Semantics-based Representation for Multimodal Interpretation in Conversational Systems

Joyce Chai
IBM T. J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532, USA
{jchai@us.ibm.com}

Abstract

To support context-based multimodal interpretation in conversational systems, we have developed a semantics-based representation to capture salient information from user inputs and the overall conversation. In particular, we present three unique characteristics: fine-grained semantic models, flexible composition of feature structures, and consistent representation at multiple levels. This representation allows our system to use rich contexts to resolve ambiguities, infer unspecified information, and improve multimodal alignment. As a result, our system is able to enhance understanding of multimodal inputs including those abbreviated, imprecise, or complex ones.

1 Introduction

Inspired by earlier works on multimodal interfaces (e.g., Bolt, 1980; Cohen el al., 1996; Wahlster, 1991; Zancanaro et al., 1997), we are currently building an intelligent infrastructure, called Responsive Information Architect (RIA) to aid users in their information-seeking process. Specifically, RIA engages users in a full-fledged multimodal conversation, where users can interact with RIA through multiple modalities (speech, text, and gesture), and RIA can act/react through automated multimedia generation (speech and graphics) (Zhou and Pan 2001). Currently, RIA is embodied in a testbed, called Real Hunter™, a real-estate application to help users find residential properties.

As a part of this effort, we are building a semantics-based multimodal interpretation framework MIND (**M**ultimodal **I**nterpretation for **N**atural **Di**alog) to identify meanings of user multimodal inputs. Traditional multimodal interpretation has been focused on integrating multimodal inputs together with limited consideration on the interaction context. In a conversation setting, user inputs could be abbreviated or imprecise. Only by combining multiple inputs together often cannot reach a full understanding. Therefore, MIND applies rich contexts (e.g., conversation context and domain context) to enhance multimodal interpretation. In support of this context-based approach, we have designed a semantics-based representation to capture salient information from user inputs and the overall conversation.

In this paper, we will first give a brief overview on multimodal interpretation in MIND. Then we will present our semantics-based representation and discuss its characteristics. Finally, we will describe the use of this representation in context-based multimodal interpretation and demonstrate that, with this representation, MIND is able to process a variety of user inputs including those ambiguous, abbreviated and complex ones.

2 Multimodal Interpretation

To interpret user multimodal inputs, MIND takes three major processes as in Figure 1: unimodal understanding, multimodal understanding, and discourse understanding. During unimodal understanding, MIND applies modality specific recognition and understanding components (e.g., a speech recognizer and a language interpreter) to identify meanings from each unimodal input, and captures those meanings in a representation called *modality unit*. During multimodal understanding, MIND combines semantic meanings of unimodal inputs (i.e., modality units), and uses contexts (e.g., conversation context and domain context) to form an overall understanding of user multimodal inputs. Such an overall understanding is then captured in a representation called *conversation unit*. Furthermore, MIND also identifies how an input relates to the overall conversation discourse through discourse understanding. In particular, MIND uses a representation called *conversation segment* to group together inputs that contribute to a same goal or sub-goal (Grosz and Sidner, 1986). The result of discourse understanding is an evolving conversation history that reflects the overall progress of a conversation.

Figure 2 shows a conversation fragment between a user and MIND. In the first user input U1, the deictic

