A Multimodal Dialogue Contribution Coding Scheme

Luis Villaseñor, Antonio Massé,
Luis A. Pineda

Department of Computer Science
Institute for Applied Mathematics and Systems (IIMAS)
National Autonomous University of México (UNAM),
AP. 20-726 Admon. No. 20.
Del. Alvaro Obregón 01000,
México, D.F.
{villasen, masse, luis}@servidor.unam.mx

Abstract

In this paper a coding scheme for annotating speech acts in the context of multimodal design tasks is presented. The scheme is an extension of the dialogue act markup in several layers scheme *DAMSL* (Allen and Core, 1997). This scheme has been used to annotate human-human task oriented conversations; however, information other than spoken language, like deictic gestures or information conveyed through external representations, such as paper or graphical screen, cannot be taken into account with DAMSL and related coding schemes. The extension proposed provides a methodology to capture deictic and graphical information common in design task oriented dialogues. The scheme is been developed in the context of the *Diálogo Inteligente Multimodal en Español* program (*DIME*) to support Spanish spoken human-computer conversations in which the system has the role of a design assistant in the kitchen design domain.

1. Introduction

One long-term goal of computational linguistics and artificial intelligence is the construction of natural language conversational systems with spoken input and output facilities. Although the goal to model human conversation with all its richness and flexibility is still far away to reach, current computational technology and hardware capabilities make feasible the construction of prototype systems capable to sustain a goal-oriented conversation in specific domains. Instances of this kind of systems are *TRAINS* and *TRIPS* (Allen et al., 97). In these systems, human-users are able to engage in a natural language and graphics interactive session with the purpose of solving specific planning or designs problems. In the case of *TRAINS*, the system helps to schedule trains work orders, and in *TRIPS* the system is an assistant to evacuate an island in a situation of emergency. Following on the lines of these two systems and in order to test whether this kind of technology can be applied to different application domains and languages, we are currently developing the program *DIME* (*Diálogos Inteligentes Multimodales en Español*).

There are several considerations that have to be taken into account for the construction of conversational working prototypes. First, the application domain should be complex enough to merit the use of a natural language assistant but, at the same time, as simple as possible to be able to model the task with current computational technology. Also, in order to handle natural language ambiguity and limit the very large amount of general knowledge that might be involved in even trivial human conversations, current conversational systems must be restricted in several dimensions. In particular, the application domain and potential goals to be satisfied through a conversation must be defined in advances and as precisely as possible. For the purpose of *DIME*, the domain is kitchen design. It is a simple task that most people can undertake without previous experience and yet the assistance of an expert can help to notice and enforce a number of design constraints to improve the functionality, look and value of kitchens. From this discussion it should be intuitive enough that a central concern for the design of conversational systems is to rely on empirical observations about both the nature of the task (e.g., beliefs and intentions of the conversational participants) and the kind of language employed through goal-oriented conversations in the application domain.

Currently, the present scheme is used to annotate a corpus consisting of about 30 dialogues in the kitchen design domain with promising results. These dialogues were collected through a Wizard of Oz experiments. Particularly, our experimental setting is aimed to obtain dialogues in Mexican Spanish where the system and a user collaborate to design a kitchen. In such dialogues either participant can refer to objects through a graphical user interface and reason about them and its geometric form or functional relations. Each session consisted of an explanation of the system to the subject, a demonstration of the system, and the solution of two tasks through a goal-oriented conversation. The first task was very simple and had the purpose to familiarize the user with the experimental setting; the second consisted in solving more complex design problem. The dialogues discussed below in this paper are the result of these experiments; there were 15 experiments run with 15 different persons aged 30 on average, most of them were computer science related students.

2. A Dialog Annotation Scheme

The main purpose of the task analysis is to identify the intentions underlying the human-user expressions. Once

the system is able to identify a specific intention it should be able to identify a specific problem to be solved or a goal to be achieved and to produce a response and actively engage in the conversational cycle. The task analysis, based on the study of the corpus, permits to identify the family of goals that are normally pursued in the design domain, and also the conversational strategies used by human-users to achieve such goals. To characterize this kind of information we employ the *DAMSL* annotation scheme (Allen and Core, 1997). Here, we note that the basic unit in *DAMSL* is the utterance; however, expressing and satisfying intentions in the course of conversations can normally be achieved through a number of conversational turns. To capture this level of aggregation we propose to extend *DAMSL* with a more structured conversational unit which, following (Clark and Schaefer, 89), we call *contribution*.

