Ontology-Based Multi-Party Meeting Understanding[*]

Matthew Purver, John Niekrasz, Stanley Peters
Center for the Study of Language and Information, Stanford University
Stanford, CA 94305, USA
{mpurver, niekrasz, peters}@csli.stanford.edu

January 3, 2005

1 Introduction

This paper describes current and planned research efforts towards developing multimodal discourse understanding for an automated personal office assistant. The research is undertaken as part of a project called The Cognitive Agent that Learns and Organizes (CALO) (see `http://www.ai.sri.com/project/CALO`). The CALO assistant is intended to aid users both personally and as a group in performing office-related tasks such as coordinating schedules, providing relevant information for completing tasks, making a record of meetings, and assisting in fulfilling decisions.

Our focus within this enterprise is on understanding, describing, and automatically participating in multimodal human-human and human-computer discourse amongst CALO users and the system itself. This aspect is functionally realized by the system in its role as a persistent presence before, during, and after meetings; firstly by helping to set up and coordinate meetings and meeting agendas; secondly by extracting detailed information about what was discussed, what the participants' actions were, and what decisions were reached; thirdly by interactively reporting on this extracted information; and eventually by interactively providing relevant and useful information or responding to queries during the meeting itself.

Currently, the system is not interactive during meetings (though this is a plan for future versions), which allows human participants to interact with each other completely naturally. Natural unscripted meeting discourses depart significantly from the standard computational dialogue understanding paradigm. Firstly, as the CALO agent is not a primary participant, the discourse cannot be automatically directed or constrained, and clarification cannot be sought by the system when ambiguity or misunderstanding is encountered. While this makes full interpretation and disambiguation more difficult, it also eliminates the requirement to perform such interpretation and disambiguation immediately. Also, these discourses are by nature highly multimodal; full understanding requires identification of speakers and addressees, along with resolution of reference to other participants and objects, and integration of both verbal and non-verbal communication (including not only gesture and eye gaze but drawing and writing).

Our general approach is to make use of the vast amounts of noisy data obtained from the physical-awareness, speech, gesture, and sketch recognition agents by organizing the information using a central, unified *multimodal discourse (MMD) ontology and knowledge base*. We couple this with a dialogue-understanding framework which maintains and shares multiple hypotheses between discourse-understanding components. Finally, through modular ontology design, we then allow integration and relation of the many hypotheses with information from sources external to the meeting and specific to the domain of discourse, the ultimate goal being to relate the highly ambiguous but physically observable interactions with the less ambiguous (and sometimes perfectly known) knowledge obtained through direct interaction with the system

[*] This work was supported by DARPA grant NBCH-D-03-0010. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

prior to and after the meeting.

Virtuality In its current incarnation, the prototype assistant monitors meetings non-interactively (although the user can interact with the system afterwards to access the extracted information for use in their other activities). Nevertheless, it is important to consider both the system as a virtual discourse participant, and to consider the physical and virtual roles it and the system components play during meetings.

Firstly, CALO has a significant physical presence in the meeting room – its sensors (microphones, stereo and panoramic cameras, electronic whiteboards and computer desktops) and their associated elements are part of the environment and become potential referents and subjects for discussion. Potential discourse referents are however not constrained to the physical. Just as agents do, referents occur on a scale from physically concrete (microphones, cameras), to partially concrete (documents or computer files being presented and discussed, diagrams on a whiteboard), and to completely virtual (abstract discussion of tasks and milestones). This is a particularly significant issue for the system because documents, presentation slides and individual bullet points are expected to be frequently discussed referents. They are the entities of which the other system components have a deeper understanding, and around which the purpose of meeting understanding revolves. They appear in the meeting in (partially) physical forms, and they are discussed conceptually through natural language.

Secondly, as the assistant (and its use) develops, it will become a potential participant in the discourse, and questions may eventually be directed towards it. It may also be designed to use its intiative to provide potentially helpful information. However, without a central physical manifestation (e.g. an avatar) which can be taken as an “addressee” in a traditional sense, we will be required to investigate new techniques for discourse understanding as the participants’ own techniques evolve for addressing CALO, whether with eye gaze, physical gesture, or spoken language.

2 Ontologies and Knowledge-Bases

The wide range of possible physical incarnations of CALO in future versions of the system, and the unknown ways in which human participants will address the system and the knowledge it contains or presents, has forced us to design a system which allows for a flexible representation of agents and referents along the virtuality continuum.

