The Need for Context in Multi-Modal Interfaces

Elise H. Turner Roy M. Turner
Charles Grunden Jason Mailman
Mark Neale John Phelps
Department of Computer Science
University of Maine
Orono, ME 04469

Abstract

Context is important for AI applications that interact with users. This is true both for natural language interfaces as well as for multi-modal interfaces. In this paper, we consider the kinds of contexts that are important in a multi-modal interface combining natural language and graphical input to describe locations. The descriptions will then be converted into queries to a geographical database system. We have identified several kinds of contexts in our preliminary study. We describe them and consider how each affects the system's interpretation of user input. Plans for future work on the project are also presented, both for implementation and for empirical studies.

Introduction

Context has long been recognized as being important for natural language processing. This includes not only such things as the discourse context (Grosz 1977, e.g.), but also such non-linguistic contextual features as the social relationship between the speakers (Holtgraves 1994). When natural language is combined with other modes of communication, such as graphical input, each mode brings into play its own particular kinds of contexts and is affected by context in its own way. Attention must be paid to all of these contexts and how they work together in order to create a final interpretation of the intended communication.

In addition to giving important information affecting the interpretation of speech and graphics at any point, contextual knowledge brought to light in the course of the communication is potentially useful for knowledge acquisition. For example, the temporal context of what is being described to an AI application using a multi-modal interface ("there used to be a house here") is important to understanding what is being described. After the session is over, information from the temporal context may also be used to augment the application's knowledge base (e.g., by adding information about the prior presence of a house where no such information was previously known).

In this paper, we will describe some contexts we have identified as being important for multi-modal interfaces. We use the term "context" here in two ways: first, to mean the collection of features comprising the current state of the world with respect to the interaction between the user and application; and second, to mean a portion of that context that it makes sense to talk about separately. So, for example, we will speak of the "current context", meaning the current state of the world, as well as contexts concerned with the discourse history, the graphical focus, and so forth.

We first describe our domain and application. We then discuss the different contexts that we have so far identified as being important for multi-modal interfaces. Our work is preliminary at this point, and our conclusions are tentative. Directions for future work are presented in the concluding section.

Sketch-and-Talk

The domain in which we are interested is multi-modal interfaces to geographical information systems (GIS). Our example application is Sketch-and-Talk, which is being created by the Department of Spatial Information Sciences and Engineering at the University of Maine. Sketch-and-Talk will serve as an interface to a database of geographic information. The system will construct database queries from spoken natural language and graphical input from the user. For example, a user might wish to find a map of a plot of land that he or she has some prior information about. The interaction with the system might look like:

> I want to find a piece of land [draws a square on a graphics tablet] in southern Penobscot County that lies on the the river [draws a wavy line northwest–southeast]. It has trees along the river [shades in an area near the wavy line], but is mostly open, old-growth fields.

From this, and possibly additional input, the application would try to build a query to the GIS database to find the parcel of land being described.

Researchers at the University of Maine have done a great deal of work in representing geographical relations and identifying a vocabulary for relationships that correspond to particular sets of representations (Mark & Egenhofer 1994; Egenhofer & Mark 1995). In addition, as shown above, speakers are expected to verbally label entities as they draw them. These labels are then attached to the spatial representations. This work provides the foundation for integrating speech and graphics.

To understand the full range of user input, however, the system must be extended to include context. When speaking, users will rely on discourse context, and other contextual aspects exploited in natural language processing, as a matter of course. When graphics are added, more aspects come into play. To begin our work with this application, we have focused on identifying phenomena that are likely to occur when multimodal input is used, and suggesting the contextual aspects which must be available in order to understand these phenomena.

Method

Because the implementation of the initial system has not yet been completed, we have begun our work by studying ten videotaped examples of members of our research group describing locations or spatial information. These examples may differ from the Sketch-and-Talk application in three significant ways:

- Not all the examples had the task of identifying a particular location.
- Sketches were drawn on a chalkboard, probably giving the user much more ease and flexibility in drawing.
- Locations were described to other humans, not to a computer. So, as expected, the descriptions included humor and were occasionally interrupted to give information of interest that was not necessarily directly related to describing the physical features of the location so it could be found in a database of geographical information.[1]

Despite the informal way in which our data has so far been collected, we believe that we have been able to identify several kinds of contexts that affect the interpretation of multi-modal interaction. We look forward to examining the data that will be collected from users of the Sketch-and-Talk system to refine our notions of the need for context in that application.

Contexts for Multi-Modal Interactions

We have identified several types of context that influence the interpretation of multi-modal communication. Many of these contexts may seem familiar because they provide information that is needed to process natural language. By dividing the information into specific contexts we can begin both to examine the effects of each kind of context as well as identify the kind of knowledge it is important to represent about the context.