A second concern of this investigation is the multimodal character of the dialogs under study. Information in the kitchen design scenario is conveyed not only through natural language but also through graphics and demonstration gestures. We can distinguish two kinds of multimodal aspects during these dialogues: the one related to permit and facilitate communication, like explaining or helping the other participant to recognize a particular object, and those related to the modification of graphical scenarios. Graphical actions modify the cognitive state of both participants and thus they must be considered as part of the dialogue.

The question of how multimodal aspects of dialogs permit or facilitate communication is not that simple. Here, we explore the hypothesis that multimodal information facilitates the process of establishing reference. The interpretation of pronouns, descriptions or even a proper name are complex inference processes; however, if a non-linguistic context is available, reference can be established often by correlating linguistic terms with objects in the context directly. According to this, an important consideration for a multimodal annotation scheme is to record whether terms occurring in multimodal dialogues are understood in relation to a linguistic context, built out of discourse information, or they rather receive a direct interpretation from graphical or other kinds of context. In our view, information provided by non-linguistic modalities is placed in a look-up table that can be accessed directly for resolution purposes. We call this table *the non-linguistic context* or simply *the graphical context*. In the same way that exceptions can be considered before the application of inference rules in many knowledge domains, like in the interpretation of irregular verb forms, the context can be thought of as a cache memory in which information provided by perception or memory is readily available for interpretation. The resolution process would look into this table before general inference rules, say for the resolution of anaphors, are employed. The simplest process of this kind is ostensive demonstration. In addition, in order to implement this strategy it is assumed that the indexical resolution is a process of constraint satisfaction through which linguistic and contextual information can be related very efficiently (Pineda and Garza, 2000). Our extensions to the *DAMSL* annotation scheme are designed to capture these intuitions.

3. Multimodal Annotation Scheme

The extension to *DAMSL* has been done in two dimensions. First we extended the number of labels to count with a mechanism to label graphical actions. Second we structured multimodal cooperative dialogues as sequence of contributions. Table 1 shows a fragment of a *DIME* dialogue where this extension is shown. It has entries for contribution labels, utterance identifiers, annotation labels, Spanish interventions with their corresponding English translations and, finally their corresponding referents. Deictic expressions within interventions are highlighted and the corresponding pointing gestures are labeled as demonstrative events. These are identified as "*evX.Y.Dd*", where X is the utterance in which the demonstration is made, Y is the number of demonstration in the utterance, and D indicates whether the gesture has been made on the 2-D or 3-D window of the multimodal interface. Intuitively, the referent of a highlighted term on the Interventions column is an individual that can be identified in the region pointed out in the graphical domain, taking into account the conceptual constraints imposed by the linguistic term. On the Referents column for every spatial demonstration a referent is explicitly stated. Here, we would like to suggest that the main contribution of multimodality to the communication process is that indexical references are available directly and no complex inferences are involved for the resolution of these terms. We also note that for the construction of a multimodal conversational system an algorithm for correlating linguistic terms with the graphical referents in the graphical context in a very effective way must be available. An algorithm for such a purpose is described in (Pineda and Garza, 2000).

Many of the references in *utt11* to *utt18*, in Table 1, are supported by overt demonstrations, which provide directly referents from the graphical domain. It can also be noticed that several demonstrations can occur within a single utterance. To capture this information we include in the annotation scheme a representation of the context as shown in Table 2, where the information that can be accessed by visual perception and sets the context for the dialogue is specified. Spatial demonstrative expressions refer always to these objects. Whenever a new object is introduced during the dialog, a new entry is added to the context by entering the demonstrative event identifier and the object. Table 2 shows the corresponding context to the dialogue fragment shown in table 1.

