In CONTEXT'01.

Representing the Graphics Context to Support Understanding Plural Anaphora in Multi-Modal Interfaces*

Elise H. Turner[1] and Roy M. Turner[1]

Computer Science Department
University of Maine
Orono, ME USA 04468-5752
{eht,rmt}@umcs.maine.edu

Abstract. Previous communication provides important context for new communication in an interaction. In natural language interfaces, the discourse context represents and maintains information about what has been said before. When other modes of communication are also used, they must also contribute to the context. In this paper, we describe how information about the graphics can be represented and maintained in the graphics context. We are particularly interested in how the graphics context can be used to support finding referents for plural anaphora.

1 Introduction

Understanding plural anaphoric references is a difficult problem for natural language interfaces. To understand anaphora, a discourse context that represents the entities that have been referred to in the discourse must be maintained. To find the correct referent, the system must create a set with the proper membership. Some sets that can serve as referents are clearly indicated in the discourse by a plural head noun (e.g., "the dogs") or the use of a conjunction (e.g., "Lassie and Clifford"). Others are not so clearly marked and must be pieced together from distinct noun phrases. This requires knowledge gained from reasoning about the world which may be time-consuming and may be difficult to focus.

In multi-modal interfaces, when users are allowed to communicate using graphics as well as natural language, correctly understanding plural anaphora can still be difficult. As with discourse, the graphical communication itself can provide important clues to the membership of the set. However, because the graphics remains visible throughout the interaction, sets can be formed from related icons that are entered into the graphics over time. Unlike relationships between entities that are added over time in the discourse, these relationships

* This work was funded in part by grant IIS–9613646 from the National Science Foundation. The authors would also like to thank Shawnee Treadwell for her work on transcibing videotapes, and the anonymous reviewers for their helpful comments on an earlier draft of this paper.

do not require time-consuming general purpose reasoning to be identified. Instead, they are relationships such as sharing the same icon that can be easily and quickly perceived when viewing the graphics. This means that, unlike the discourse context, the graphics context must support relationships between entities that may only be known much after the first entity is drawn.

In this paper, we will describe how the graphics context can be represented to support finding the proper referent of "these." We begin in Section 2 by discussing how understanding "these" in natural language interfaces differs from understanding "these" when graphics are also used. In Section 3 we illustrate how "these" can be understood with an example from a videotaped session of a speaker describing a location. We present our representation of the graphics context in Section 4.

Before we begin our discussion, we need to specify how some terms will be used throughout the paper. We will use *anaphora* for expressions that refer to entities that have been referred to previously, either in discourse or graphics. We will use *discourse* to refer only to spoken or written communication. We will use *communication* to include both discourse and graphics. We will call the extended multi-modal communication between the user and the system the *interaction*.

2 Understanding "These" in Multi-Modal Interfaces

In order to understand plural anaphora, a system must have two types of information about entities that have been referenced in the communication. First, it must know which entities are currently *available* as referents to anaphoric expressions. Second, it must know the *membership criteria* that identifies which of the available entities belong in the referent set.

In natural language systems, entities that have been referenced in the discourse are stored in the *discourse context*. Many systems use simple history lists for the discourse context [1]. Entities are placed on the list in the order in which they are referenced in the discourse. When the system must find the referent for an anaphoric expression, it searches the history list, considering the most recently referenced entities first. Items are removed from the history list after a specified period of time. This is useful because people have limited short term memory and may not remember entities referenced much earlier. However, when a set must be formed with entities that have been mentioned throughout the discourse, they may not all be available at the time the plural anaphora is used. Similar problems occur with methods of representing the discourse context that are designed to better reflect the structure of the discourse (e.g., [2–4]). These methods document the progress of the discourse by way of its topics. Entities referenced while the topic is being discussed are associated with the topic in the discourse context. Some schemes distinguish certain entities as being in higher focus than others. Only entities that are in high focus can be referenced by a pronoun. As the discourse moves from topic to topic, the entities referenced in some previous topics may be referred to anaphorically. Other previous topics may be closed in such a way that prohibits anaphoric reference to their entities.

Although a speaker can return to a closed topic, it is unclear whether and for how long entities associated with a topic can be referenced anaphorically.