In this section, we describe each type of context and suggest the knowledge that should be represented about it. Our idea is that we will explicitly represent both the contexts the application must deal with as well as the contextual knowledge useful to associate with those representations.

Discourse Context. The discourse context contains all of the entities that are mentioned in the discourse. It is important for understanding a variety of phenomena, such as referring expressions and clue words. This context is broken into several subparts, or *discourse segments.* Discourse segments are made up of contiguous utterances that are related to the same topic. A model of how a speaker can move between these segments is required to properly model the discourse context. Many techniques already exist for creating the discourse context and moving between its segments (Grosz 1977; Reichman 1985; Grosz & Sidner 1986, e.g.), and any of these could be adopted for our system.

Graphics Context. The graphics context includes all of the entities that have been drawn and their spatial relations. For our work with Sketch-and-Talk, we will use the entity and relation representations used by that project (Egenhofer & Herring 1990).

We have found that, like discourse, the graphics context should be divided into *graphics spaces.* We have seen indications that users consider the graphics context to be subdivided. Users speak of the "the area around *(some entity)*". They also deviate from their established order of drawing to draw certain related objects. For example, a user who has been drawing entities from left to right may deviate from this pattern to draw all of the outbuildings surrounding a house. Entities in the graphics spaces are often all related to a single entity or function. For example "where we fished" may constitute a graphics space. We have seen another indication that graphics spaces organize the graphics context. Users can easily refer to a graphics space with a single reference, for example, by pointing or referring to the most significant entity (in our example, the house). Clearly, the graphics spaces and discourse segments will be closely related because users are expected to talk as they draw.

Future work on this project will include discovering exactly what constitutes a graphics space and how a speaker/drawer moves between them. Part of this will also include determining what kinds of contextual knowledge it makes sense to represent as part of the representation of the contexts.

Task Context. This context provides information related to the task that the user is pursuing. Knowledge that would be useful for the application to have about this context includes likely goals of the user as well as standard procedures to achieve those goals. The task context influences the flow of the communication (Grosz & Sidner 1986), as well as helping to identify important entities and concepts.

[1] We are not ruling out the possibility that some of this might happen as well during interactions with the application.

In Sketch-and-Talk, the task will be to create a database query. In this task context, we would expect the user to make queries as straightforward and clear as possible. This was also true when videotaped "users" were working on the assigned task. However, we saw two additional task contexts that were interspersed with the assigned tasks in our examples. First, there was a *"chit-chat" context* in which users put aside the task of describing a location to interact with or entertain the observers. This context was often marked by the user turning away from the chalkboard. In this context, users told jokes or personal stories related to the location. They seldom referred to their drawings, and, when they did, only pointed to a single, specific location. We would not expect to see this context in Sketch-and-Talk, but it may be important in other multi-modal interfaces (e.g., those supporting computer-supported cooperative work).

Second, there is a *drawing correction context* that is a subtype of the description task context. We believe this is an important context to be studied for multimodal interactions. The user and application are in this context when drawings are being corrected. It is important to recognize this context so that the representation of the graphical input is changed properly. Recognizing this context is important for the application understanding the user's speech. During correction, speech is often mumbled and has the feel of "talking to oneself". Consequently, it is likely that it will not be correctly understood by the speech recognition system. Fortunately, however, other than being used to help recognize the shift in context, understanding the speech may not be crucial. Realizing that it is in this context may allow the interface to know that it can safely ignore most if not all of this the error-prone speech.

Target Location Context. In Sketch-and-Talk, the kind of location that is the target of the query also constitutes an important context. The interface can use its (or its application's) world knowledge about the location type to aid in interpreting the user's input; different target location contexts will likely give rise to different interpretations. Although world knowledge has long been considered an essential part of the contextual knowledge needed for understanding natural language, it is also clear that it is necessary for interpreting graphics. For example, a curving line may indicate a road in the context of describing a residential neighborhood and a stream in the context of describing an undeveloped forest. Similarly, if the interface understands the target location context to be "a farm", it can use its knowledge about farms to understand the user's reference to something being near "the barn", even if no barn has previously been mentioned or sketched.

Since the identity of the target location unfolds as the task is being carried out, Sketch-and-Talk must be able to decide the target location context as it is being discussed. Some users may be able to provide information specific enough so that the target location context can be represented in detail (e.g., "I'm looking for the main entrance to Acadia National Park."). However, we expect most users to identify only the type of location as they begin a session with Sketch-and-Talk (e.g., "I'm interested in a forested lot", or "I'm looking for a house in a neighborhood near a school."). The representation of the target location context should include more detail as the target location is refined.