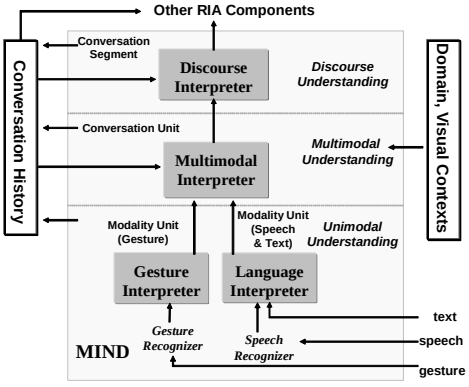


Figure 1. MIND components

	A collection of houses are shown on the map of Irvington
U1:	*Speech*: How much is this? *Gesture*: Point to the screen (not directly on any object)
R1:	*Speech*: Which house are you interested in? *Graphics*: Highlight two house icons
U2:	*Speech*: The green one.
R2:	*Speech*: The green house costs 250,000 dollars.
U3:	*Speech*: What about this one? *Gesture*: Point to a house icon on the screen
R3:	*Speech*: This house costs 320,000 dollars. *Graphics*: Highlight the house icon and show a picture
U4:	*Speech*: Show me houses with this style around here *Gesture*: Point to a position east of Irvington on the map
R4:	*Speech*: This is a Victorian style house. I find seven Victorian houses in White Plains. *Graphics*: Show seven houses in White Plains
U5:	*Speech*: Compare these two houses with the previous house. *Graphics*: Point to the corner of the screen where two house icons are displayed
R5:	*Speech*: Here is the comparison chart. *Graphics*: Show a chart

Figure 2. A conversation fragment

gesture (shown in Figure 3) is ambiguous. It is not clear which object the user is pointing at: two houses nearby or the town of Irvington[1]. The third user input U3 by itself is incomplete since the purpose of the input is not specified. Furthermore, in U4, a single deictic gesture overlaps (in terms of time) with both "this style" and "here" from the speech input, it is hard to determine which one of those two references should be aligned and fused with the gesture. Finally, U5 is also complex since multiple objects ("these two houses") specified in the speech input need to be unified with a single deictic gesture.

This example shows that user multimodal inputs exhibit a wide range of varieties. They could be abbreviated, ambiguous or complex. Fusing inputs together often cannot reach a full understanding. To process these inputs, contexts are important.

3 Semantics-based Representation

To support context-based multimodal interpretation, both representation of user inputs and representation of contexts are crucial. Currently, MIND uses three types of contexts: domain context, conversation context, and visual context. The domain context provides domain knowledge. The conversation context reflects the progress of the overall conversation. The visual context gives the detailed semantic and syntactic structures of visual objects and their relations. In this paper, we focus on representing user inputs and the conversation context. In particular, we discuss two aspects of representation: semantic models that capture salient information and structures that represent those semantic models.

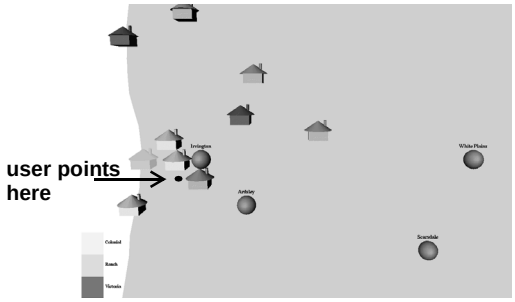


Figure 3. An example of graphics output

1 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.

3.1 Semantic Models

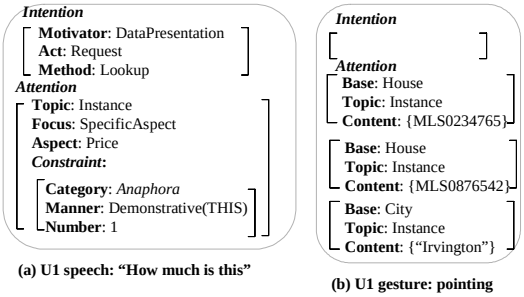
When two people participate in a conversation, their understanding of each other's purposes forms strong constraints on how the conversation is going to proceed. Especially, in a conversation centered around information seeking, understanding each other's information needs is crucial. Information needs can be characterized by two main aspects: motivation for seeking the information of interest and the information sought itself. Thus, MIND uses an intention model to capture the first aspect and an attention model to capture the second. Furthermore, since users can use different ways to specify their information of interest, MIND also uses a constraint model to capture different types of constraints that are important for information seeking.

3.1.1 Intention and Attention

Intention describes the purpose of a message. In an information seeking environment, intention indicates the motivation or task related to the information of interest. An intention is modeled by three dimensions: Motivator indicating one of the three high level purposes: DataPresentation, DataAnalysis (e.g., comparison), and ExceptionHandling (e.g., clarification), Act specifying whether the input is a request or a reply, and Method indicating a specific task, e.g., Search (activating the relevant objects based on some criteria) or Lookup (evaluating/retrieving attributes of objects).