The purpose of this dialog fragment is to reach an agreement about an action performance. During the presentation phase (utt11-15) a graphical action and its corresponding referents are explicitly identified. During the acceptance phase (utt16-18) the graphical action is performed and the new graphical state is evaluated. Note that in order to achieve successfully this action all graphical referents involved, as well as their desired graphical properties, must be unambiguously determined.

Contribution-Task Level	Contribution-Description Level		Utt Id	DAMSL labels	Interventions	Translations	Referents
Presentation = placement	Presentation = action-specification		u: utt11	Task / Influence-on-listener = Action-directive / ∅	Puedes poner este <sil> **este estante** (**ev11.1.3d**) lo puedes poner eh también en **esta pared** (**ev11.2.2d**) pero o se[a] más o menos a **esta altura** (**ev11.3.3d**) en **la pared de este lado** (**ev11.4.3d**) e-en **la pared del fondo** (**ev11.5.2d**)	Can you put this <sil> **this shelf** (**ev11.1.3d**) can you put it erm also on **this wall** (**ev11.2.2d**) but I mean more or less at **this high** (**ev11.3.3d**) on **the wall of this side** (**ev11.4.3d**) on **the back-wall**. (**ev11.5.2d**)	furn.2 wall.3 region.1 wall.3 wall.3
			s: utt12	Task / ∅ / understanding = Ack(utt11)	Ok	Ok	
	Acceptance = action-specification	Presentation = attributes-disambiguation	s: utt13	Task / Info-request=yes, Influence-on-speaker = offer / ∅	¿quieres que ponga **este estante** (**ev13.1.3d**) en **esta esquina** (**ev13.2.2d**)?	Do you want me that to put **this shelf** (**ev13.1.3d**) in **this corner** (**ev13.2.2d**)?	furn.2 region.2
		Acceptance = attributes-disambiguation	u: utt14	Task / ∅ / Answer(utt13), Agreement=Accept	Sí	Yes	
			s: utt15	Task / understanding = Ack(utt14), agreement=Accept / influence-on-speaker = Commit	Ok	Ok	
Acceptance = placement	Presentation = action-accomplishment		s: utt16	Task / ∅ / information-relations = Action-accomplishment(utt11)	*<conjunto de acciones para colocar el estante>*	<sequence of actions to put the shelf>	move(furn.2, region.2)
			s: utt17	Task / Info-request=yes / ∅	¿ Así está bien ?	Is that all right ?	
	Acceptance = action-accomplishment		u: utt18	Task / Answer(utt17) / agreement=Accept	Sí, así esta bien	Yes, that's fine	

Table 1: Multimodal annotation

Introduction	Referents
Initial Context	wall.1 = left wall wall.2 = right wall wall.3 = back wall window.1 = back window ...

	furn.1 = stove furn.2 = shelf furn.3 = refrigerator ...
ev11.3.3d	region.1 = medium height wall.1
ev13.2.2d	region.2 = corner wall.1 & wall.3

Table 2: The graphical context.

3.1. Conversational contribution

Following (Clark and Schaefer, 89) we suggest to extend *DAMSL* with the notion of contribution. A contribution implicates that a common belief between the conversational participants has been established. This can be thought of in two steps: the *contribution presentation* and the *contribution acceptance*. In the presentation a proposition is expressed by one of the conversational participants; in the acceptance step the addressee acknowledges the message and provides evidence that he or she has grasped its propositional or semantic content. A common belief is established when both the presentation and acceptance are successfully accomplished.

In our dialogs, the user expresses an order as clearly as possible during the presentation phase, favoring its interpretation. However the presentation is performed spontaneously, using incorrect words, hesitations, repairs, etc, complicating this task. In case the user auto-detects a possible source of ambiguity, she would present an immediate correction to facilitate interpretation. Therefore, the presentation phase is more than uttering a dialog act; it is the creation of a communicative structure (Clark and Schaefer, 89). Complications arise during the acceptance phase in which the system must be able to indicate its level of understanding. Moreover, the system could have different understanding levels for different parts of the user presentation. The system must show its understanding level through interventions during the acceptance phase. These interventions are called *understanding evidences*.

The need of these evidences is even greater given the interaction features. On the one hand there is no direct visual contact between interlocutors (as in normal telephone communications) which increases the need for a continuous feedback to confirm that a message has been perceived; on the other, the user is unaware of the computer's understanding capabilities. Let alone the fact that task oriented dialogs require clear evidences of understanding.

