Context-based Multimodal Input Understanding in Conversational Systems

Joyce Chai, Shimei Pan, Michelle X. Zhou and Keith Houck
IBM T. J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532, USA
{jchai, shimei, mzhou, khouck}@us.ibm.com

Abstract

*In a multimodal human-machine conversation, user inputs are often abbreviated or imprecise. Sometimes, only fusing multimodal inputs together cannot derive a complete understanding. To address these inadequacies, we are building a semantics-based multimodal interpretation framework called MIND (**M**ultimodal **I**nterpretation for **N**atural **D**ialog). The unique feature of MIND is the use of a variety of contexts (e.g., domain context and conversation context) to enhance multimodal fusion. In this paper, we present a semantic rich modeling scheme and a context-based approach that enable MIND to gain a full understanding of user inputs, including those ambiguous and incomplete ones.*

1. Introduction

Multimodal interfaces allow human to interact with machines through multiple modalities such as speech, gesture, and gaze. Studies showed that these interfaces support a more effective human-computer interaction, for example, by reducing task completion time and task errors rate [11]. Inspired by the earlier work (e.g., [2, 4, 8, 13]), we are building an intelligent infrastructure, called Responsive Information Architect (RIA), which can engage users in a multimodal conversation. Currently, RIA is embodied in a testbed, called Real Hunter™, a real-estate application for helping users find residential properties.

Figure 1 shows RIA's main components. A user can interact with RIA using multiple input channels, such as speech and gesture. First, a multimodal interpreter exploits various contexts (e.g., conversation history) to produce an interpretation frame that captures the meanings of user inputs. Based on the interpretation frame, a conversation facilitator decides how RIA should act by generating a set of conversation acts (e.g., Describe information to the user). Upon receiving the conversation acts, a presentation broker sketches a presentation draft that expresses the outline of a multimedia presentation. Based on this draft, a language designer and a visual designer work together to author a multimedia blueprint that contains fully coordinated and detailed multimedia presentation. The blueprint is then sent to a producer to be realized. To support all components described above, an information server supplies various contextual information, including domain data (e.g., houses and cities for a real-estate application), a conversation history (e.g., detailed conversation exchanges between RIA and a user), a user model (e.g., user profiles), and an environment model (e.g., device capabilities).

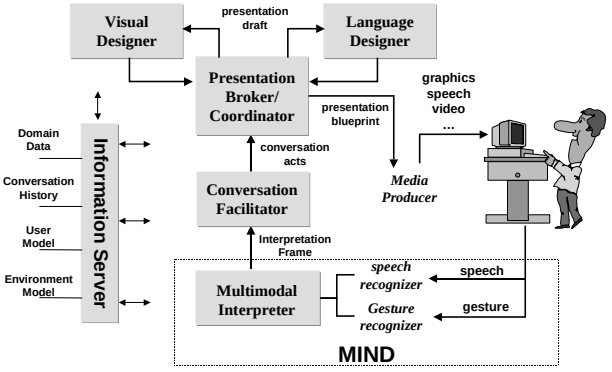

Figure 1. RIA Infrastructure

Our focus in this paper is on the interpretation of multimodal user inputs. Specifically, we are developing a semantics-based multimodal interpretation framework called MIND (**M**ultimodal **I**nterpreter for **N**atural **D**ialog). Most existing works on multimodal interpretation focus on interpreting user inputs through modality integration (e.g., merging speech with gesture) (e.g.,[2, 4, 8]) without considering interaction contexts (although they have been used extensively in spoken dialog systems [1, 14]). In a conversation setting, user inputs are often imprecise or abbreviated. Only integrating meanings from individual modalities together sometimes cannot reach a full understanding of those inputs. Therefore, MIND applies a context-based approach that uses a variety of contexts (e.g., domain context and conversation context) to enhance multimodal fusion.