Attention relates to objects, relations that are salient at each point of a conversation. In an information seeking environment, it relates to the information sought. An attention model is characterized by six dimensions. Base indicates the semantic type of the information of interest (e.g., House, School, or City which are defined in our domain ontology). Topic specifies the granularity of the information of interest (e.g., Instance or Collection). Focus identifies the scope of the topic as to whether it is about a particular feature (i.e., SpecficAspect) or about all main features (i.e., MainAspect). Aspect provides specific features of the topic. Constraint describes constraints to be satisfied (described later). Content points to the actual data. The intention and attention models were derived based on preliminary studies of user information needs in seeking for residential properties. The details are described in (Chai et al., 2002).

For example, Figure 4(a-b) shows the Intention and Attention identified from U1 speech and gesture input respectively. Intention in Figure 4(a) indicates the user is requesting RIA (Act: Request) to present her some data (Motivator: DataPresentation) about attributes of



(a) U1 speech: "How much is this" **(b) U1 gesture: pointing**

Figure 4. Intention and Attention for U1 unimodal inputs

certain object(s) (Method: Lookup). The Attention indicates that the information of interest is about the price (Aspect: Price) of a certain object (Focus: Instance). The exact object is not known but is referred by a demonstrative "this" (in Constraint). Intention in Figure 4(b) does not have any information since the high level purpose and the specific task cannot be identified from the gesture input. Furthermore, because of the ambiguity of the deictic gesture, three Attentions are identified. The first two Attentions are about house instances MLS0234765 and MLS0876542 (ID from Multiple Listing Service) and the third is about the town of Irvington.

3.1.2 Constraints

In an information seeking environment, based on the conversation context and the graphic display, users can refer to objects using different types of references, for example, through temporal or spatial relations, visual cues, or simply a deictic gesture. Furthermore, users can also search for objects using different constraints on data properties. Therefore, MIND models two major types of constraints: reference constraints and data constraints. Reference constraints characterize different types of references. Data constraints specify relations of data properties. A summary of our constraint model is shown in Figure 5. Both reference constraints and data constraints are characterized by six dimensions. Category sub-categorizes constraints (described later). Manner indicates the specific way such a constraint is expressed. Aspect indicates a feature (features) this constraint is concerned about. Relation specifies the relation to be satisfied between the object of interest and other objects or values. Anchor provides a particular value, object or a reference point this constraint relates to. Number specifies cardinal numbers that are associated with the constraint.

Reference Constraints

Reference constraints are further categorized into four categories: Anaphora, Temporal, Visual, and Spatial. An anaphora reference can be expressed through pronouns such as "it" or "them" (Pronoun), demonstratives such as "this" or "these" (Demonstrative), here or there (Here/There), or proper names such as "Lynhurst" (ProperNoun). An example is shown in Figure 4(a), where a demonstrative "this" (Manner: Demonstrative-This) is used in the utterance "this house" to refer to a single house object (Number: 1). Note that

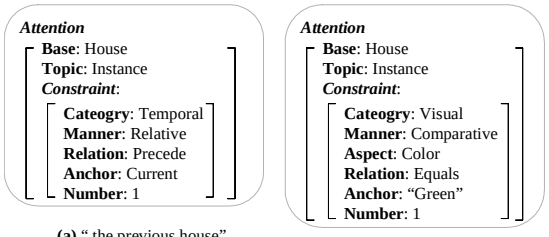
	Category	Manner	Aspect	Relation	Anchor	Number
Reference Constraints	Anaphora	Demonstrative, Pronoun, Here/There, ProperNoun,	-	-	-	Multiple, Cardinal-number (e.g., 1, 2)
	Temporal	Relative, Absolute	-	Procede, Succeed, Ordinal (e.g., first)	Current, Object	
	Spatial	Relative, Absolute	-	Orientation (e.g., Left, Right)	DisplayFrame, FocusFrame, Object	
	Visual	Comparative	Visual-Properties (e.g., Color, Highlight)	Equals	DataValue, ValueOfObject, Object	
Data Constraints	Attributive	Comparative, Superlative, Fuzzy	Data Features (e.g., Price, Size)	Less-Than, Equals, Greater-Than	DataValue, ValueOfObject, Object	

Figure 5. Constraint model

Manner also keeps track of the specific type of the term. The subtle difference between terms can provide additional cues for resolving references. For example, the different use of "this" and "that" may indicate the recency of the referent in the user mental model of the discourse, or the closeness of the referent to the user's visual focus.