Entities do not become unavailable over time in the graphics context. Discourse is ephemeral. Speech "goes away" as it is spoken, leaving the hearer with only a mental representation of what was said. Although written text does not actually disappear as it is read, the reader is not expected to have to re-read passages in order to understand anaphoric references. Graphical communication, on the other hand, remains accessible throughout the entire interaction. In our videotaped examples, graphical communication was only made inaccessible if it was erased. Although we have not studied erasures, we believe they effectively eliminate the erased entity from the graphics context. In other communication, graphics may be occluded (e.g., covered by other windows in the system) or otherwise made temporarily unavailable. However, to communicate by pointing, the speaker must believe that the hearer will be able to access the graphics (e.g., by moving the window). In these ways, graphics are available for the user to refer to, and to reinterpret, at any time later in the interaction. Consequently, the system's representation of the graphics in the graphics context cannot simply point to entities in the system's knowledge base, but must store the information that will allow the system, like the user, to reinterpret those graphics.

Because the graphics can be viewed and reinterpreted throughout the interaction, it is easy for speakers to add new entries to existing sets. By a simple pointing gesture the speaker is able to draw the user's attention to all of the entities in an area. The speaker can then add entities to the area and relate them to nearby entities – no matter how distant the last reference to those entities or how unrelated their associated topic. The graphics is also repository for shared knowledge for the interlocutors who can see it and who have understood as it was built. As a result, with graphics a speaker can point to an icon and say "these" and expect the user to be able build a set of all similar items that have been referenced throughout the interaction. Using discourse alone, the speaker would have to refer to the same set of *X's* using an expression like "all of the *X's*".

In addition to differences in availability of entities, the graphics context also differs from the discourse context in the ways that entities can be identified as members of certain sets. Both can rely on general purpose knowledge and reasoning to make the connections that group entities into sets. However, this requires significant effort and can easily lead to miscommunication. Membership criteria can be interpreted more efficiently and more effectively if they rely on standard relationships that can be easily apprehended from the communication.

For discourse, this is through the use of plurals and conjunctions. Sets derived from these sources can be added immediately to the discourse context. It is more difficult to add new entities to these sets, requiring that the set be available and that the hearer recognize, usually with the aid of additional reasoning, that the new entity should be a set member. For graphics, the mode of communication increases the ease of creating sets. As discussed above, all entities remain available throughout the interaction. More important, entities can be grouped together

in several ways that are easily perceived by looking at the graphics. Specifically, entities can be related by location and by icon. In addition, a speaker can refer to a set by pointing to one of its members in the graphics, but relying on the discourse context to help understand the referent. For example, among many buildings, a speaker might draw several buildings using the same icon while saying "There are some new buildings in this area." She might next point to one and say "These are all up to code." The referent set can be best identified as the new buildings using information from the discourse.

In the current work we are more interested in constructing the proper set for a plural reference than in the deictic nature of "these." Consequently, we have focused our work on gestural usages when a speaker is pointing to a member of a set and when information from the graphics context is needed to find the proper referent. We believe the graphics context that we are developing to support creating appropriate sets can provide a foundation for handling deixis in multi-modal interfaces. Specifically, we believe that the support for relating entities by location will be useful for understanding gestural usages of "here." In the future, we will focus more on "these" as a deictic expression. In particular, we would like to explore how the gesture used (e.g., pointing to a single object vs. pointing to several objects vs. a sweeping gesture) helps to indicate the breadth of the set in terms of the relationships that we have identified.

3 Example

Our work on context will be used to support Sketch-and-Talk, a multi-modal interface to geographical data being developed by the Department of Spatial Information Science and Engineering at the University of Maine. With Sketch-and-Talk, users will describe locations using speech and graphics. To begin understanding the role of context in these interactions before the system was implemented, we examined videotaped sessions of humans describing locations. Descriptions were provided by students and faculty from a research seminar class studying the use of context for understanding multi-modal communication. Class participants chose a location and described it using a chalk board for graphics. This gave the participants greater flexibility than they would have with existing software for recognizing graphical input.

The following example includes communication that helped motivate our representation of the graphics context, including a gestural usage of "these." It was transcribed from our videotaped sessions and is representative of them. A fragment of the discourse is shown in Fig. 1. The letters appearing in the text correspond to the labels on icons in the figures showing the graphics. These labels were not included in the original drawing. Labels appear after the word that was being said as the speaker began to draw the corresponding icon.