Specifically, MIND supports three major processes: unimodal understanding, multimodal understanding, and discourse understanding (Figure 2). First, in unimodal understanding, an array of recognizers (e.g., a speech recognizer) convert input signals (e.g., speech signals) to modality-specific outputs (e.g., text). These outputs are then processed by modality-specific interpreters (e.g., a natural language interpreter). As a result, the meanings of each unimodal input are captured by a unimodal interpretation frame†. Based on these meanings, during the multimodal understanding process, a multimodal integrator

† We use IBM ViaVoice to perform speech recognition, and a statistics-based natural language understanding component [7] to process the natural language sentences. For gestures, we have developed a simple geometry-based gesture recognition and understanding component.

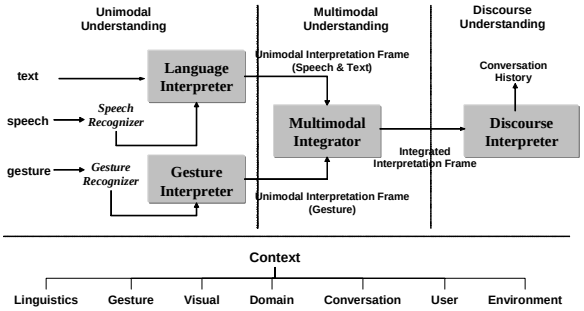

Figure 2. MIND components

uses proper contextual information to infer and create an integrated interpretation frame. This frame captures the overall meanings of the multimodal inputs. In addition to understanding each user input, MIND also captures the overall progress of a conversation and thus establishes a rich conversation context through discourse understanding. Based on earlier works on discourse interpretation ([10, 12]), MIND captures how a particular user input is related to the whole conversation, for example, whether the current input contributes to an existing conversation topic or it initiates a new one.

In this paper, we focus on the multimodal understanding process. In particular, we present two aspects of MIND. The first is a fine-grained semantic model that characterizes the meanings of user inputs and the overall conversation. The second is an integrated interpretation approach that identifies the semantics of user inputs using a wide variety of contexts (e.g., conversation history and domain knowledge). We first start with an example scenario to better explain the functions of MIND.

2. Example Scenario

Table 1 logs a conversation fragment between a user Joe and RIA. Joe initiates the conversation by asking for

Table 1. A conversation fragment.

Joe:	*Speech:* Show me houses in Irvington. (**U1**)
RIA:	*Speech:* Here are the houses you requested. *Graphics:* Show a collection of houses on the map (**R1**)
Joe:	*Speech:* What's the cost? *Gesture:* Point to the screen (**U2**)
RIA:	*Speech:* Which house are you interested in? *Graphics*: Highlight two candidate houses (**R2**)
Joe:	*Speech:* The red one (**U3**)
RIA:	*Speech:* The asking price of this red house is 350,000 dollars. *Graphics:* Highlight the red house and show the price (**R3**)
Joe:	*Speech:* And the size? (**U4**)
RIA:	*Speech:* The size of this house is 2000 square feet (**R4**)
Joe:	*Speech:* This one? *Gesture:* Put a question mark on top of a house icon (**U5**)
RIA:	*Speech:* The size of this house is 2200 square feet. (**R5**) *Graphics:* Highlight the house icon
Joe:	*Speech:* By the way, where is the train station? (**U6**)
RIA:	*Speech:* Here is the train station in Irvington. *Graphics:* Indicate the train station on the map (**R6**)
Joe:	*Speech:* OK.Then...how many bedrooms does it have? (**U7**)
RIA:	*Speech:* This house has four bedrooms.(**R7**)