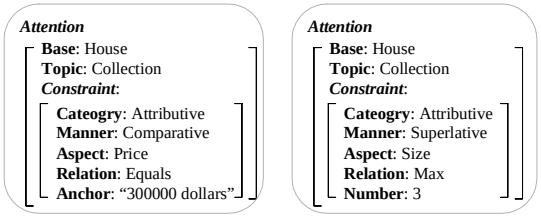
Temporal references use temporal relations to refer to entities that occurred in the prior conversation. Manner is characterized by Relative and Absolute. Relative indicates a temporal relation with respect to a certain point in a conversation, and Absolute specifies a temporal relation regarding to the whole interaction. Relation indicates the temporal relations (e.g., Precede or Succeed) or ordinal relations (e.g., first). Anchor indicates a reference point. For example, as in Figure 6(a), a Relative temporal constraint is used since "the previous house" refers to the house that precedes the current focus (Anchor: Current) in the conversation history. On the other hand, in the input: "the first house you showed me," an Absolute temporal constraint is used since the user is interested in the first house shown to her at the beginning of the entire conversation.

Spatial references describe entities on the graphic display in terms of their spatial relations. Manner is again characterized by Absolute and Relative. Absolute indicates that entities are specified through orientations (e.g., left or right, captured by Relation) with respect to the whole display screen (Anchor: DisplayFrame). In contrast, Relative specifies that entities are described through orientations with respect to a particular sub-frame (Anchor: FocusFrame, e.g., an area



(a) " the previous house" **(b)** " the green house"

Figure 6. Temporal and visual reference constraints



(a) "houses under 300,000 dollars" **(a)** "3 largest houses"

Figure 7. Attributive data constraints

with highlighted objects) or another object.

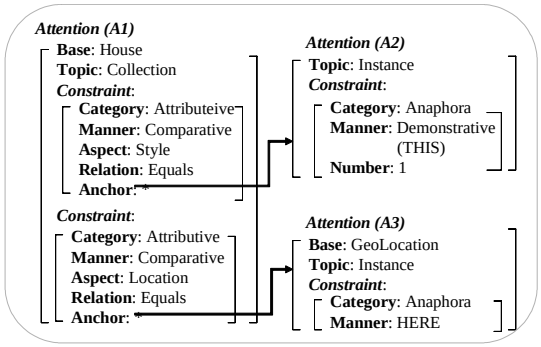
Visual references describe entities on the graphic output using visual properties (such as displaying colors or shapes) or visual techniques (such as highlight). Manner of Comparative indicates a visual entity is compared with another value (captured by Anchor). Aspect indicates the visual entity used (such as Color and Shape, which are defined in our domain ontology). Relation specifies the relation to be satisfied between the visual entity and some value. For example, constraint used in the input "the green house" is shown in Figure 6(b). It is worth mentioning that during reference resolution, the color Green will be further mapped to the internal color encoding used by graphics generation.

Data Constraints

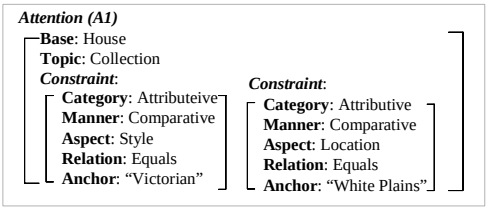
Data constraints describe objects in terms of their actual data attributes (Category: Attributive). The Manner of Comparative indicates the constraint is about a comparative relation between (aspects of) the desired entities with other entities or values. Superlative indicates the constraint is about minimum or maximum requirement(s) for particular attribute(s). Fuzzy indicates a fuzzy description on the attributes (e.g., "cheap house"). For example, for the input "houses under 300,000 dollars" in Figure 7(a), Manner is Comparative since the constraint is about a "less than" relationship (Relation: Less-Than) between the price (Aspect: Price) of the desired object(s) and a particular value (Anchor: "300000 dollars"). For the input "3 largest houses" in Figure 7(b), Manner is Superlative since it is about the maximum (Relation: Max) requirement on the size of the houses (Aspect: Size).

