Performance Evaluation and Error Analysis for Multimodal Reference Resolution in a Conversation System

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

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

Abstract

Multimodal reference resolution is a process that automatically identifies what users refer to during multimodal human-machine conversation. Given the substantial work on multimodal reference resolution; it is important to evaluate the current state of the art, understand the limitations, and identify directions for future improvement. We conducted a series of user studies to evaluate the capability of reference resolution in a multimodal conversation system. This paper analyzes the main error sources during real-time human-machine interaction and presents key strategies for designing robust multimodal reference resolution algorithms.

1 Introduction[*]

Multimodal systems enable users to interact with computers through multiple modalities such as speech, gesture, and gaze (Bolt 1980; Cassell et al., 1999; Cohen et al., 1996; Chai et al., 2002; Johnston et al., 2002). One important aspect of building multimodal systems is for the system to understand the meanings of multimodal user inputs. A key element of this understanding process is reference resolution. *Reference resolution* is a process that finds the most proper referents to referring expressions. To resolve multimodal references, many approaches have been developed, from the use of a focus space model (Neal et al., 1998), a centering framework (Zancanaro et al, 1997), contextual factors (Huls et al., 1995); to recent approaches using unification (Johnston, 1998), finite state machines (Johnston and Bangalore 2000), and context-based rules (Kehler 2000).

Given the substantial work in this area; it is important to evaluate the state of the art, understand the limitations, and identify directions for future improvement. We conducted a series of user studies to evaluate the capability of reference resolution in a multimodal conversation system. In particular, this paper examines two important aspects: (1) algorithm requirements for handling a variety of references, and (2) technology requirements for achieving good real-time performance. In the following sections, we first give a brief description of our system. Then we analyze the main error sources during real-time human-machine interaction and discuss the key strategies for designing robust reference resolution algorithms.

2 System Description

We implemented a multimodal conversation system to study multimodal user referring behavior and to evaluate reference resolution algorithms. Users can use both speech and manual gestures (e.g., point and circle) to interact with a map-based graphic interface to find information about real estate properties.

As shown in Figure 1, our system applies a semantic fusion approach that combines the semantic information identified from each modality. A key characteristic of the system is that, in addition to fusing information from different modalities, our system systematically incorporates information from the conversation context (e.g., the focus of attention from prior conversation), the

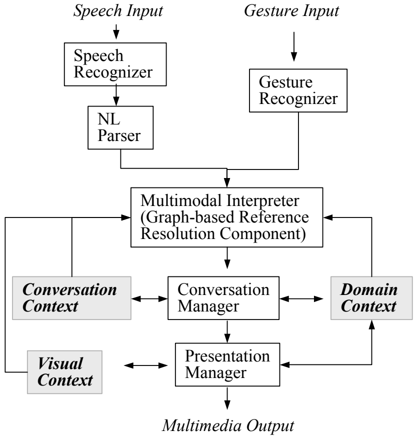

Figure 1: Overview of the system

[*] This work was supported by grant IIS-0347548 from the National Science Foundation and grant IRGP-03-42111 from Michigan State University.

visual context (e.g., objects on the screen that are in the visual focus), and the domain context (i.e., the domain knowledge).

The reference resolution approach is based on a graph-matching algorithm. Specifically, two attribute relational graphs are used (Tsai and Fu, 1979). One graph is called *referring graph* that captures referring expressions from speech utterances. Each node, corresponding to one referring expression, consists of the semantic information extracted from the expression and the temporal information when the expression is uttered. Each edge represents the semantic and temporal relation between two referring expressions. The second graph is called *referent graph* that represents all potential referents (including objects selected by the gesture, objects in the conversation history, and objects in the visual focus). Each node captures the semantic and temporal information about a potential referent (e.g., the time when the potential referent is selected by a gesture). Each edge captures the semantic and temporal relations between two potential referents. Given these graph representations, the reference resolution problem becomes a graph-matching problem (Gold and Rangarajan, 1996). The goal is to find a match between the referring graph and the referent graph that achieves the maximum compatibility between the two graphs. The details of this approach are described in (Chai et al., 2004).