It is precisely by means of evidence of understanding that we recognize an agreement through the dialog. In our multimodal design domain, a typical conversation can be structured in three agreement levels. The first level is reached when there is evidence that an intervention has been perceived (e.g. when we notice that the interlocutor is paying attention); the second level is achieved when evidence of an appropriate interpretation are observed (e.g. when we notice the interlocutor understands what he or she is intended to do). Finally, a third level of agreement is found where there is evidence of an adequate performance of the requested action. With *DAMSL* it is possible to label the first level of agreement (through the information level). With the contribution model we propose to capture the second and third levels. The dialog can be seen as a sequence of contributions oriented to perform a graphical action. Each contribution involves the specification and performance of a graphical action. The specification involves establishing an order and the referents to act upon, and the performance involves the evaluation of its own result.

In this way contributions are presented in two levels. The external level focused on the performance of the action, *the contribution task level*. When a contribution of this level is over, an agreement about the effects of the requested action has been reached. The inner contribution level –*the contribution description level*– provides us with a clarifying process of the action to be performed as well as a rectification process of the action's effects. In other words, the internal contribution level (presentation and acceptance of the action specification) is in turn a presentation phase of an external contribution focused on the current graphical action. We can assume, for the purpose of our simple design domain, that such kind of actions are determined or specified once the *contribution presentation* has been successfully accomplished and the action performance can be considered a part of the *contribution acceptance*. Then, in a simple conversational model, a conversation can be thought of as a sequence of contributions oriented towards an action performance. The standard contribution is established in two steps. In the first step, one or several sub-contributions are fully oriented to determine the action to be performed. In the second, several sub-contributions are oriented towards performing and evaluating the action.

Table 1 shows an annotated dialog consisting of one contribution with two contribution levels: the contribution task level and the contribution description level. The contribution task level is divided in turn in the presentation and acceptance phases of the "placement" contribution. The contribution description level is composed by two sub-contributions. The first one is used to specify the action to be performed, labeled *action-specificatio*n with their corresponding presentation and acceptance phases. The second sub-contribution establishes the agreement about the effects of the action, and is called *action-accomplishment*. Notice that in the acceptance phase of the first sub-contribution there is a embedded contribution, labeled *attributes-disambiguation*, which resolves an ambiguity introduced in the presentation phase.

4. Conclusions

In this paper we present two extensions to DAMSL scheme that allow us to label utterances in multimodal contexts. The first extension is based on the idea of contribution proposed by (Clark and Schaefer, 89) such that task-oriented dialogues can be modeled as contribution sequences. The second extension consists on the definition a structure where the contextual information useful for the interpretation of spatial deixis is recorded. This extension is based on the intuition that multimodal information supports effective communication by providing referents directly from the context, as discussed in (Pineda and Garza, 2000).

Acknowledgements

This work was done under partial support of CONACyT program for the development of computer science grant 31128A. We express special thanks to Prof. Jean Caelen for his comments. We also thank Miguel Salas Z. for technical support and to the following people for segmenting and transcribing the recorded dialogs: Sergio Santana S., Nora la Serna P., Linda Rosales M., Esmeralda Uraga S., Guadalupe Martínez G., Israel López R., Ivan Meza R and Castalia Negrete P

5. References

Allen, J., Schubert, L., Ferguson, G., Heeman, P., Hwang, C., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M. and Traum, D., 1994. *The Trains Project: A Case of Study in Building a Conversational Planning Agent.* Trains Technical Note 94-3, Computer Science Department, University of Rochester.

Allen, J. and Core. M., 1997. *Draft of DAMSL: Dialog Act Markup in Several Layers*. pp 32. http://www.cs.rochester.edu:80/research/trains/annotation/RevisedManual/RevisedManual.html

Clark, H and Schaefer, E., 1989. Contributing to discourse. *Cognitive Science*, 13, 259-294.

Pineda, L.A. and Garza, G., 2000. A Model for Multimodal Reference Resolution. *Computational Linguistics*, 26(2), 139-193.