The refined characterization of different constraints provides rich cues for MIND to identify objects of interest. In an information seeking environment, the objects sought can come from different sources. They could be entities that have been described earlier in the conversation, entities that are visible on the display, or entities that have never been mentioned or seen but exist in a database. Thus, fine-grained constraints allow MIND to determine where and how to find the information of interest. For example, temporal constraints help MIND navigate the conversation history by providing guidance on where to start, which direction to follow in the conversation history, and how many to look for.

Our fine-grained semantic models of intention, attention and constraints characterize user information needs and therefore enable the system to come up with an intelligent response. Furthermore, these models are domain independent and can be applied to any information seeking applications (for structured information).



(a) Attention structure in the modality unit for U4 speech input



(b) Attention structure in the conversation unit for U4 speech input

Figure 8. Attention structures for U4

3.1.3 Representing User Inputs

Given the semantic models of intention, attention and constraints, MIND represents those models using a combination of feature structures (Carpenter, 1992). This representation is inspired by the earlier works (Johnston et al., 1997; Johnston, 1998) and offers a flexibility to accommodate complex inputs. Specifically, MIND represents intention, attention and constraints identified from user inputs as a result of both unimodal understanding and multimodal understanding.

During unimodal understanding, MIND applies a decision tree based semantic parser on natural language inputs (Jelinek et al., 1994) to identify salient information. For the gesture input, MIND applies a simple geometry-based recognizer. As a result, information from each unimodal input is represented in a modality unit. We have seen several modality units (in Figure 4, Figure 6, and Figure 7), where intention, attention and constraints are represented in feature structures. Note that only features that can be instantiated by information from the user input are included in the feature structure. For example, since the exact object cannot be identified from U1 speech input, the Content feature is not included in its Attention structure (Figure 4a). In addition to intention, attention and constraints, a modality unit also keeps a time stamp that indicates when a particular input takes place. This time information is used for multimodal alignment which we do not discuss here.

Depending on the complexity of user inputs, the representation can be composed by a flexible combi-

nation of different feature structures. Specifically, an attention structure may have a constraint structure as its feature, and on the other hand, a constraint structure may also include another attention structure.

For example, U4 in Figure 2 is a complex input, where the speech input "what about houses with this style around here" consists of multiple objects with different relations. The modality unit created for U4 speech input is shown in Figure 8(a). The Attention feature structure (A1) contains two attributive constraints indicating that the objects of interest are a collection of houses that satisfy two attributive constraints. The first constraint is about the style (Aspect: Style), and the second is about the location. Both of these constraints are related to other objects (Manner: Comparative), which are represented by Attention structures A2 and A3 through Anchor respectively. A2 indicates an unknown object that is referred by a Demonstrative reference constraint (this style), and A3 indicates a geographic location object referred by HERE. Since these two references are overlapped with a single deictic gesture, it is hard to decide which one should be unified with the gesture input. We will show in Section 4.3 that the fine-grained representation in Figure 8(a) allows MIND to use contexts to resolve these two references and improve alignment.

During multimodal understanding, MIND combines information from modality units together and generates a conversation unit that represents the overall meaning of user multimodal inputs. A conversation unit also has the same type of intention and attention feature structures, as well as the feature structure for data constraints. Since references are resolved during the multimodal understanding process, the reference constraints are no longer present in conversation units. For example, once two references in Figure 8(a) are resolved during multimodal understanding (details are described in Section 4.3), and MIND identifies "this style" is "Victorian" and "here" is "White Plains", it creates a conversation unit representing the overall meanings of this input in Figure 8(b).