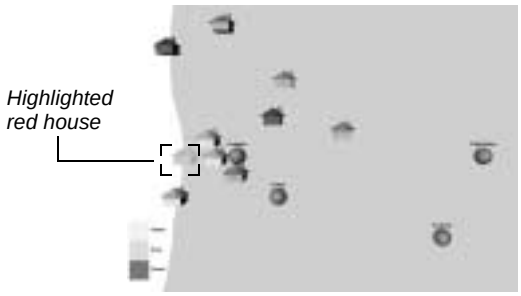

Figure 3. A snapshot of graphics output

houses in Irvington (U1), and RIA replies by showing a group of desired houses (R1). Based on the generated visual display, Joe points to the screen (a position between two houses) and asks for the price (U2). In this case, it is not clear which object Joe is pointing at. There are three candidates: two houses nearby and the town of Irvington†. Using our domain knowledge, MIND can rule out the town of Irvington, since Joe is asking for a price. However, MIND still can not determine which of the two house candidates is the desired one. To clarify this ambiguity, RIA highlights both houses and asks Joe to pinpoint the house of interest (R2).

Again, Joe's reply (U3) alone would be ambiguous, since there are multiple red objects on the screen. However, using the conversation history (R2) and the visual properties (Figure 3), MIND is able to infer that Joe is referring to the highlighted red house. Joe continues on to ask for the size (U4). This request by itself is incomplete, since Joe did not explicitly specify the object of interest (house). Nevertheless, MIND understands that Joe is asking for the size of the same red house based on the conversation history (U2–3). Joe moves on to inquire about another house (U5). This input by itself does not indicate exactly what Joe wants. Again, using the conversation history (U4), MIND recognizes that Joe is most likely asking for *the size* of another house. Next Joe switches to asking for the location of a train station (U6). According to our domain knowledge, train stations are always related to towns. Although Joe did not specify the town at this turn, MIND is able to conclude that the relevant town is Irvington using the conversation history (U1). Finally Joe asks about the number of bedrooms (U7). Based on the current visual context (one house still being highlighted from U5), MIND infers that Joe now returns to the previously explored house.

3. Semantics-based Modeling

To enable a full understanding of user multimodal inputs, we use a set of semantic features to model not only semantic aspects of user inputs at each turn of a conversation, but also the overall progress of the conversation.

3.1 Modeling User inputs

In support of multimodal conversation, MIND has

† The generated display has multiple layers, where the house icons are on top of the Irvington town map. Thus this deictic gesture could either refer to the town of Irvington or houses.

(a) Intention
Act: Request
Motivator: DataPresentation
Type: Describe

(b) Attention
Base: House
Topic: Instance
Focus: SpecificAspect (^Topic)
Aspect: Price
Constraint: < >
Content: [MLS0187652 | MLS0889234]

(c) Interpretation Status
SyntacticComplete: Attentional-ContentAmbiguity
SemanticComplete: TRUE

(d) Modality Decomposition
Modality: ^SpeechInput
Modality: ^GestureInput

(e) Presentation Preference
Directive: <Summary>
Media: <Multimedia>
Device: <Desktop>
Style: < >

Figure 4. The interpretation of a multimodal input U2[1]

1. Symbol ^ indicates a pointer and < > indicates no information conc ing this parameter has been identified from the user input.

two goals. First, MIND must understand the meanings of user inputs precisely so that the conversation facilitator (Figure 1) can decide how the system should act. Second, MIND needs to capture the user input styles (e.g., using a particular verbal expression or gesturing in a particular way) or user communicative preferences (e.g., preferring a verbal vs. a visual presentation). The captured information helps the multimedia generation components (visual or language designers in Figure 1) create more effective and tailored system responses. To accomplish both goals, MIND characterizes five aspects of a user input: intention, attention, interpretation status, presentation preference, and modality decomposition.

