Aspects of Context for Understanding Multi-Modal Communication

Elise H. Turner, Roy M. Turner, John Phelps, Mark Neal, Charles Grunden and Jason Mailman

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 aspects of context that are important in a multi-modal interface combining natural language and graphical input to describe locations. We have identified several aspects of contexts in our preliminary study. We describe them here and discuss plans for future work.

1 Introduction

Context has long been recognized as necessary to understanding natural language communication. Several components of context have been identified and studied, including discourse context [1] and user models [2]. We consider context to be anything that is required to understand an utterance beyond syntax and semantics. We call the components of context *contextual aspects* or simply *aspects*.

In this paper, we will discuss contextual aspects which we have identified for understanding user input in speech and graphics. A distinct aspect was recognized for one of several reasons. First, all of the knowledge contained in the aspect might be connected by a theme. Second, different aspects would be relevant to, or function differently when, handling different phenomena. Third, some parts of the context were divided into separate aspects because they are managed differently. Similarly, some aspects were separated from others due to the duration of the information in the aspect (i.e., is the information relevant during a single session or across many sessions?). Fourth, some aspects have been separated out, for now, because they are well-studied elsewhere. This is the case for the discourse aspect.

Our work is to be applied to Sketch-and-Talk, a multi-modal interface to geographical information systems that is being created by Max Egenhofer and his colleagues in the Department of Spatial Information Science and Engineering at the University of Maine. The system will construct database queries from spoken natural language and graphical input from the user. 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. This preliminary work has lead us to identify several contextual aspects that affect the interpretation of multi-modal interaction. These aspects are presented below. More detail can be found in [3, 4].

2 Contextual Aspects for Multi-Modal Interactions

Discourse Aspect. The discourse aspect is known in natural language processing as the *discourse context*. It contains all of the entities that are mentioned in the discourse. 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. Many techniques already exist for creating the discourse context and moving between its segments [1], and any of these could be adopted for our system.

Graphics Aspect. The graphics aspect 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 [5].

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. Users also draw detailed views of particular regions of the location and move between the overview and detailed views during a session. Entities in a graphics space are often all related to a single entity or function. For example "where we fished" may constitute a graphics space. Also, users can easily refer to a graphics space with a single reference, for example, by pointing or referring to the most significant entity.

Clearly, the graphics spaces and discourse segments will be closely related because users are expected to talk as they draw. For now, we keep the discourse and graphics aspects separate to take advantage of the work that has been done to develop representations and management algorithms for the discourse aspect. In future work, we plan to explore the relationship between these two aspects. This includes determining if they are truly separate. Future work will also include discovering exactly what constitutes a graphics space and how a speaker/drawer moves between them.

Task Aspect. This aspect provides information related to the task that the user is pursuing. For our application, the representation of this aspect will include likely goals of the user as well as procedures for achieving those goals. In addition, we saw evidence of a *social interaction task aspect*, in which users put aside the task of describing a location to interact with or entertain the observers, and a *drawing correction task aspect*. The task context influences the flow of the communication [6, 7, 1], as well as helping to identify important entities and concepts. The information represented for this aspect will vary, depending on the task.

Location Aspect. In Sketch-and-Talk, the kind of location that is the target of the query also constitutes an important aspect of the context. We expect the application to have world knowledge about the location that it can access to build or respond to database queries. The location aspect brings this information into the context. Other types of world knowledge will be needed to understand

the speech, and, at times, the graphics. However, we create a separate aspect for location because it is so important for interpreting symbols and for managing the graphics aspects.

Since the identity of the target location often unfolds as the task is being carried out, Sketch-and-Talk must be able to determine the location aspect as it is being discussed. The representation of the location aspect then includes more detail as the location is described by piecing together representations. For example, the current location context may be a forested lot. If picnic tables are added to the sketch by the user, then the current location aspect must be merged with the context of a picnic area.

User Aspect. Knowledge of the user's goals, beliefs, level of expertise, style of interaction, and idiosyncrasies, traditionally stored in *user models* [2, 8], constitute the user aspect. In our application, the idiosyncrasies and style of interaction 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. Consequently, individual styles of the interaction work like conventions in natural language. 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 Aspect. Locations change over time. While drawing, the user may refer to different features of the location that existed at different times. The system may also be aware of differences in the location at different times. In order to understand what the user is saying and drawing, the interface needs to model and track the temporal context. Thus we have identified a temporal contextual aspect as a separate kind of aspect.

A user's description of a location has a *primary temporal aspect*. 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 aspect will be the only one that is needed to interpret the user's input. Other temporal aspects may be invoked if the user has seen the location at different times.

Legend Aspect. We have noticed that occasionally users provide a legend for symbols that they will use during a particular session. This information also defines the aspect of the context in which those symbols have those meanings. The legend aspect is applicable only during the current session. This distinguishes it from information about what the symbol denotes that can be consistently associated with users, locations, or tasks across multiple sessions.

Environment Aspect. The environment that the user is in also affects the interaction with the system and should be represented separately. The environment aspect includes the user's location, the equipment used, and the presence of observers or other participants in the session.

3 Discussion

Lenat [9] discusses some of the problems with monolithic context representations based on experience with the Cyc program. He delineates twelve dimensions of

"context-space" in response, four of which are similar to two of our contextual aspects. His two spatial dimensions, "Absolute Place" and "Type Of Place", are related to our location aspect, and his two temporal context dimensions, "Absolute Time" and "Type Of Time", are related to our temporal aspect (below). It is difficult to see, however, where the remainder of the contextual aspects we have identified would fit in his framework.

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

4 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IRI-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. The authors also wish to thank Patrick Brézillon for his helpful comment on an earlier version of this paper.

References

1. B. J. Grosz and C. L. Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.
2. C. L. Paris. *The Use of Explicit User Models in Text Generation: Tailoring to a User's Level of Expertise*. PhD thesis, Columbia University, October 1987.
3. E. H. Turner, R. M. Turner, C. Grunden, J. Mailman, M. Neale, and J. Phelps. The need for context in multi-modal interfaces. In *Workshop Notes for the 1999 AAAI Workshop on Reasoning in Context for AI Applications*, AAAI Technical Report (ISBN 1-57735-098-7), pages 91–95, Orlando, FL, July 1999. AAAI Press.
4. E. H. Turner, R. M. Turner, J. Phelps, M. Neal, C. Grunden, and J. Mailman. Aspects of context for understanding multi-modal communication. Technical Report 99-01, Department of Computer Science, University of Maine, 1999.
5. M. Egenhofer and J. Herring. Fourth international symposium on spatial data handling. pages 803–813, 1990.
6. B. J. Grosz. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the Fifth International Conference on Artificial Intelligence*, pages 67–76, Los Altos, California, 1977. William Kaufmann, Inc.
7. R. Reichman. *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics (An ATN Model)*. The MIT Press, Cambridge, Mass, 1985.
8. S. Carberry. Modeling the user's plans and goals. *Computational Linguistics*, 14(3):23–37, 1988.
9. D. Lenat. The dimensions of context-space. Published on-line at URL http://www.cyc.com/context-space.rtf, accessed March 28, 1999., 1998.