3.2 Representing Conversation Context

MIND uses a conversation history to represent the conversation context based on the goals or sub-goals of user inputs and RIA outputs. For example, in the conversation fragment mentioned earlier (Figure 2), the first user input (U1) initiates a goal of looking up the price of a particular house. Due to the ambiguous gesture input, in the next turn, RIA (R2) initiates a sub-goal of disambiguating the house of interest. This sub-goal contributes to the goal initiated by U1. Once the user replies with the house of interest (U2), the sub-goal is fulfilled. Then RIA gives the price information (R2), and the goal initiated by U1 is accompolished. To reflect this progress, our conversation history is a hierarchical structure which consists of conversation segments and conversation units (in Figure 9). As mentioned earlier, a conversation unit records user (rectangle U1, U2) or RIA (rectangle R1, R2) overall meanings at a single turn in the conversation. These units can be grouped together to form a

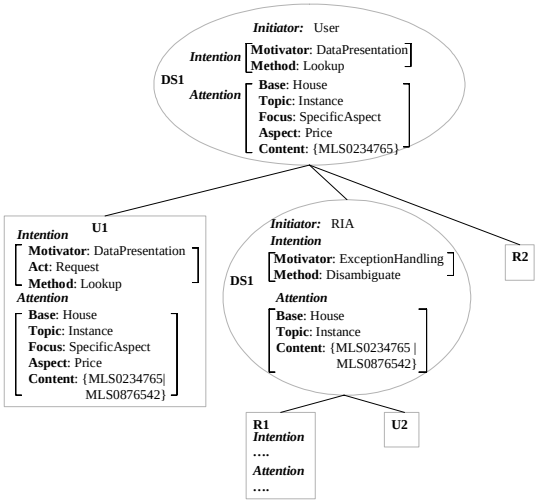


Figure 9. A fragment of a conversation history

conversation segment (oval DS1, DS2) based on their goals and sub-goals. Furthermore, a conversation segment contains not only intention and attention, but also other information such as the conversation initiating participant (Initiator). In addition to conversation segments and conversation units, a conversation history also maintains different relations between segments and between units. Details can be found in (Chai et al., 2002).

Another main characteristic of our representation is the consistent representation of intention and attention across different levels. Just like modality units and conversation units, conversation segments also consist of the same type of intention and attention feature structures (as shown in Figure 9). This consistent representation not only supports unification based multimodal fusion, but also enables context-based inference to enhance interpretation (described later).

We have described our semantics-based representation and presented three characteristics: fine-grained semantic models, flexible composition, and consistent representation. Next we will show that how this representation is used effectively in the multimodal interpretation process.

4 The Use of Representation in Multimodal Interpretation

As mentioned earlier, multimodal interpretation in MIND consists of three processes: unimodal understanding, multimodal understanding and discourse understanding. Here we focus on multimodal understanding. The key difference between MIND and earlier works is the use of rich contexts to improve understanding. Specifically, multimodal understanding consists of two sub-processes: multimodal fusion and context-based inference. Multimodal fusion fuses intention and attention structures (from modality units) for unimodal inputs and forms a combined representation. Context-based inference uses rich con-

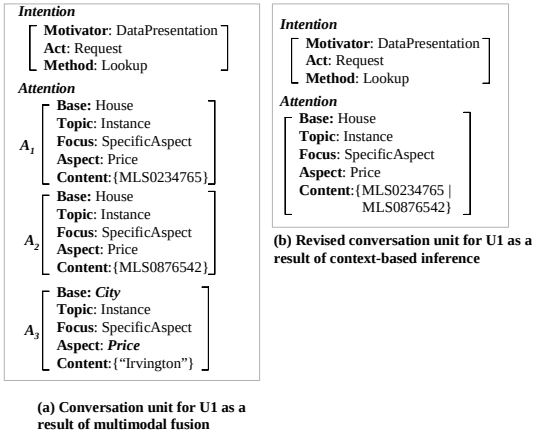


Figure 10. Resolving ambiguity for U1

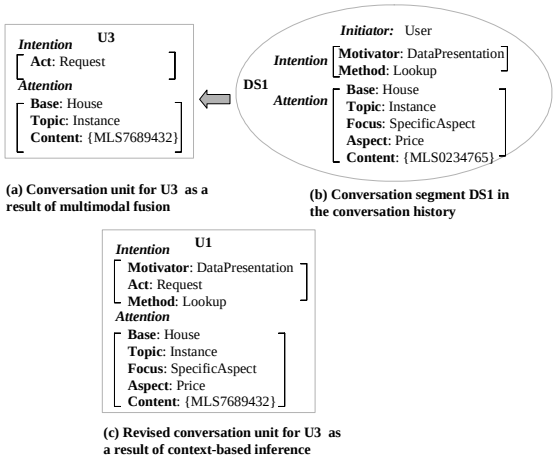


Figure 11. Deriving unspecified information for U3

texts to improve interpretation by resolving ambiguities, deriving unspecified information, and improving alignment.