3.1.1. Intention. *Intention* describes the purpose of a user input [6]. We characterize three aspects of intention: Motivator, Act, and Type. Motivator captures the purpose of an interaction. Since we focus on information-seeking applications, MIND currently distinguishes three top-level purposes: DataPresentation, DataAnalysis (e.g., comparison), and ExceptionHandling (e.g., disambiguation). Act indicates one of the three user actions: request, reply, and inform. Request specifies that a user is making an information request (e.g., asking for a collection of houses in U1 Table 1). Reply indicates that the user is responding to a previous RIA request (e.g., confirming the house of interest in U3). Unlike Request or Reply, Inform states that a user is simply providing RIA with specific information, such as personal profiles or interests. Furthermore, MIND also distinguishes different types of Request. For example, one user may request RIA to Describe the desired information, such as the price of a house, while the other may request RIA simply to Identify the desired information (e.g., show a train station on the screen). Intention is modeled not only to support conversation, but also to facilitate multimedia generation. Specifically, Motivator and Type together direct RIA in its response generation. For example, RIA would consider Describe and Identify two different data presentation directives [15]. Figure 4(a) shows the Intention identified from the user input U2 (Table 1). It indicates that the user is asking RIA to present him with some information. The information to be presented is captured in Attention.

3.1.2. Attention. While Intention indicates the purpose of a user input, Attention captures the content of a user input with six features. Base specifies the semantic category of the content (e.g., all houses in our application belong to the House category). Topic indicates whether the user is concerned with a concept, a relation, an instance, or a collection of instances. For example, in U1 (Table 1) the user is interested in a collection of House, while in U2 he is interested in a specific instance. Focus further narrows down the scope of the content to distinguish whether the user is interested in a topic as a whole or just specific aspects of the topic. For example, in U2 the user focuses only on *one* specific aspect (price) of a house instance. Aspect enumerates the actual topical features that the user is interested in (e.g., the price in U2). Constraint holds the user constraints or preferences placed on the topic. For example, in U1 the user is only interested in the houses (Topic) located in Irvington (Constraint). Content points to the actual data in our database. Figure 4(b) shows the Attention identified for the user input U2. It states that the user is interested in the price of a house instance, MLS0187652 or MLS0889234 (house ids from the Multiple Listing Service). As discussed later, our fine-grained modeling of Attention provides MIND the ability to discern subtle changes in user interaction (e.g., a user may focus on one topic but explore different aspects of the topic). This in turn helps MIND assess the overall progress of a conversation.

3.1.3. Interpretation Status. InterpretationStatus provides an overall assessment on how well MIND understands an input. This information is particularly helpful in guiding RIA's next move. Currently, it includes two features. SyntacticCompleteness assesses whether there is any unknown or ambiguous information in the interpretation result. SemanticCompleteness indicates whether the interpretation result makes sense. Using the status, MIND can inform other RIA components whether a certain exception has risen. For example, SyntacticCompleteness in Figure 4c indicates that there is an ambiguity concerning Content in Attention, since MIND cannot determine whether the user is interested in MLS0187652 or MLS0889234. Based on this status, RIA would ask a clarification question to disambiguate the two houses (e.g., R2 in Table 1).

3.1.4. Presentation Preference. During a human-computer interaction, a user may indicate what type of responses she prefers. Currently, MIND captures user preferences along four dimensions. Directive specifies the high-level presentation goal (e.g., preferring a summary to details). Media indicates the preferred presentation medium (e.g., verbal vs. visual). Style describes what general formats should be used (e.g., using a chart vs. a diagram to illustrate information). Device states what devices would be used in the presentation (e.g., phone or PDA). Using the captured presentation preferences, RIA can generate multimedia presentations that are tailored to individual users and their goals. For example, Figure 4(e) records the user preferences from U2. Since the user did not explicitly specify any preferences, MIND uses the default values to represent those preferences. Presentation preferences can either directly derived from user inputs or inferred based on user and environment contexts.