3 Performance Evaluation and Analysis

We conducted several user studies to evaluate the performance of real time reference resolution using the graph-based approach. Eleven subjects participated in these studies. Each of them was asked to interact with the system using both speech and gestures (point and circle) to accomplish five tasks. For example, one task was to find the least expensive house in the most populated town. The voice from each subject was trained individually to minimize speech recognition errors.

3.1 Performance Evaluation

Table 1 summarizes the referring behavior observed in the studies and the performance of the system. The columns indicate whether there was no gesture, one gesture (point or circle), or multiple gestures involved in the input. The rows indicate the type of referring expressions in the speech utterances. Each table entry shows the system performance on resolving a particular combination of speech and gesture inputs. For example, the entry at <S2, G4> indicates that 35 inputs consist of demonstrative singular noun phrases (as the referring expressions) and a single circle gesture. Out of these inputs, 27 were correctly recognized and eight were incorrectly recognized by the speech recognizer. Out of the 27 correctly recognized inputs, 26 were correctly assigned referents by the system. Out of the eight incorrectly recognized inputs, references in two inputs were correctly resolved.

Consistent with earlier findings (Kehler 2000), the majority of user references were simple which only involved one referring expression and one gesture as shown in Table 1 (i.e., S1 to S8, with column G2 and G4). However, we have also found that 14% (31/219) of the inputs were complex, which involved multiple referring expressions from speech utterances (see the row S9). Some of these inputs did not have any accompanied gesture (e.g., <S9, G1>). Some were accompanied by one gesture (e.g., <S9, G4>) or multiple gestures (e.g., <S9, G3> and <S9, G5>). The referents to these referring expressions could come from user's gestures, or from the conversation context, or from the graphic display. To resolve these types of references, the graph-based approach is effective by simultaneously considering the semantic, temporal, and

	G1 No Gesture	**G2** One Point	**G3** Multiple Points	**G4** One Circle	**G5** Multiple Circles	**G6** Points and Circles	**Total Num**
S1: the (adj)*(N | Ns)	1(1), 1(0)	5(5), 3(0)	0	1(0), 1(1)	0	0(0), 1(0)	7(6), 6(1)
S2: (this|that) (adj*) N	3(3), 1(0)	29(28), 16(9)	3(2), 0(0)	27(26), 8(2)	1(1), 0(0)	1(1), 2(0)	64(61), 27(11)
S3: (these|those)(num)*(adj)*Ns	0	0	0	19(18), 12(3)	0	3(3), 2(0)	22(21), 14(3)
S4: it|this|that| (this|that|the)(adj)*one	1(0), 2(1)	2(2), 6(3)	0	4(2), 4(1)	0	0	7(4), 12(5)
S5:(these|those)(num)*(adj)*(ones)*|them	0	0	0	1(1), 1(0)	0	0	1(1), 1(0)
S6: here|there	0(0), 1(1)	1(1), 1(0)	0	6(6), 0(0)	0	0	7(7), 2(1)
S7: empty expression	1(1), 0(0)	1(1), 0(0)	0	2(0), 0(0)	0	0	4(2), 0(0)
S8: proper nouns	1(0), 0(0)	1(0), 3(2)	0(0), 3(2)	0(0), 3(0)	0	0(0), 3(0)	2(0) 12(4)
S9: multiple expressions	0(0), 1(0)	0	4(2), 0(0)	3(1), 7(1)	8(6), 5(0)	0(0), 3(0)	15(9), 16(1)
Total Num	7(5), 6(2)	39(37),29(14)	7(4), 3(2)	63(54), 36(8)	9(7), 5(0)	4(4), 11(0)	129(111), 90(26)

Table 1: Performance evaluation of the graph-matching approach to multimodal reference resolution. In each entry form "a(b), c(d)", "a" indicates the number of inputs in which the referring expressions were correctly recognized by the speech recognizer; "b" indicates the number of inputs in which the referring expressions were correctly recognized and were correctly resolved; "c" indicates the number of inputs in which the referring expressions were not correctly recognized; "d" indicates the number of inputs in which the referring expressions were not correctly recognized, but were correctly resolved. The sum of "a" and "c" gives the total number of inputs with a particular combination of speech and gesture.

contextual constraints.