The speaker has chosen to describe Appledore Island, which is the home of the Shoals Marine Laboratory. He first introduces the topic and draws the island (Utterance 1, Icon A). Next, he draws many of the buildings that house the Shoals Marine Laboratory, starting with Kiggins Commons (Icon B, and

1. (A) The Shoals Marine Lab out on Appledore Island down south.
2. And Appledore is an island that's kind of got a wasp waist to it.
3. It's going to be two islands in a few years if they're not careful.
4. Um, we have the Kiggins (B) Commons here which has a nice porch (C) out back overlooking like a little ravine (D) are and the rest, the rest of the island is out here.
5. There's a laboratory (E) over here and classroom (F,G) buldings.
6. And (H) there (I) are dorms (J) back here.
7. And (K) also out here.
   discussion of particular buildings that were discussed above
8. Ok, um, we would do a lot of our work over in this lab.
9. It was a big nature lab.
10. It had a couple of nice features.
11. One is it (L) overlooked, ah, across the little, the little area over here overlooked Starr Island which (M) had some of the old hotel era uh era huge hotels on it.
12. It was essentially a resort, not really a resort, a retreat run by the Unitarians and Congregationalists.
13. So, you could look over there and see that.
14. You could also see the sunsets, that was nice.
15. Portsmouth's down here. Portsmouth, New Hampshire, six miles that way.
16. Across there was also another (N) building here, uh, and here (O) or so.
17. Across the ravine there was some...
18. Most of what we were interested in was the inter-tidal which was all around.
19. Um, down here on this part, um, there's a harbor here, too, sorry, Kinsb the ship, the boat would park here, the Kingsbury.
20. And, us, then they divided the area up into transects, places where people would do their experiments.
    discussion of transects
21. Ok, this is low, high, very low, and back up to pretty high.
22. There are some cliffs down here.
23. Uh, around here was an area that people had started building a cairn (P) on, this is a ritual sort of thing, people take rocks up there.
24. But this is where Celia Thaxter, I don't know if you know who that was,
    discussion of Celia Thaxter
25. She's sit up here and she'd look out at Portsmouth down here.
26. Across the way here was where (Q) people would play volleyball.
27. And then around here was the remains of Celia Thaxter's garden (R), or Celia Thaxter's garden and the remains of her (S) house.
    discussion of area around Celia Thaxter's house
28. There were trails back between all these (T–X) through the scrub.

Fig. 1. Discourse example.

porch Icon C) in Utterance 4. When the first square (Icon B) is drawn, a set for entities represented by a square is formed. Each time a new entity is drawn as a square icon, it is added to this set.

In the same utterance after discussing the Commons, he draws a "ravine area" and notes that it separates the Lab from "the rest of the island...out here". This indicates that a set should be formed for this area. This set should contain all entities currently in this area as well as those that will be added later. The speaker then goes on to talk about the buildings of the Lab. Fig. 2 shows the graphics at this point.

Fig. 2. Drawing through discussion of the lab buildings.

Next, the speaker talks about and draws two areas that provide interesting views from the lab. This includes a description, in Utterance 11, of Starr Island (Icon L) and the hotel on it (Icon M). The icon for the hotel is clearly a rectangle instead of a square, so a new set is created for icons that are large rectangles. The speaker then returns to the topic of the buildings at the lab and adds two more (Utterance 16, Icons I and O). The icons, the locations of the buildings, and the

fact that only Laboratory buildings have been discussed so far all lend evidence to understanding these buildings to be laboratory buildings. Recognizing that only laboratory buildings have been discussed would require some reasoning by the hearer. However, this is the only means of establishing the connection between the buildings using only speech.[1] In that mode, no set would be formed that contained all of the buildings at the time of understanding. With graphics, both the icons and the locations would cause the newly-referenced buildings to join sets containing the existing buildings. The icon set that these buildings join is the one created at Utterance 4, and the location set is the one created at Utterance 16. Though identified as related by icon or location instead of by organization, these sets will allow us to accurately understand a future reference to the buildings of the Laboratory.

The speaker next discusses the part of the island across the ravine. This includes a discussion of Celia Thaxter's garden and house (Utterance 26) and a discussion of the volleyball courts (Utterance 27). These are drawn with square icons like those used for the laboratory buildings, but larger (Icons O–S). Fig. 3 shows the graphics at this point.

At Utterance 28, the speaker draws trails between some Lab buildings, as shown in Fig. 4. He begins drawing trails as he says "these," at that point drawing the trail (Icon T) between two buildings. He does not connect all buildings with the trail, yet two independent reviewers of the videotape understood "these" to refer to all laboratory buildings. When asked, the speaker reported that he intended that interpretation.

The speaker has linked two square icons (Icons B and N). We would expect "these" to indicate an existing set, and the set of entities drawn as square icons has already been formed. "These," as opposed to "all these" or "those," also indicates that the entities in the set are near to each other. The ravine area is the smallest identified area that contains the linked icons. We find the referent to "these" by taking the intersection of these sets.

