Combining Semantic and Temporal Constraints for Multimodal Integration in Conversation Systems

Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48864
jchai@cse.msu.edu

Pengyu Hong
Department of Statistics
Harvard University
Cambridge, MA 02138
hong@stat.harvard.edu

Michelle X. Zhou
IBM T. J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
mzhou@us.ibm.com

Abstract

In a multimodal conversation, user referring patterns could be complex, involving multiple referring expressions from speech utterances and multiple gestures. To resolve those references, multimodal integration based on semantic constraints is insufficient. In this paper, we describe a graph-based probabilistic approach that simultaneously combines both semantic and temporal constraints to achieve a high performance.

1 Introduction

Multimodal conversation systems allow users to converse with systems through multiple modalities such as speech, gesture and gaze (Cohen et al., 1996; Wahlster, 1998). In such an environment, not only are more interaction modalities available, but also richer contexts are established during the interaction. Understanding user inputs, for example, what users refer to is important. Previous work on multimodal reference resolution includes the use of a focus space model (Neal et al., 1998), the centering framework (Zancanaro et al., 1997), context factors (Huls et al., 1995), and rules (Kehler 2000). These previous approaches focus on semantics constraints without fully addressing temporal constraints. In a user study[1], we found that the majority of user referring behavior involved one referring expression and one gesture (as in [S2, G2] in Table 1). The earlier approaches worked well for these types of references. However, we found that 14.1% of the inputs were complex, which involved multiple referring expressions from speech utterances and multiple gestures (S3 in Table 1). To resolve those complex references, we have to not only apply semantic constraints, but also apply temporal constraints at the same time.

	G1 no gest.	***G2*** one gest	***G3*** mul. gest.	**Total Num**
S1: no expression	2	1	0	3
S2: one expression	12	117	2	131
S3: mul. expressions	1	6	15	22
Total Num	15	124	17	156

Table 1: Referring patterns from the user study

For example, Figure 1 shows three inputs where the number of referring expressions is the same and the number of gestures is the same. The speech utterances and gestures are aligned along the time axis. The first case (Figure 1a) and the second case (Figure 1b) have the same speech utterance but different temporal alignment between the gestures and the speech input. The second case and the third case (Figure 1c) have a similar alignment, but the third case provides an additional constraint on the number of referents (from the word "two").

Although all three cases are similar, but the objects they refer to are quite different in each case. In the first case, most likely "this" refers to the house selected by the first point gesture and "these houses" refers to two houses selected by the other two gestures. In the second case, "this" most likely refers to the highlighted house on the display and "these houses" refer to three houses selected by the gestures. In the third case, "this" most likely refers to the house selected by the first point gesture and "these two houses" refers to two houses selected by the other two point gestures.

[1] We are developing a system that helps users find real estate properties. So here we use real estate as the testing domain.

Resolving these complex cases requires

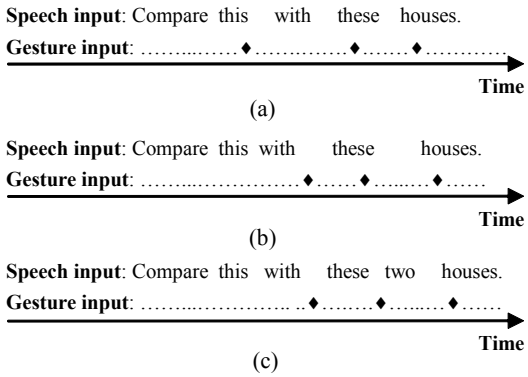

Figure 1. Three multimodal inputs under the same interaction context. The timings of the point gestures are denoted by "♦".

simultaneously satisfying semantic constraints from inputs and the interaction contexts, and the temporal constraints between speech and gesture.

2 Graph-based Approach

We use a probabilistic approach based on attributed relational graphs (ARGs) to combine semantic and temporal constraints for reference resolution. First, ARGs can adequately capture the semantic and temporal information (for both referring expressions and potential referents). Second, the graph match mechanism allows a simultaneous application of temporal constraints and semantic constraints. Specifically, we use two attributed relational graphs (ARGs). One graph corresponds to all referring expressions in the speech utterances, called the referring graph. The other graph corresponds to all potential referents (either coming from gestures or contexts), called the referent graph. By finding the best match between the referring graph and the referent graph, we can find the most possible referent(s) to each referring expression.