3.2 Error Analysis

As shown in Table 1, out of the total 219 inputs, 137 inputs had their referents correctly identified (A complex input with multiple referring expressions was considered correctly resolved only if the referents to all the referring expressions were correctly identified). For the remaining 82 inputs in which the referents were not correctly identified, the errors mainly came from five sources as summarized in Table 2.

	Percentage
Speech recognition errors	55%
Language understanding errors	20%
Out of synchronization	13%
Disfluency	7%
Others	5%

Table 2: The distribution of error sources

A poor performance in speech recognition is a major error source. Although we have trained each user's voice individually, the speech recognition rate is still very low. Only 59% (129/219) of inputs had correctly recognized referring expressions. This is partly due to the fact that more than half of our subjects are non-native speakers. Fusing inputs from multiple modalities together can sometimes compensate for the recognition errors (Oviatt 1996). Among 90 inputs in which referring expressions were incorrectly recognized, 26 of them were correctly assigned referents due to the mutual disambiguation. However, poor speech recognition still accounted for 55% of the total errors. A mechanism to reduce the recognition errors, especially by utilizing information from other modalities will be important to provide a robust solution for real time multimodal reference resolution.

The second source of errors (20% of the total errors) came from insufficient language understanding, especially the out-of-vocabularies. For example, "area" was not in our vocabulary. So the additional semantic constraint expressed by "area" was not captured. Therefore, the system could not identify whether a house or a town was referred when the user uttered "this area". It is important for the system to have a capability of acquire knowledge (e.g., vocabulary) dynamically by utilizing information from other modalities and the interaction context. Furthermore, the errors also came from a lack of understanding of spatial relations (as in "the house just close to the red one") and superlatives (as in "the most expensive house"). Algorithms to align visual features to resolve spatial references as described in (Gorniak and Roy 2003) are desirable.

Among all errors, 13% came from unsynchronized inputs. Currently, we use an idle status (i.e., 2 seconds with no input from either speech or gesture) as the boundary to delimit an interaction turn. There are two types of out of synchronization. The first type is unsynchronized inputs from the user (such as a big pause between speech and gesture) and the other comes from the underlying system implementation. The system captures speech inputs and gesture inputs from two different servers through TCP/IP protocol. A communication delay sometimes split one synchronized input into two separate turns of inputs (i.e., one turn was speech input alone and the other turn was gesture input alone). A better engineering mechanism to synchronize inputs is desired.

The disfluencies from the users also accounted for about 7% of the total errors. Recent findings indicated that gesture patterns could be used as an additional source to identify different types of speech disfluencies during human-human conversation (Chen et al., 2002). As expected, speech disfluencies did not occur that much in our studies. Based on our limited cases, we found that gesture patterns could be indicators of speech disfluencies when they did occur. For example, if a user says "show me the red house (point to house A), the green house (still point to the house A)", then the behavior of pointing to the same house with different speech description usually indicates a repair. Furthermore, gestures also involve disfluencies, for example, repeatedly pointing to an object is a gesture repetition. Failure in identifying these disfluencies caused problems with reference resolution. It is important to have a mechanism that can identify these disfluencies using multimodal information.

The remaining 5% errors came from the implementation of our approach in order to reduce the complexity of graph matching. Currently, the referent graph only consists of potential referents from gestures, objects from the prior conversation, and the objects in the visual focus (i.e., highlighted on the screen). Therefore, it is insufficient to handle cases where users only use proper names (without any gestures) to refer to objects visible on the screen.

From the error analysis, we learned that variations in user inputs (e.g., variations in vocabulary and synchronization patterns), disfluencies in speech utterances, and even small changes in the input quality or the environment could seriously impair the real-time performance. The future research effort should be devoted to developing adaptive approaches for reference resolution to deal with unexpected inputs (e.g., inputs that are outside of system knowledge).