4.1 Resolving Ambiguities

User inputs could be ambiguous. For example, in U1, the deictic gesture is not directly on a particular object. Fusing intention and attention structures from each individual inputs presents some ambiguities. For example, in Figure 4(b), there are three Attention structures for U1 gesture input. Each of them can be unified with the Attention structure from U1 speech input (in Figure 4a). The result of fusion is shown in Figure 10(a). Since the reference constraint in the speech input (Number: 1 in Figure 4a) indicates that only one attention structure is allowed, MIND uses contexts to eliminate inconsistent structures. In this case, A3 in Figure 10(a) indicates the information of interest is about the price of the city Irvington. Based on the domain knowledge that the city object cannot have the price feature, A3 is filtered out. As a result, both A1 and A2 are potential interpretation. Therefore, the Content in those structures are combined using a disjunctive relation as in Figure 10(b). Based on this revised conversation unit, RIA is able to arrange the follow-up question to further disambiguate the house of interest (R2 in Figure 2). This example shows that, modeling semantic information by fine-grained dimensions supports the use of domain knowledge in context-based inference, and can therefore resolve some ambiguities.

4.2 Deriving Unspecified Information

In a conversation setting, user inputs are often abbreviated. Users tend to only provide new information when it is their turn to interact. Sometimes, fusing individual modalities together still cannot provide overall meanings of those inputs. For example, after multimodal fusion, the conversation unit for U3 ("What about this one") does not give enough information on what the user exactly wants. The motivation and task of this input is not known as in Figure 11(a). Only based on the conversation context, is MIND able to identify the overall meaning of this input. In this case, based on the most recent conversation segment (DS1) in Figure 9 (also as in Figure 11b), MIND is able to derive Motivator and Method features from DS1 to update the conversation unit for U3 (Figure 11c). As a result, this revised conversation unit provides the overall meaning that the user is interested in finding out the price information about another house MLS7689432. Note that it is important to maintain a hierarchical conversation history based on goals and subgoals. Without such a hierarchical structure, MIND would not be able to infer the motivation of U3. Furthermore, because of the consistent representation of intention and attention at both the discourse level (in conversation segments) and the input level (in conversation units), MIND is able to directly use conversation context to infer unspecified information and enhance interpretation.

4.3 Improving Alignment

In a multimodal environment, users could use different ways to coordinate their speech and gesture inputs. In some cases, one reference/object mentioned in the speech input coordinates with one deictic gesture (U1, U3). In other cases, several references/ objects in the speech input are coordinated with one deictic gesture (U4, U5). In the latter cases, only using time stamps often cannot accurately align and fuse the respective attention structures from each modality. Therefore, MIND uses contexts to improve alignment based on our semantics-based representation. For example, from the speech input in U4 ("show me houses with this style around here"), three Attention structures are generated as shown in Figure 8(a). From the gesture input, only one Attention structure is generated which corresponds to the city of White Plains. Since the gesture input overlaps with both "this style" (corresponding to A2) and "here" (corresponding to A3), there is no obvious temporal relation indicating which of these two references should be unified with the deictic gesture. In fact, both A2 and A3 are potential candidates. Based on the domain context that a city cannot have a feature Style, MIND determines that the deictic gesture is actually resolving the refer-