This is not simply a matter of moving though a hierarchy of location contexts. Instead, we expect the context to be pieced together from representations of existing contexts. For example, the current target location context may be a forested lot. If picnic tables are added to the sketch by the user, then the current context must be merged with the target location context of a picnic area. If, instead, logging roads are added, then the forested lot context must be merged with the logging site target location context. An important area of future work for this project is how to merge contexts by merging the program's corresponding contextual knowledge.

User Context. Properties of the user also define a useful context. Knowledge useful to have about the user includes his or her goals, beliefs, level of expertise, style of interaction, and idiosyncrasies. This contextual knowledge can then be used by the system to help it understand what the user is trying to communicate. It is likely that the system will benefit from explicitly representing both kinds of users as well as particular users. Contextual knowledge associated with the former can provide predictions and information about how a new user will interact with the system. Contextual knowledge associated with the latter can provide more specific information, perhaps gained from the system's own history of interacting with the user, about how a particular user differs from general expectations. In traditional approaches to interaction, this kind of information is often stored in *user models* (Paris 1987; Carberry 1988, e.g.), which can be viewed as a kind of explicit context representation.

In our application, the style of interaction and idiosyncrasies of the user are particularly interesting. In multi-modal communication, unlike natural language communication, conventions are not necessarily shared by the community of users. Instead, individuals develop their own styles of interacting. For example, some research group members labeled most of the entities in their drawing as a matter of course. Others only labeled entities that had important names that were critical to the description of the location. In the latter group, the fact of the labeling itself had significance; in the former group, it had none. In such cases, the styles of the interaction work like conventions in natural language (Grice 1975). Part of our work on the user context will be to better understand particular behaviors of users and the roles they play in interpreting the input.

Temporal Context. As do many other things an application might be concerned with, locations change over time. A user's description of a location has a *primary temporal context.* If the user has only seen the location at one time, or is describing features of a prototype location that he or she would like to find, this primary temporal context will be the only one that is needed to interpret the user's input.

Often, however, a user may describe a location, or events related to a location, at several different time periods during the course of a session with the system. For example, when a user is describing a favorite spot for family vacations from his or her childhood, the period of those vacations will most likely be the primary temporal context. Other temporal contexts may be invoked if the user has seen the location at different times. For example, if the user has returned to the family vacation spot for a short trip, the user may include information about changes to the location. In this case, the user may indicate that a field he or she referred to earlier (in the primary context) has changed by saying, "There's a house on it now", and, possibly, drawing a symbol for the house. Later, the user may refer again to that location as "the field". If the statement about the house is seen as a correction to the description rather than information about a different context, the later reference to "the field" will not be properly understood.

A representation of temporal context will need to include the description of the location as well as the time at which the description is valid. Other temporal context representations can inherit information from the representation of the primary context. In our example, there would be only two such contexts: the time of family vacations and the time of the most recent visit. The time of family vacations would, itself, cover a wide range of time. If instead an area were being developed over time, then more temporal contexts would be involved and the time that they cover would be more specific. In dealing with additional contexts, it will be important to decide how knowledge can be inherited between representations of secondary temporal contexts. For example, it may be appropriate to allow information to be inherited from earlier context representations by later ones.

Temporal contexts will be particularly important for the Sketch-and-Talk application. Geographical databases may contain maps and other descriptions of locations that differ based on time. To retrieve the correct information, it will be necessary to match a query from a single temporal context to information that was collected at the corresponding time. By representing temporal contexts as descriptions of the locations as they existed at a particular time, the correct temporal context can be chosen to use to create the database query.

Legend Context. We have noticed that occasionally users provide a legend for symbols that they will use during a particular session with Sketch-and-Talk. This information also defines a context, in particular, the context in which those symbols have those meanings. In our example, these legends were not provided all at once, before the description of the location began. Instead, users would identify the meaning of the symbol for the session. This identification could be quite explicit (e.g., "areas that I'm drawing with double lines are tidal areas"). In this case, the symbol almost certainly was given its identity for only the current session. In other cases, a symbol is identified implicitly. For example, a user may draw a square for a building, identify it as a building, and then continue to draw squares without identifying them as buildings. In the latter case, it is less clear whether the square will be used as a building for a single session or is the user's preferred symbol for buildings in this task. Currently, we believe the legend context is applicable only for one session. This distinguishes it from information about symbols that can be consistently associated with users or tasks across multiple sessions.