An ARG consists of a set of nodes and a set of edges. For example, Figure 2(a) is the referring graph for the speech utterance in Figure 1(c). There are two nodes corresponding to two referring expressions "this" and "these two houses" respectively. Each node encodes the semantic and temporal information of the corresponding referring expression such as the semantic type of the potential referent, the number, the start and end time the expression was uttered, etc. The edge between two nodes indicates the semantic and temporal relations between these two expressions. Similarly, Figure 2(b) is the referent graph for the input in Figure 1(c). This referent graph consists of four sub-graphs. Three sub-graphs correspond to three gestures respectively. Each node in these sub-graphs corresponds to one object selected by the gesture. Each node encodes the semantic and temporal information of the selected object, as well as the probability this object is actually selected. There is also a sub-graph corresponding to the interaction context. Each node in this sub-graph represents an object in the focus in the last interaction turn. The sub-graphs are connected via semantic type and temporal relations.

With the ARG representations described above, the reference resolution problem becomes matching the referent graph with the referring graph. Suppose we have two graphs to be matched:

- The referent graph $G_c = \langle \{a_x\}, \{r_{xy}\} \rangle$, where $\{a_x\}$ is the node list and $\{r_{xy}\}$ is the edge list. The edge r_{xy} connects nodes a_x and a_y.

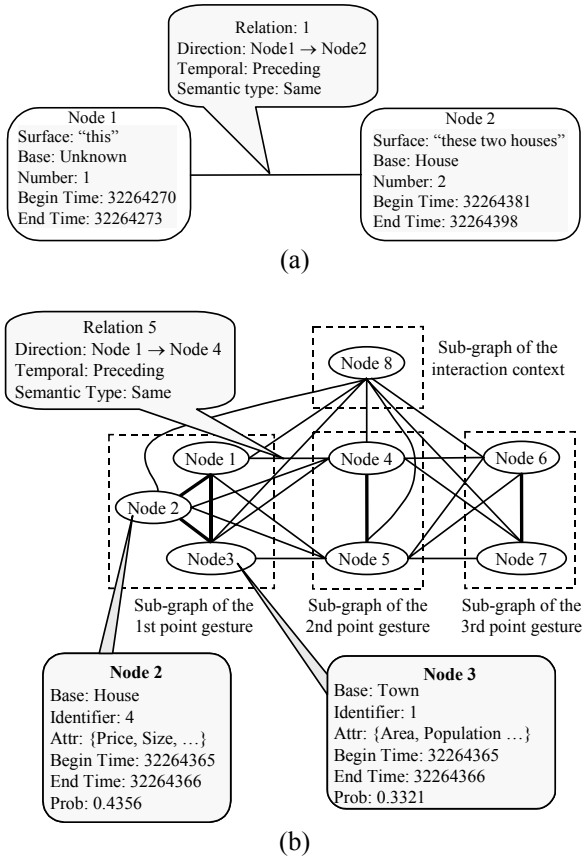

Figure 2. The ARG representation for references in Figure 1(c). (a) The referring graph (b) The referent graph, where dashed rectangles represent sub-graphs.

- The referring graph $G_s = \langle \{\alpha_m\}, \{\gamma_{mn}\} \rangle$, where $\{\alpha_m\}$ is the node list and $\{\gamma_{mn}\}$ is the edge list. The edge γ_{mn} connects nodes α_m and α_n.

The match process is to maximize the following function:

$$Q(G_c, G_s) = \sum_x \sum_m P(a_x, \alpha_m) \zeta(a_x, \alpha_m) + \sum_x \sum_y \sum_m \sum_n P(a_x, \alpha_m) P(a_y, \alpha_n) \psi(r_{xy}, \gamma_{mn}) \quad (1)$$

with respect to $P(a_x, \alpha_m)$, the matching probabilities between the referent node a_x and the referring node α_m.