3.3 Design Strategies

The evaluation also indicates three important strategies in designing effective algorithms for multimodal reference resolution. The first strategy concerns with how to handle temporal relations. Consistent with the previous findings (Oviatt et al, 1997), in most cases (85%) in our study, gestures occurred before the referring expressions were uttered. However, we did find some exceptions. In 7% of cases, there was no overlap between speech and gesture and speech were uttered before gestures occurred. Furthermore, one user could have different temporal behavior at different stages in one interaction. In our study, five users exhibited varied temporal alignment during the interaction. Therefore, to accommodate different temporal variations, incorporating relative temporal relations between different modalities based on temporal closeness is preferred over incorporating absolute temporal relations or temporal orders.

Second, in a multimodal conversation, the potential objects referred to by a user could come from different sources. They could be the objects gestured at, objects in the visual focus (e.g., highlighted), objects visible on the screen, or objects mentioned in a prior conversation. It is important for reference resolution algorithms to simultaneously combine semantic, temporal, and contextual constraints. This is particularly important for complex inputs that involve multiple referring expressions and multiple gestures as described earlier.

Third, depending on the interface design and the underlying architecture for multimodal systems, different types of uncertainties occur during the process of input interpretation. For example, in our interface, each house icon is built on top of the town icon. Therefore, a pointing gesture could result in several possible objects. Once a touch screen is used, a finger point may result in different possibilities. Furthermore, most systems like ours are based on the pipelined architecture as shown in Figure1. The pipelined processes can potentially lose low probability information (e.g., recognized alternatives with low probabilities) that could be very crucial when incorporated with other modalities and the interaction context. Therefore, it is important to retain information at different levels and systematically incorporate the imprecise information.

4 Conclusion

This paper presents an evaluation of graph-based multimodal reference resolution in a conversational system. The evaluation indicates that, the real-time performance is largely dependent on speech recognition performance, language processing capability, disfluency detection from both speech and gesture, as well as the system engineering issues. Furthermore, the studies identify three important strategies for robust multimodal reference resolution algorithms: (1) using relative temporal constraints based on temporal closeness, (2) combining temporal, semantic, and contextual constraints simultaneously, and (3) incorporating imprecise information. A successful approach will need to consider both algorithmic requirements and technology limitations.

Acknowledgement

The authors would like to thank Keith Houck and Michelle Zhou at IBM T. J. Watson Research Center for their support in developing the system, and the anonymous reviewers for their helpful comments and suggestions.

References

Bolt, R.A. 1980. Put that there: Voice and Gesture at the Graphics Interface. *Computer Graphics*14(3): 262-270.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H. and Yan, H. 1999. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the CHI'99 Conference*, pp. 520-527. Pittsburgh, PA.

Chai, J. Y., Hong, P., and Zhou, M. X. 2004. A probabilistic approach to reference resolution in multimodal user interfaces*, Proceedings of 9th International Conference on Intelligent User Interfaces (IUI)*: 70-77. Madeira, Portugal, January.

Chai, J., Pan, S., Zhou, M., and Houck, K. 2002. Context-based Multimodal Interpretation in Conversational Systems. *Fourth International Conference on Multimodal Interfaces.*

Chen, L., Harper, M. and Quek, F. 2002. Gesture patterns during speech repairs. *Proceedings of International Conference on Multimodal Interfaces (ICMI).*

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

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

Gorniak, P. and Roy, D. 2003.Grounded Semantic Composition for Visual Scenes. *Journal of Artificial Intelligence Research.*

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

Johnston, M. 1998. Unification-based Multimodal parsing, *Proceedings of COLING-ACL.*

Johnston, M. and Bangalore, S. 2000. Finite-state multimodal parsing and understanding. *Proceedings of COLING.*

Johnston, M., Bangalore, S., Visireddy G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., and Maloor, P. 2002. MATCH: An Architecture for Multimodal Dialog Systems, in *Proceedings of ACL.*

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

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

Oviatt, S., DeAngeli, A., and Kuhn, K. 1997. Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction, In *Proceedings of Conference on Human Factors in Computing Systems: CHI '97.*

Tsai, W.H. and Fu, K.S. 1979. Error-correcting isomorphism of attributed relational graphs for pattern analysis. *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, pp. 757–768.

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