Specific Symbol Context. Some symbols, or types of symbols, create contexts that extend beyond the objects that they represent. We saw at least two examples of this in our data. First, some shapes were meant to be interpreted as being more meaningful than others. Although drawings were expected to be approximate, squares and straight lines were expected by the user to be interpreted as squares and straight lines. When one user drew a square, but did not intend the property to be square, he explicitly stated that he did not care about the shape of the property. Roads were drawn as curved lines, except in one case where city blocks were drawn. This defines a context in which symbols have certain semantics, similar to the legend context. The chief difference is that this kind of context is usually not explicitly marked by the user. Information from the symbol's context can be overridden. For example, a building may be drawn as a square in the context of any task not related specifically to the architecture of the building. So, a square representing a building may not lead to the interpretation of the building being square.

Second, graphical objects that can be seen as containers are expected to contain the thing that they are meant to locate. For example, circles, squares, and curved lines which intersect with the boundary of the location form containers. If a user says, "there are woods on the property" and draws a container, the woods are expected to be inside the container.

The drawings that we are studying are not detailed. In our examples, a small set of symbols was used for all of the examples. Consequently, through empirical study, we expect to be able to create rich representations of the context created by the use of each symbol.

Environment Context. This relates to the environment that the user is in. This includes the user's location, the equipment used, and the presence of observers or other participants in the session. It is quite likely that the environment context will be constructed from many subcontexts which must be merged. The expected differences between our videotaped examples and examples from Sketch-and-Talk make the need for the environment context clear. For example, the equipment used for drawing in Sketch-and-Talk will be more limiting than drawing on a blackboard. This will impact knowledge from the specific symbol contexts and knowledge about drawing from many other contexts. In addition, Sketch-and-Talk users will be interacting only with a machine instead of with a group of colleagues and friends. Consequently, we would expect the "chit-chat" task context to disappear completely.

Conclusion

In this paper, we have discussed some preliminary work we have done on identifying contexts and contextual knowledge important to multi-modal interfaces. We have so far identified the following contexts, based on examining videotapes of research group members simultaneously talking about and drawing locations: discourse, graphics, task, target location, user, temporal, legend, specific symbol, and environment contexts.

Our future work on this project will follow two paths. We will perform empirical studies of users interacting with the initial versions of the Sketch-and-Talk system to provide more rigorous data about the kinds of contexts important to multi-modal interfaces. This data will allow us to refine and extend the initial set of contexts we have identified. We will also pay close attention to the kinds of knowledge the users seem to have about the various contexts. This will help determine what knowledge should be associated with representations of the contexts. Based on the sessions with users, we will also identify contextual knowledge the interface would need in order to process the users' input.

Simultaneously, we will begin to design an approach to explicitly representing and using contexts and contextual knowledge in multi-modal interfaces. This will involve considering such issues as: what contextual knowledge is useful to represent, what distinguishes one context from another, how contexts (and thus, their explicit representations) are related, and how contextual knowledge can be merged to capture the interaction of multiple contexts. One possibility is to base the mechanism on one already devised by a member of our research group (Turner 1998); other possibilities will also be considered. To test our work, the resulting mechanism will be implemented, integrated with the Sketch-and-Talk system, and evaluated empirically.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-9613646. The authors would like to thank Max Egenhofer and the other members of UM's Department of Spatial Information Science and Engineering who are developing the Sketch-and-Talk system.

References

Carberry, S. 1988. Modeling the user's plans and goals. *Computational Linguistics* 14(3):23–37.

Egenhofer, M., and Herring, J. 1990. Fourth international symposium on spatial data handling. 803–813.

Egenhofer, M., and Mark, D. 1995. Spatial information theory - a theoretical basis for gis, international conference cosit '95, lecture notes in computer science 988. Berlin: Spring-Verlag.

Grice, H. P. 1975. Logic and conversation. In Cole, P., and Morgan, J. L., eds., *Syntax and Semantics*, volume 3. New York: Academic Press.

Grosz, B. J., and Sidner, C. L. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

Grosz, B. J. 1977. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the Fifth International Conference on Artificial Intelligence*, 67–76. Los Altos, California: William Kaufmann, Inc.

Holtgraves, T. 1994. Communication in context: Effects of speaker status on the comprehension of indirect requests. *Journal of Experimental Psychology* 20(5):1205–1218.

Mark, D., and Egenhofer, M. 1994. Calibrating the meanings fo spatial predicates from natural language: Line - region relations. In Waugh, T., and Healey, R., eds., *Sixth International Symposium on Spatial Data Handling*, 538–553.

Paris, C. L. 1987. *The Use of Explicit User Models in Text Generation: Tailoring to a User's Level of Expertise.* Ph.D. Dissertation, Columbia University.

Reichman, R. 1985. *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics (An ATN Model).* Cambridge, Mass: The MIT Press.

Turner, R. M. 1998. Context-mediated behavior for intelligent agents. *International Journal of Human–Computer Studies* 48(3):307–330.