The function $Q(G_c, G_s)$ measures the degree of the overall match between the referent graph and the referring graph. This function not only considers the similarities between nodes as indicated by the function $\zeta(a_x, \alpha_m)$, but also considers the similarities between edges as indicated by the function $\psi(r_{xy}, \gamma_{mn})$. Both node similarity and edge similarity functions are further defined by a combination of semantic and temporal constraints. For example, $\zeta(a_x, \alpha_m) = Sem(a_x, \alpha_m) Tem(a_x, \alpha_m)$, where $Sem(a_x, \alpha_m)$ measures the semantic compatibility by determining whether the semantic categories of a_x and α_m are the same, whether their attributes are compatible, and so on. $Tem(a_x, \alpha_m)$ measures the temporal alignment and is empirically defined as follows:

$$Tem(a_x, \alpha_m) = \begin{cases} \exp\left(-\frac{|time(a_x) - time(\alpha_m)|}{2000}\right), & a_x \text{ is from gesture} \\ 0.1, & a_x \text{ is from context} \end{cases}$$

To maximize (1), we modified the graduated assignment algorithm (Gold and Rangarajan, 1996). When the algorithm converges, $P(a_x, \alpha_m)$ gives us the matching probabilities. Details are described in a separate paper.

3 Discussion

During the study, we collected 156 inputs. The system assigned time stamps to each recognized word in the utterance, and each gesture. Figure 3 shows an example of an input that consisted of two gesture inputs and a speech utterance "compare this house with this house". The first two lines represent two gestures. Each line gives information about when the gesture started and ended, as well as the selected objects with their probabilities. These data provided us information on how the speech and gesture were aligned (to the accuracy of milliseconds). These data will help us further validate the temporal compatibility function used in the matching process.

Input received on port 3334: 67275921 67277343 2 69 23 3 1 2 67275921 67277343 39218 10000 0 0 255 0.28571 70 23 2 2 2 67275921 67277343 39218 10000 0 0 255 1.

Input received on port 3334: 67278140 67279078 2 71 24 4 1 2 67278140 67279078 797 10000 255 0 0 0.74545 72 24 3 2 2 67278140 67279078 797 10000 0 0 255 1.

speech input: compare_67273821 this_67274160 House_67274490 with_67275547 this_67275847 House_67276096

Figure 3. Gesture and speech data

We described an approach that uses graph matching algorithm to combine semantic and temporal constraints for reference resolution. The study showed that this approach worked quite well (93% accuracy) when the referring expressions were correctly recognized by the ASR. In the future, we plan to incorporate spatial constraints.

References

P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J.Clow. 1996. Quickset: Multimodal Interaction for Distributed Applications. *Proceedings of ACM Multimedia*, 31-40.

S. Gold and A. Rangarajan. 1996. A graduated assignment algorithm for graph matching, *IEEE Trans. Pattern Analysis and Machine Intelligence*, *vol*. 18, *no*. 4. 377–388.

C. Huls, E. Bos, and W. Classen. 1995. Automatic Referent Resolution of Deictic and Anaphoric Expressions. *Computational Linguistics,* 21(1):59-79.

A. Kehler. 2000. Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction, *Proceedings of AAAI'00.* 685-689.

J. G. Neal, C. Y. Thielman, Z. Dobes, S. M. Haller, and S. C. Shapiro. 1998. Natural Language with Integrated Deictic and Graphic Gestures. *Intelligent User Interfaces, M. Maybury and W. Wahlster (eds.)*, 38-51.

W. H. Tsai and K. S. Fu. 1979. Error-correcting isomorphism of attributed relational graphs for pattern analysis. *IEEE Trans. Sys., Man and Cyb.*, vol. 9, 757–768.

W. Wahlster. 1998. User and Discourse Models for Multimodal Communication*, Intelligent User Interfaces, M. Maybury and W. Wahlster (eds.)*, 359-370.

M. Zancanaro, O. Stock, and C. Strapparava. 1997. Multimodal Interaction for Information Access: Exploiting Cohesion. *Computational Intelligence* 13(7):439-464.