The set of Laboratory buildings could also be created by reasoning about the information given in the discourse about the buildings. However, the icons can so the same work more efficiently. We believe the icons that a speaker chooses indicates how he or she is dividing up the world. In general, entities that are seen by the speaker as related in an important way are drawn using the same icon. It is also possible for slight, but regular, variations in the icon to represent subclasses in the icon class. Icons often reflect the shape of the entity, but more often are used for this sort of grouping. There are exceptions to this. Speakers may embellish an icon. For example, the porch (Icon C) added to Kiggins Commons (Icon B) is an embellishment. It should not change the type of icon for Kiggins Commons and should only be included in an icon set with similar icons that are also used as embellishments. Some speakers changed icons when drawing more of an often repeated type of entity. Speakers usually changed from

[1] We assume the buildings would be referenced by speech only. However, that speech might include additional information that would more easily link the newly-referenced buildings to the other Lab buildings.

Fig. 3. Drawing just prior to trails being added.

a complicated icon or one that was more time-consuming to draw to X's or dots. When speakers did this, they also were not as precise about the objects in other ways. Specifically, they were often not precise about the number of objects or their exact location. In this way, entities drawn as X's or dots had a diminished status. Speakers also reused icons. The separate uses were usually clearly identified in the discourse. Occasionally, there was a clear point in the discourse when the icon changed meaning. More often, either none of the entities represented by the icon were so important that they needed to be identified in the graphics (i.e., the icon created the set of "other things"), or all but one use of the icon followed a standard interpretation of that icon (e.g., two parallel lines to designate roads and to designate inter-tidal regions).

The general rule is upheld in our example. Squares were used to indicate "the places lab members frequented." Smaller squares represented buildings and larger squares represented other areas. In discussing understanding "these," we did not divide the square icon class into subclasses. This would have made understanding no harder because the buildings alone would form a group. A different icon was used to indicate the cairn area (Icon P), which was a place of special significance; no icons were used to indicate transects, each of which was only frequented by a few lab members. The speaker also used a different icon for buildings that lab members did not visit. In Utterance 11, the speaker uses a rectangle (Icon M) instead of a square to indicate a hotel. After the fragment shown in the figure, the speaker refers to houses that were owned by lobstermen and that were not part of the Lab. He draws squares for these houses and then draws X's over the squares.

4 Representing the Graphical Context

In addition to the drawing itself, the interface needs to represent two kinds of information about the graphics context: the graphical objects present and various ways of grouping those objects into sets that could be the referents for plural references.

Graphical objects can be thought of as augmented icons. A graphical object is the icon drawn by the user plus additional information that is either given by the user or derived by the system, including information about what object in the world the icon stands for. In Fig. 4, graphical objects include the dormitories, the laboratory, Kiggins Commons, the cairn, and the ravine. In drawings of other kinds of locations, graphical objects such as roads, rivers, lakes, and so forth, would be expected.

Fig. 5 shows a graphical object represented as a frame. The *icon* slot holds a frame-based representation of the icon that was drawn. In addition to the bitmap of the icon, that frame will contain information about the shape (e.g., a square), the approximate size, the icon's location, and the legend used to interpret the icon.

The *legend* is an aspect of the current context that the system tracks (see [5]). It contains information that, like a map legend, links icons to meanings.

Fig. 4. Drawing for entire discourse fragment.

The legend contextual aspect may contain information that is user- and task-independent, such as the fact that squares often denote buildings. It may also contain information about the way icons are used in the current task. For example, as mentioned above, in some drawings in our protocols, both squares and X's were used for buildings; this information would be stored in the legend.[2] Iconography that is idiosyncratic for the user will come from contextual information having to do with the user [5].

```
^g-dorm1:
  isa: ^graphical-object
  icon: ^icon1
  object: ^dormitory1
  inferences: nil
```

Fig. 5. A graphical object representing one of the dormitories.

The *object* slot contains a pointer to the frame representing the object this graphical object stands for. In the figure, the graphical object refers stands for a particular instance of "dormitory".

The *inferences* slot contains any inferences that were made in order to fill in the other slots.

In order to resolve plural anaphoric references, candidate referent sets must be constructed and considered. In our approach, these sets are created and maintained as objects are added to the drawing based on relationships that are likely to be useful and that make set creation and maintenance relatively easy. The sets are explicitly represented and associated with the graphics context. At the time a plural reference is made, they can be quickly examined to determine the referent. In our approach so far, we are considering three kinds of sets: location sets, icon sets, and discourse-related sets. Function sets and class sets may be added if there is sufficient support for them in the discourse or graphics. However, since there is support for class sets in the knowledge base, they may be created only when needed.

A *location set* is based on spatial relationships between icons in the drawing. These are comprised of graphical objects that are all within regions of the drawing that have either been explicitly identified or that have been inferred. Regions in the drawing of the Laboratory include the island itself, Starr Island, the region north of the ravine, the region south of the ravine, the area of cliffs, the area where poets and Impressionists "hung out", and the cluster of dorms near the laboratory building. Each of these regions would correspond to a location set containing all the graphical objects contained within that region.