3.1.5. Modality Decomposition. ModalityDecomposition (Figure 4d) maintains a reference to the interpretation

Gesture Input **(a)**	**(b)**	**(c)**	**(d)**	Speech Input **(e)**	**(f)**
Intention	**Attention (A1)**	**(A2)**	**(A3)**	**Intention**	**Attention**
Act: < > **Motivator**: < > **Type**: Refer **SurfaceAct**: Point **TimeInterval**: [..]	**Base**: House **Topic**: Instance **Focus**: < > **Aspect**: < > **Constraint**: < > **Content**: [MLS0187652]	**Base**: House **Topic**: Instance **Focus**:< > **Aspect**: < > **Constraint**: < > **Content**: [MLS0889234]	**Base**: City **Topic**: Instance **Focus**: < > **Aspect**: < > **Constraint**: < > **Content**: [Irvington]	**Act**: Request **Motivator**: DataPresentation **Type**: Describe **SurfaceAct**: Inquire **TimeInterval**: [..]	**Base**: < > **Topic**: Instance **Focus**: SpecificAspect(^Topic) **Aspect**: Price { <**SynCat**: Noun> <**Realization**: "cost">} **Constraint**: [ReferredBy THIS] **Content**: < >

Figure 5. Separate interpretation of two unimodal inputs in U2.

result for each unimodal input, such as the gesture input in Figure 5(a–d) and the speech input in Figure 5(e–f). In addition to the meanings of each unimodal input (Intention and Attention), MIND also captures modality-specific characteristics from the inputs such as time intervals during which the actions take place. In particular, MIND uses SurfaceAct to distinguish different types of gesture/speech acts. For example, there is an Inquire speech act (Figure 5e) and a Point gesture act (Figure 5a). Furthermore, MIND captures the syntactic form of a speech input, including the syntactic category (SynCat) and the actual language realization (Realization) of important concepts (e.g., Topic and Aspect). For example, Aspect price is realized using a noun cost (Figure 5f). Using such information, RIA can adapt itself to user input styles (e.g., using similar vocabulary).

3.2 Discourse-level Modeling.

In addition to modeling user inputs at each conversation turn, we also model the entire progress of a conversation to provide a rich conversation context based on Grosz and Synder's conversation theory [1986].

3.2.1. Conversation Unit and Segment. Our conversation history has two main elements: conversation units and conversation segments. A *conversation unit* records user or RIA actions at a single turn of a conversation. These units can be grouped together to form a *segment* (e.g., based on their intentions and sub-intentions). Figure 6 depicts the hierarchical conversation history that outlines the first eight turns of the conversation in Table 1. This structure contains eight units (rectangles U1–4 for the user, R1–4 for RIA) and three segments (ovals DS1–3)

Specifically, a user conversation unit contains the interpretation result of a user input discussed in the last section. A RIA unit contains the automatically generated multimedia response, including the semantic and syntactic structures of a multimedia presentation [15]. A segment has five features: Intention, Attention, Initiator, Addressee, and State. The Intention and Attention are similar to those modeled in the turns (see DS1, U1 and R1 in Figure 6). In addition, Initiator indicates the conversation initiating participant (e.g., Initiator is User in DS1). Addressee indicates the recipient of the conversation (e.g., Addressee is RIA in DS1). Finally, State reflects the current state of a segment: active, accomplished or suspended. For example, after U3 DS1 is still active, but DS3 is already accomplished since its purpose of disambiguating the content has been fulfilled.