Our approach is ontology-based: we use a central ontology of multimodal discourse to describe all communicative actions performed in simultaneously physical and conceptual levels, from the lowest level of basic perceptual data through to higher levels of symbolic interpretations of these data. For example, we provide specifications for raw video and audio data, extracted elementary physical characteristics (e.g. people’s locations, head and arm orientations, gaze directions and utterance transcriptions), and symbolic interpreted actions like looking at something, drawing a line, and asking a question or making a proposal. The ontology provides a *lingua franca*, encoded in a formal description logic, with which the individual understanding components share knowledge about the discourse and integrate reasoning and inference capabilities.

Importantly, this MMD ontology contains only information relating to the communicative activity involved in the meeting. Referents or concepts having to do with the subject matter under discussion are described in a separate domain ontology which is formally linked though independent; specific conversational structures (modes of conduct) that might be specific to a particular discourse type such as corporate decision-making meetings or human-computer information-seeking dialogues are placed in an application-specific component; and information about surface lexical items themselves are confined to a language-specific taxonomy (see Flycht-Eriksson (1999) for a discussion of common modularizations of dialogue system knowledge). This allows the MMD component to be maximally independent of domain, language or application, while ensuring that domain-specific referents are equally available to all interpretation components.

In addition to the ontology specification, we have developed a persistent temporal knowledge base system called *KronoBase*, which is used both as a repository of knowledge collected by the component agents and as a manager of meta-information about the knowledge itself. Knowledge is asserted in a form which conforms to that which is specified by the ontology, but this knowledge will often be speculative or incomplete (as produced from the viewpoint of individual components). KronoBase maintains this speculative information in the form of probabilities and underspecified logical structures, allowing later learning via reinforcement or supplementary information. In addition, it maintains reference to the source and time of the assertion and the context in which it was asserted, thus enabling access to a complete history of the knowledge state.

This results in a generic framework for persistent, collaborative interpretation and reinterpretation, which is necessary for linking the extracted knowledge to the interactions which happen before the meeting, after the meeting, and during other meetings. For these extra-meeting interactions, we are developing a question-answering dialogue agent *Meeting Reviewer* to allow a user to query information about the discourse history itself: not only what decisions were made and when, but who made them, who (dis)agreed with them, and whether they were later modified. Allowing the user to interact with and correct the system if answers are wrong can directly provide it with information to adjust and re-learn its recently acquired information and understanding algorithms. This is a fundamental aspect of how the system aligns the information obtained in the non-interactive meeting environment with that which is obtained in the virtual interaction outside the meeting.

3 Multi-Party Discourse Understanding

The relatively free subject domain prevents the use of a constrained grammar for semantic interpretation. Instead, we intend to use a more robust approach based on shallow chunk parsing followed by semantic construction governed by the lexical and domain ontologies, with pragmatic interpretation then being guided by constraints provided by the domain and MMD ontologies themselves together with the knowledge-base's current model of discourse context (see e.g. Ludwig et al., 2002; Milward and Beveridge, 2004). The centrality of the ontologies allows understanding components to be to a large degree domain-independent: lexical entries, names, concepts and their combinatory possibilities are all specified within the domain and lexical ontologies rather than the generic processing rules.

It is therefore vital that the domain ontology reflect the full possibilities of multi-party discourse. All possible participants and referents (together with their properties) must be represented, including the CALO agent and its physical attributes, and physical and virtual elements such as electronic documents and slide presentations. In order to achieve successful interpretation and disambiguation, the ontology must also reflect the possible and impossible relations between these entities, and the roles they can or are likely to play in these relations. This can aid us in interpretation and disambiguation: information from sources outside the meeting can tell us who is likely to play a particular role in the meeting, and thus perhaps to be a likely speaker or addressee of a certain agenda item or action item assignment; similarly the record of a past meeting may tell us who is likely to be the speaker of a new report on an old action item from that meeting.

