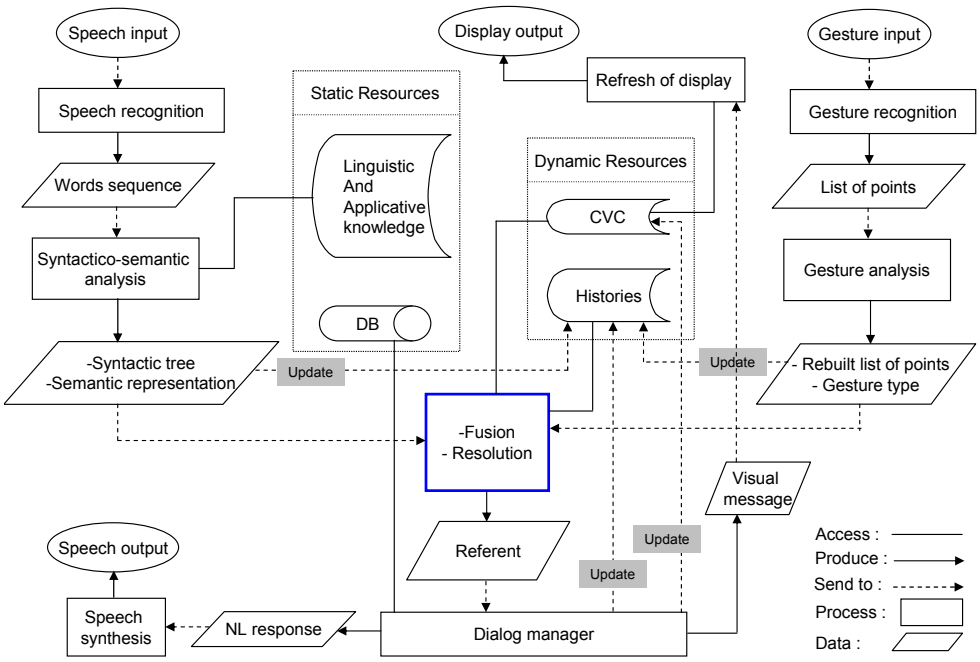

Figure 2: Data flow processing architecture

Firstly, if there is a gesture in input, gesture processing algorithms [1] are applied to find the designated referent object. These algorithms use, in addition to the gesture type, the semantic representation, if available, of the user's utterance (the semantic representation is not available if the semantic analysis fails or if the user produced a gesture alone (this is possible in some specific dialog situations)).

In the case of an input without gesture, we determine the *RE* type. Depending on this type, an appropriate processing method is applied to find the referent. Hence, if the *RE* type is a DD employed in first mention (the example 5), the object referent is found by accessing modeled CVC. The determination of the referent in this case consists of matching the NL description (available in the *RE*) with the objects' characteristics in the CVC to find the most relevant and likely to the user. In the case of anaphoric *RE*, the referent is found in the modeled histories [4].

For all cases of failure a specific code is produced and sent to the dialog manager to ask for clarification or to indicate failure to the user (cf. section 4.1).

5. KNOWLEDGE MODELING

We have shown above the necessity of using knowledge during different *RE* processing steps. Knowledge mainly concerns the natural language and the visual context modalities and allows the system to deal with inputs including complex NL utterances and imprecise gestures. For example, knowledge about the displayed objects allows the system to process metaphoric designations to these objects and to disambiguate gestures.

In addition to the NL and visual context modeling detailed below, we model histories, which represent the multimodal dialog, in a centralized, structured and synchronized manner. 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. The structure follows the Georal dialog grammar [2].

5.1 Natural Language Modeling

We propose a modeling which allows the incorporation of lexical, syntactic, semantic, pragmatic, and applicative aspects for the processing of the NL utterances. This modeling is based on a context free grammar which incorporates lexical, semantic, pragmatic and applicative rules. These rules are organized by taking into account the existing relationships between linguistic expressions associated to *RE*s, the object representations on the screen and the real world. In fact, we concluded after an experimental study that linguistic expressions have to be classified following three viewpoints: the first one concerns linguistic expressions which apply to any object, for example, the linguistic expressions *near to, close to* can precede the designation of any displayed object such as the REs *near to this object; near to this line; near to this point; near to this river; near to this city;* etc. The second one concerns the linguistic expressions which apply to a kind of geometric object, for example, *along, inside* can precede objects of line and zone geometric form respectively such as the REs *along this line; along this river; along this road; inside this zone;* etc. The third viewpoint concerns linguistic expressions which apply to applicative objects, for example, *in the mouth of, in the downtown of* can precede objects of river and city type respectively such as the REs *in the mouth of this river; in the downtown of this city;* etc. Our modeling is based on the corpus analysis we have collected during an experimental study which aimed at observing how users phrase their requests in front of the displayed map.

We have developed the grammar while given heed to the genericity of the kind of application we are dealing with. Obviously, a part of the rules is application dependent. In geographical information search systems, objects from the real world which correspond to forests, rivers (roads and coasts) and cities are represented on the map by zone, line (or polyline) and point geometric forms.