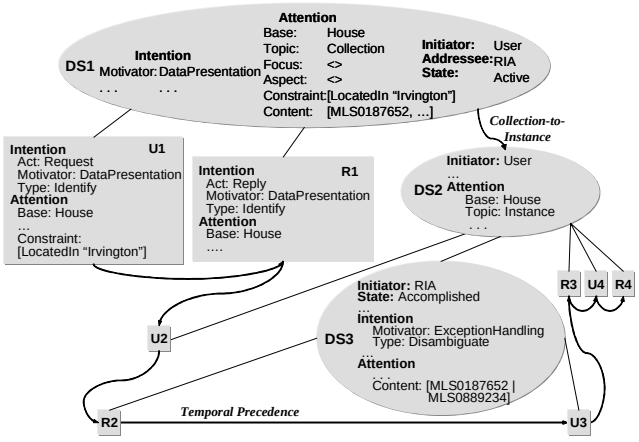

Figure 6. Fragment of a discourse structure

3.2.2. Discourse Relations. To model the progress in a conversation, MIND captures two types of relations in the discourse: structural relations and transitional relations. Structural relations reveal the intention/sub-intention structure between the purposes of conversation segments. For example, in Figure 6, DS3 is a sub-intention of DS2, since ExceptionHandling (Motivator of DS3) is for the purpose of DataPresentation (Motivator of DS2) of a particular house. Transitional relations specify transitions between conversation segments and between conversation units as the conversation unfolds. Currently, two types of relations are identified between segments: intention switch and attention switch. The attention switch is further categorized by eight types of data transitional relations such as Collection-to-Instance and Instance-to-Aspect. For example, the attention is switched from a collection of houses in DS1 to a specific house in DS2 (Figure 6) through Collection-to-Instance. Data transitional relations allow MIND to capture user data exploration patterns. Such patterns in turn can help RIA decide potential data navigation paths and provide users with an efficient information-seeking environment. In addition to segment relations, there is also a temporal-precedence relation between conversation units that preserves the sequence of conversation.

4. Context-based Multimodal Understanding

Based on the semantic model described above, MIND uses a wide variety of contexts to interpret the rich semantics of user inputs. In a conversation setting, users often give partial information at a particular turn. Traditional multimodal understanding that focuses on multimodal integration is often inadequate to achieve a full understanding of those inputs. For example, in U5 (Table 1) it is not clear what exactly the user wants by just merging the two inputs together. To address these inadequacies, MIND adds context-based inference on top of multimodal fusion.

Our approach allows MIND to use rich contextual information to infer the unspecified information (e.g., the exact intention in U5) and resolve ambiguities from the user input (e.g., the gestural ambiguities in U2). In particular, MIND applies two operations: fusion and inference to achieve multimodal understanding.

4.1 Fusion

Fusion creates an integrated representation by combining multiple unimodal inputs. In this process, MIND first merges intention structures using a set of rules. For example, one rule asserts that when combining two intentions together, if one is only for referral purpose (e.g., the gesture of U2 in Figure 5a), then the other (e.g., the speech of U2 in Figure 5e) serves as the combined intention (e.g., the integrated Intention of U2 in Figure 4a). The rational behind this rule is that a referral action without any overall purpose most likely complements another action that carries a main communicative intention. Thus, this communicative intention is the intention after fusion. Once intentions are merged, MIND uses unification to merge the corresponding attention structures. For example, in U2 MIND produces two combined attention structures by unifying the Attention from the speech (Figure 5f) with each Attention from the gesture (Figure 5b-d). The result of fusion is shown in Figure 7. In this combined representation, there is an ambiguity about which of the two attention structures is the true interpretation (Figure 7b, c). Furthermore, within the attention structure for House, there is an additional ambiguity on the exact object (Content in Figure 7b). This example shows that integration resulting from unification based multimodal fusion is not adequate to resolve ambiguities. We will show later that some ambiguities can be resolved based on contexts.

For simple user inputs, attention fusion is straightforward. However, it may become complicated when multiple attentions from one input need to be unified with multiple attentions from another input. To fuse these inputs, MIND first applies temporal constraints to align the attentions identified from each modality. This alignment can be easily performed when there is an overlapping or a clear temporal binding between a gesture and a particular phrase in the speech. However, in a situation where a gesture is followed (preceded) by a phrase without an obvious temporal association as in "tell me more about the red house (deictic gesture 1) this house (deictic gesture 2) the blue house," MIND uses contexts to determine which two of the three objects (the red house, this house, and the blue house) mentioned in the speech should be unified with the attentions from the gesture.

Modality integration in most existing multimodal systems is speech driven and relies on the assumption that speech always carries the main act, and others are complementary [2, 3]. In contrast, our modality integration is based on the semantic contents of inputs rather than their forms of modalities. Thus, as Quickset [8], MIND supports all modalities equally. For example, the gesture input in U5 is the main act, while the speech input is the complementary act for reference.