CALO as a Participant Further interesting questions are raised by the planned participation of the CALO agent itself in post-meeting interaction and in the meeting discourse itself. Given the virtual nature of the CALO agent, and its lack of a central physical presence, it seems likely that its interaction will have different properties from the rest of the human-human discourse: addressing CALO cannot be signalled by directing eye gaze, for example – it seems more likely that CALO-directed questions will be explicitly linguistically signalled as such. More generally, the question we need to ask is: what properties will differ between human-human and human-CALO interaction, and what properties will stay the same? It seems likely that the general rules of linguistic interpretation, ellipsis resolution, and discourse coherence will stay the same; however, it also seems likely that CALO will be governed by different discourse obligations to other participants (questions). Use of an obligation-based model to drive CALO's contributions (Poesio and Traum, 1997, e.g.) must therefore take this into account. It may even be the case that low-level interpretation will differ: for example, (Ginzburg and Fernández, 2005) show that bare answers are often further separated from their questions in multi-party discourse than in two-party discourse, and we expect that this may hold for our human-human meetings in general. However, for CALO-directed questions, it may not be acceptable or coherent.

4 Multimodal Understanding and Mutual Disambiguation

Our ontology-based approach allows the principled fusion of multimodal communicative information. As the various components use the same ontology to describe the possible objects and concepts being referred to (although in different modalities and/or languages), the hypotheses of these components can be combined in the knowledge-base to supplement, reinforce, or contradict each other. The basic properties of the ontological classes can also give information about likely reference or selectional restrictions, helping with disambiguation of gesture or linguistic interpretation.

An example might be the utterance "I think we should move that milestone back 6 months", co-occurring with a pointing gesture referring to a point on a project plan diagram being displayed on a projection screen. The deictic reference to "that milestone" is unresolvable by speech alone, and the location being pointed to cannot be resolved to a high-enough precision by vision alone. Each component produces either an underspecified or probabilistic analysis; the central ontology and knowledge base allows these to be combined, either directly by inference rules explicitly encoded as part of the ontology, or by making them centrally available to a third interpretive agent which performs the combination.

In this type of systematic knowledge exchange, each component's input domain is characterized as a subset of the ontology, and each component's range of output is as well. This establishes an ontologically formalized relationship between the types of knowledge the components may produce and consume, which in turn establishes a hierarchy of interpretation: from low-level perceptual actions to mid-level symbolic physical interpretation to high-level interpretations of discourse structure and information exchange.

We anticipate multimodal resolution of this kind being very important in overall meeting understanding. The approach here seems generally applicable: it could be used for combining eye gaze data with linguistic or gestural reference and/or with addressee identification (Traum, 2004; Jovanovic and op den Akker, 2004); or for combining linguistic semantic interpretation with physical or virtual gestures towards (virtual) documents and/or their sub-parts. However, in order to apply it efficiently, several important questions must be answered. Firstly, how exactly can we generalise the simple example of deictic reference to a milestone on a diagram to physical and virtual objects at varying degrees of granularity (documents, paragraphs, bullet points etc.)? How is reference to such virtual objects actually expressed gesturally and verbally? Secondly, what degree of co-occurrence can we expect? How closely temporally aligned are references in multiple modalities likely to be, and does this change between reference to physical and virtual objects, or between human-human discourse and human-CALO discourse? Does eye gaze or gesture really correspond with addressee? We are currently in the process of planning experiments to help answer some of these questions.

References

Annika Flycht-Eriksson. Representing knowledge of dialogue, domain, task and user in dialogue systems – how and why? *Electronic Transactions on Artificial Intelligence*, 3(2):5–32, 1999.

Jonathan Ginzburg and Raquel Fernández. Conversational acts and non-sentential utterances in multilogue. In *Proceedings of the 6th International Workshop on Computational Semantics (IWCS-6)*, Tilburg, January 2005. To appear.

Natasa Jovanovic and Rieks op den Akker. Towards automatic addressee identification in multi-party dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, MA, 2004. Association for Computational Linguistics.

Bernd Ludwig, Kerstin Bücher, and Günther Görz. Corega tabs: Mapping semantics onto pragmatics. In G. Görz, V. Haarslev, C. Lutz, and R. Möller, editors, *Proceedings of the KI-2002 Workshop on Applications of Description Logics*, Aachen, 2002.

David Milward and Martin Beveridge. Ontologies and the structure of dialogue. In J. Ginzburg and E. Vallduví, editors, *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, pages 69–77, Barcelona, July 2004.

Massimo Poesio and David Traum. Conversational actions and discourse situations. *Computational Intelligence*, 13(3), 1997.

David Traum. Issues in multi-party dialogues. In F. Dignum, editor, *Advances in Agent Communication*, pages 201–211. Springer-Verlag, 2004.