We propose a taxonomy of these objects based on three

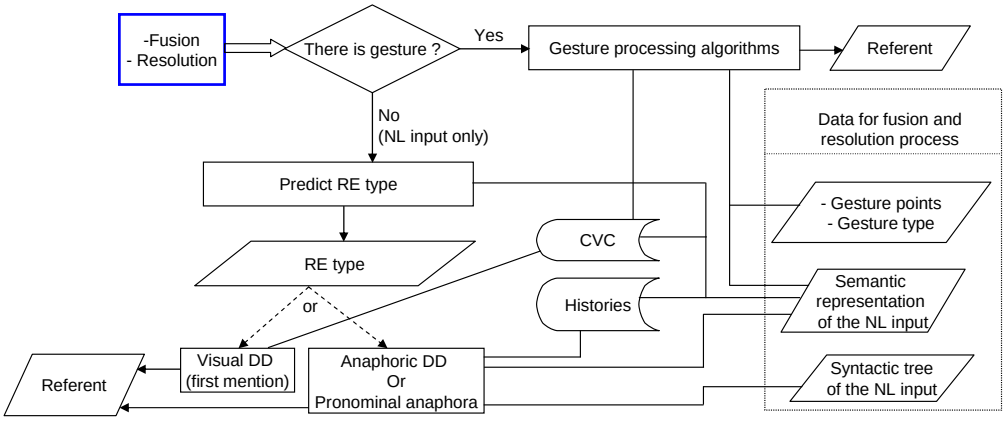

Figure 3: Fusion and resolution processes steps overview

hierarchical levels (noted L_1, L_2 and L_3): L_1={object}, L_2={zone, line (polyline), point}, L_3={forest, river, road, coast, city}. So we obtain nine nodes in the hierarchy (object, zone, line, point, forest, river, road, coast and city). We make relationships, on different hierarchical levels, between nodes and linguistic expression sets according to their classification of use. Grammar rules for analysis are obtained by combining nodes and sets. Hence, this grammar could be used for other applications by updating (adding or deleting nodes) the hierarchy.

The grammar developed contains semantic filters by its construction. Semantic filters consist of verifying the coherence within the NL input between linguistic expressions and the object references, for example, the grammar does not accept utterances like *in this river* or *at the mouth of this road*. We added to the grammar semantic and applicative interpretation rules to compute the semantic representation, for example, the semantic representation of the example 3 is $Request(hotel, along(this, river))$. This interpretation step serves as an early reference processing because it translates some words depending on their meaning in the application. This interpretation facilitates the matching between RE and the visual objects. For example, the semantic interpretation of the word *surroundings* is *suburbs* if it precedes the word *city*. We have developed this grammar using the SRGS [8] and the SISR [9] specifications.

5.2 Visual Context Modeling

We call the "common visual context" the map displayed on the screen. The modeling of this display is important to understand the user inputs. It consists of the internal representation of the displayed objects. We associate a characteristics vector to each displayed object. A vector contains the name displayed on the screen (it can be different to the database name), color, form, size, coordinates, and salience. The salience [3] of an object consists of its visual and contextual weight. To determine these weights, we use a salience distribution algorithm on objects in the common visual context [1]. Salience is used in the case of designation gesture ambiguity. We have modeled the CVC in XML.

6. CONCLUSION

We have proposed and described a knowledge modeling as well as the associated architecture for multimodal input processing and fusion for the kind of geographical search information application. The aim of the fusion is to determine the object designated by the user (referent). The knowledge modeling is mainly concerned with the natural language and the visual context. The natural language modeling is based on a corpus analysis. This modeling is generic and incorporate semantic and pragmatic considerations. The visual context modeling take into account objects' characteristics (salience, ...). Part of the knowledge is dynamically updated to take into account the interaction history. In the proposed architecture, each input modality is processed separately by using the modeled knowledge, producing intermediate structures. Next, by using dynamic knowledge, a fusion of these structures allows the determination of the referent aimed at. The steps of this process take into account the possible combinations of modalities as well as the clues carried by each modality (linguistic clues, gesture type). The proposed architecture and the modeling are implemented in our system.

7. REFERENCES

[1] A. Choumane and J. Siroux. Interpretation of multimodal designation with imprecise gesture. In *IE07*, pages 232–238, Germany, 2007.

[2] A. Choumane and J. Siroux. A model for multimodal representation and processing for reference resolution. In *WMISI '07*, pages 39–42. ACM, 2007.

[3] F. Landragin. Referring to objects with spoken and haptic modalities. In *ICMI '02*, page 99. IEEE Computer Society, 2002.

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

[5] S. Qu and J. Y. Chai. Salience modeling based on non-verbal modalities for spoken language understanding. In *ICMI '06*, pages 193–200. ACM, 2006.

[6] J. Siroux, M. Guyomard, F. Multon, and C. Rémondeau. Multimodal references in georal tactile. In *Workshop of ACL Meeting*, Spain, 1997.

[7] R. Vieira and M. Poesio. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26 (4):539–593, 2000.

[8] W3C. *Speech Recognition Grammar Specification*. http://www.w3.org/TR/speech-grammar/, March 2004.

[9] W3C. *Semantic Interpretation for Speech Recognition*. http://www.w3.org/TR/semantic-interpretation/, April 2007.