(a)	(b)	(c)
Intention	**Attention**	**Attention**
Motivator: DataPresentation **Act**: Request **Type**: Describe	**Base**: House **Topic**: Instance **Focus**: SpecificAspect **Aspect**: Price **Content**: [MLS0187652 | MLS0889234]	**Base**: City **Topic**: Instance **Focus**: SpecificAspect **Aspect**: Price **Content**: [Irvington]

Figure 7. Combined interpretation as a result of multimodal fusion in U2.

4.2 Inference

Inference identifies user unspecified information and resolves input ambiguities using contexts. In a conversation, users often supply abbreviated or imprecise inputs at a particular turn, e.g., abbreviated inputs given in U3, U4, U5, and the imprecise gesture input in U2 (Table 1). Moreover, the abbreviated inputs often foster ambiguities in interpretation. To derive a thorough understanding from the partial user inputs and resolve ambiguities, MIND exploits various contexts.

The domain context provides domain knowledge and is particularly useful in resolving input ambiguities. For example, fusion inputs in U2 which has imprecise gesture results in ambiguities (Figure 7). To resolve the ambiguity whether the attention is a city object or a house object, MIND uses the domain context. In this case, MIND eliminates the city candidate, since cities cannot have an attribute of price. As a result, MIND understands that the user is asking about the House.

In addition to the domain context, the conversation context also provides MIND with a useful context to derive the information not specified in the user inputs. In an information seeking environment, users tend to only explicitly or implicitly specify the new or changed aspects of their information of interest without repeating those that have been mentioned earlier in the conversation. Therefore, some required but unspecified information in a particular user input can be inferred from the conversation context. For example, the user did not explicitly specify the object of interest in U4 since he has provided such information in U3. However, MIND uses the conversation context and infers that the missing object in U4 is the house mentioned in U3. In another example U5, the user specified another house but did not mention the interested aspect of this new house. Again, based on the conversation context, MIND recognizes that the user is interested in the size aspect of the new house.

RIA's conversation history is inherently a complex structure with fine-grained information (e.g., Figure 6). However, with our hierarchical structure of conversation units and segments, MIND is able to traverse the conversation history efficiently. In our example scenario, the conversation between U1 and R5 contributes to exploring houses in Irvington. U6 starts a new segment, in which the user asked for the location of a train station, but did not specify the relevant town name. However, MIND is able to infer that the relevant town is Irvington directly from DS1, since DS1 captures the town name Irvington. Without the segment structure, MIND would have to traverse all previous 10 turns to resolve the town reference.

As RIA provides a rich visual environment for users to interact with, users may refer to objects on the screen by their spatial (e.g., the house at the left corner) or percep-

tual attributes (e.g., the red house). To resolve these spatial/perceptual references, MIND exploits the visual context, which provides the detailed semantic and syntactic structures of visual objects and their relations. More specifically, visual encoding automatically generated for each object is maintained as a part of the system conversation unit in the conversation history. During reference resolution, MIND would identify potential candidates by mapping the referring expressions with the internal visual representation. For example, the object which is highlighted on the screen (R5) has an internal representation that associates the visual property Highlight with the object identifier. This allows MIND to correctly resolve referents for *it* in U7. In this reference resolution process, based on the Centering Theory [5], MIND first identifies the referent most likely to be the train station since it is the preferred center in the previous utterance. However, according to the domain knowledge, such a referent is ruled out since the train station does not have the attribute of bedrooms. Nevertheless, based on the visual context, MIND recognizes a highlighted house on the screen. An earlier study indicates that objects in the visual focus are often referred by pronouns, rather than by full noun phrases or deictic gestures [9]. Therefore, MIND considers the object in the visual focus (i.e., the highlighted house) as a potential referent. In this case, since the highlighted house is the only candidate that satisfies the domain constraint, MIND resolves the pronoun *it* in U7 to be that house. Without the visual context, the referent in U7 would not be resolved.