[2] In such cases, the use of one kind of icon rather than the other may indicate that the user considers there to be two distinct sets of the same kind of object.

The system only creates location sets that have been marked by the user through discourse or graphics. For example, a location set is created for dorms H–J at Utterance 6 when the user indicates this set with the words "back here".

Spatial regions can nest and overlap; their corresponding location sets will have similar subset-superset and intersection relationships. For example, as shown in Fig. 6, it is possible to derive several location sets in the area near the three dorms. One is comprised of the three graphical objects representing the dorms and another, larger set consists of those objects and the graphical object representing the laboratory. Others include sets consisting of all graphical objects in the northern half of the island and of all graphical objects on the island. When the user points near the dorms and says "these", the system should consider all of these regions as possibilities for possible referents.

Fig. 6. Several different regions of the Shoals Marine Laboratory sketch.

Most of the work of location set creation and maintenance will be done using simple heuristics as new icons are added. This will speed the process of finding referents for plural references. However, there may be times when new location sets need to be created based on gestures or utterances by the user. For example, the user might use a sweeping gesture while saying "around here"; the system may interpret this as the definition of a new region, figure out what that region encompasses, then create a new location set to represent graphical objects within that area.

Location sets are represented as frames. Fig. 7 shows a representation of a location set that includes the three dorms by the laboratory. The representation of a location set includes the graphical objects that are members, the region that

corresponds to the set, any label supplied by the user, any inferences that led to filling other slots, and a list of related sets. The related sets in this case are other location sets that are spatially related to this one. For example, the location set shown would be related to the set containing the three dorms and the laboratory; both sets would be based on clustering icons, but the latter would be a larger cluster.

```
^locset1:
  isa: ^location-set
  members: (^dorm1 ^dorm2 ^dorm3)
  region: ^region1
  label: "dorms"
  inferences:
  related-sets:  (^locset2 ...)
```

Fig. 7. A location set containing the three dorms by the laboratory.

The second kind of set is the *icon set*. For now, an icon set is a set of graphical objects that have the same icon. The problems noted above of using the same icon for different objects, or different icons for the same object, are a subject for future work. We also leave for future work the problem of determining when icon subsets should be formed (e.g., when there should be separate sets for small squares and large squares as well as for all squares). The current representation of an icon set is consequently simple: just a slot for the icon class itself (e.g., "squares") and a slot containing the members of the set, that is, the graphical objects that were drawn using icons that match the icon class (e.g., graphical objects representing buildings, etc.). These icon sets are easy to create and maintain.

The third kind of set is the *discourse-related set*. This set is based on the discourse context in which the objects were introduced or discussed. For example, if the user talks about being attacked by a gull while walking along the path from Kiggins Commons to the laboratory, the corresponding discourse-related set would include the graphical objects mentioned. Later, if the user says "that bird attacked me around here" and gestures near the path, the discourse could be used to determine what "that bird" refers to, and the appropriate discourse-related set from the graphics context could be used to determine what "around here" means—in this context, somewhere along the path, and not just somewhere near where the user pointed.

The representation of discourse-related sets is simple: just the members of the set and a pointer to the discourse context that defined the set. Although in principle the information contained in discourse-related contexts could be generated from the discourse context, as we have mentioned, that is ephemeral, while the graphics context is not. It makes sense to have graphical objects that were grouped by the discourse remain grouped together for later anaphora resolution.

5 Conclusion

When discourse and graphics can be used for communication in a multi-modal interface, both must contribute to the context. In this paper, we have described our work on creating a graphics context. The graphics context differs from the discourse context because the graphics are visible throughout the interaction and because a much richer set of relationships between objects can be readily apprehended from the graphics. These relationships often become the membership criteria for sets that serve as the referents for plural anaphora. We have discussed the need for a graphics context that is separate from the discourse context and have shown a representation for the graphics context that will support finding referents for plural anaphora such as "these."

References

1. J. Allen. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, Inc., Reading, Mass., 1987.
2. 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.
3. R. Reichman. *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics (An ATN Model)*. The MIT Press, Cambridge, Mass, 1985.
4. B. J. Grosz and C. L. Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.
5. E. H. Turner, R. M. Turner, J. Phelps, C. Grunden, M. Neale, and J. Mailman. Aspects of context for understanding multi-modal communication. In *Lecture Notes in Artificial Intelligence 1688: Modeling and Using Context* (Proceedings of the 1999 International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-99), Trento, Italy). Springer–Verlag, 1999.