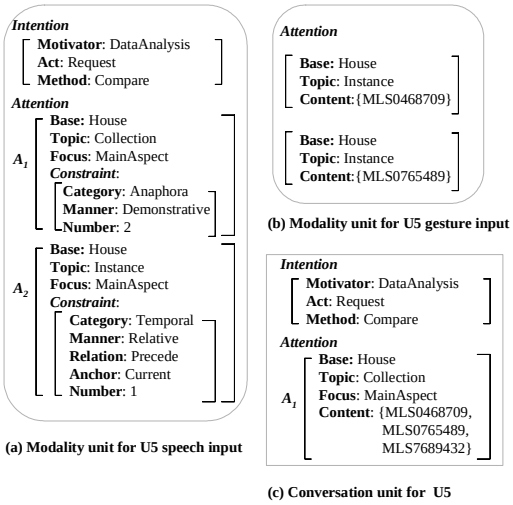


Figure 12. Improving alignment for U5

ence of "here". To resolve the reference of "this style", MIND uses the visual context which indicates a house is highlighted on the screen. A recent study (Kehler, 2000) shows that objects in the visual focus are often referred by pronouns, rather than by full noun phrases or deictic gestures. Based on this study, MIND is able to infer that most likely "this style" refers to the style of the highlighted house (MLS7689432). Suppose the style is "Victorian", then MIND is able to figure out that the overall meaning of U4 is looking for houses with a Victorian style and located in White Plains (as shown in Figure 8b).

Furthermore, for U5 ("Comparing these two houses with the previous house"), there are two Attention structures (A1 and A2) created for the speech input as in Figure 12(a). A1 corresponds to "these two houses", where the Number feature in the reference constraint is set 2. Although there is only one deictic gesture which points to two potential houses (Figure 12b), MIND is able to figure out that this deictic gesture is actually referring to a group of two houses rather than an ambiguous single house. Although the gesture input in U5 is the same kind as that in U1, because of the fine-grained information captured from the speech input (i.e., Number feature), MIND processes them differently. For the second reference of "previous house" (A2 in Figure 12a), based on the information captured in the temporal constraint, MIND searches the conversation history and finds the most recent house explored (MLS7689432). Therefore, MIND is able to reach an overall understanding of U5 that the user is interested in comparing three houses (as in Figure 12c).

5 Conclusion

To facilitate multimodal interpretation in conversational systems, we have developed a semantics-based representation to capture salient information from user inputs and the overall conversation. In this paper, we have presented three unique characteristics of our representation. First, our representation is based on fine grained semantic models of intention, attention and constraints that are important in information seeking conversation. Second, our representation is composed by a flexible combination of feature structures and thus supports complex user inputs. Third, our representation of intention and attention is consistent at different levels and therefore facilitates context-based interpretation. This semantics-based representation allows MIND to use contexts to resolve ambiguities, derive unspecified information and improve alignment. As a result, MIND is able to process a large variety of user inputs including those incomplete, ambiguous or complex ones.

6 Acknowledgement

The author would like to thank Shimei Pan and Michelle Zhou for their contributions on semantic models.

References

Bolt, R. (1980) Voice and gesture at the graphics interface. Computer Graphics, pages 262-270.

Carpenter, R. (1992) The logic of typed feature structures. Cambridge University Press.

Chai, J.; Pan, S.; and Zhou, M. X. (2002) MIND: A Semantics-based multimodal interpretation framework for conversational systems. To appear in Proceedings of International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialog Systems.

Cohen, P.; Johnston, M.; McGee, D.; S. Oviatt, S.; Pittman, J.; Smith, I.; Chen, L; and Clow, J. (1996) Quickset: Multimodal interaction for distributed applications. Proc. ACM MM'96, pages 31-40.

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

Jelinek, F.; Lafferty, J.; Magerman, D. M.; Mercer, R. and Roukos, S. (1994) Decision tree parsing using a hidden derivation model. Proc. Darpa Speech and Natural Language Workshop.

Johnston, M.; Cohen, P. R.; McGee, D.; Oviatt, S. L.; Pittman, J. A.; and Smith, I. (1997) Unification based multimodal integration. Proc. 35th ACL, pages 281-288.

Johnston, M. (1998) Unification-based multimodal parsing. Proc. COLING-ACL'98.

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

Wahlster, W. (1998) User and discourse models for multimodal communication. In M. Maybury and W. Wahlster, editors, Intelligent User Interfaces, pages 359-370.

Zancanaro, M.; Stock, O.; and Strapparava, C. (1997) Multimodal interaction for information access: Exploiting cohesion. Computational Intelligence, 13(4):439-464.

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