5. Implementation and Evaluation

We have developed MIND as a research prototype. The modeling scheme and the context-based interpretation approach are implemented in Java. The prototype is currently running on Linux.

Our initial semantic models and interpretation algorithms were driven by a user study we conducted. In this study, one of our colleagues acted as RIA and interacted with users to help them find real estate in Westchester county. The analysis of the content and the flow of the interaction indicates that our semantic models and interpretation approaches are adequate to support these interactions. After MIND was implemented, we conducted a series of testing on multimodal fusion and context-based inference (focusing on domain and conversation contexts). Half of the trials were specifically designed to contain ambiguous or abbreviated inputs. Since the focus of the testing was not on our language model, we designed the speech inputs so that they could be parsed successfully by our language understanding components. The testing showed that once the user speech input was correctly recognized and parsed, 90% of trials were correctly interpreted based on our multimodal interpretation approach. However, speech recognition is a bottleneck in MIND. To improve the robustness of MIND, we need to enhance the accuracy of speech recognition and improve the coverage of the language model. We plan to do more vigorous evaluations in the future.

6. Conclusions and Future Work

In a multimodal conversation, user inputs could be ambiguous or abbreviated. Only fusing multimodal inputs together sometimes cannot reach a full understanding. Therefore, we have built a context based multimodal interpreter MIND that applies rich contexts to enhance multimodal fusion. In particular, MIND has two unique features. The first is a fine-grained semantic model that characterizes the meanings of user inputs and the overall conversation from multiple dimensions. The second is an integrated interpretation approach that identifies the semantics of user inputs using a wide variety of contexts. These features enable MIND to achieve a deep understanding of user inputs. Our future work includes exploring learning techniques to incorporate confidence factors to further enhance input interpretation.

7. References

[1] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, G. L., and A. Stent. Toward conversational human computer interaction. *AI Magazine*, 22(4):27–37, 2001.

[2] R. A. Bolt. Voice and gesture at the graphics interface. *Computer Graphics*, pages 262–270, 1980.

[3] J. Burger and R. Marshall. The application of natural language models to intelligent multimedia. In M. Maybury, editor, *Intelligent Multimedia Interfaces*, pages 429–440. MIT Press, 1993.

[4] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: Multimodal interaction for distributed applications. *Proc. ACM MM'96*, pages 31–40, 1996.

[5] B. J. Grosz, A. K. Joshi, and S. Weinstein. Towards a computational theory of discourse interpretation. *Computational Linguistics*, 21(2):203–225, 1995.

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

[7] F. Jelinek, J. Lafferty, D. M. Magerman, R. Mercer, and S. Roukos. Decision tree parsing using a hidden derivation model. *Proc. Darpa Speech and Natural Language Workshop*, March 1994.

[8] M. Johnsonton. Unification-based multimodal parsing. *Proc. COLING-ACL'98*, 1998.

[9] A. Kehler. Cognitive status and form of reference in multimodal human-computer interaction. *Proc. AAAI'01*, pages 685–689, 2000.

[10] K. Lochbaum. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572, 1998.

[11] S. Oviatt. Multimodal interfaces for dynamic interactive maps. *Proc. CHI'96*, 1996.

[12] C. Rich and C. Sidner. Collagen: A collaboration manager for software interface agents. *User Modeling and User-Adapted Interaction*, 1998.

[13] W. Wahlster. User and discourse models for multimodal communication. In M. Maybury and W. Wahlster, editors, *Intelligent User Interfaces*, pages 359–370. 1998.

[14] W. Wahlster. Mobile speech-to-speech translation of spontaneous dialogs: An overview of the final verbmobil system. *Verbmobile*, pages 3–21, 2000.

[15] M. X. Zhou and S. Pan. Automated authoring of coherent multimedia discourse for conversation systems. *Proc. ACM MM'01*, pages 555–559, 2001