Ontology-Based Discourse Understanding for a Persistent Meeting Assistant*

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

Introduction

This paper describes current research efforts towards automatic understanding of multimodal discourse for a persistent personal office assistant. The assistant aids users in performing office-related tasks such as coordinating schedules with other users, providing relevant information for completing tasks, making a record of meetings, and assisting in fulfilling the decisions made in the meetings. Our focus within this enterprise is on *meeting understanding* – extracting detailed information about what was discussed, what the participants' actions were, what decisions were reached, and the action items assigned. The assitant monitors meetings non-interactively, although the user can interact with the system afterwards to access the extracted information for use in their other activities.

Natural multi-human meetings pose several significant challenges for an automatic discourse understanding system. Unconstrained verbal interactions generate noisy speech signals which increase errors and reduce confidence in speech recognition, which in turn propagates ambiguity to other components; the relatively unrestricted subject domain limits the utility of constrained lexicons and grammars for interpretation and forces the use of online learning of new words, concepts, and modes of interaction; participants' use of multiple communicative modalities means much of the discourse is unimodally ambiguous, requiring integrated multimodal interpretation. Given the complexity of this task and the highly ambiguous component interpretations, we believe that a flexible, unifying *multimodal discourse ontology* coupled with a generalizable framework for sharing hypotheses between understanding components is not only central to approaching these challenges, but provides in itself a means of tackling some of the more difficult problems of understanding in a persistent, dynamic and multimodal context.

A Multimodal Discourse Ontology

As a first step, we have constructed an ontology of multimodal discourse. An ontology, as widely defined in knowledge engineering, is an "explicit specification of a conceptualization" (Gruber 1993). For the multimodal discourse (hereafter, MMD) ontology we describe here, the conceptualization describes all communicative actions performed during multimodal discourse, 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 components share knowledge about the discourse and integrate reasoning and inference capabilities.

Importantly, the MMD ontology contains only information relating to the communicative activity involved in the meeting. Concepts having to do with the subject matter under discussion are kept in a separate domain ontology; 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.

KronoBase: A Temporal Knowledge Base

In addition to the ontology specification described above, we have developed a persistent temporal knowledge base system called *KronoBase*, which is used for the exchange of information gathered from the perceptual and interpretive activities performed by the components during the meeting.

The role of KronoBase is 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 specula-

*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.

tive 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.

Multimodal Fusion

One fundamental example of this kind of collaborative interpretive activity is the understanding of discourse acts that are performed through the use of multiple distinct modalities. For example, consider the utterance "I think we should move that milestone back 6 months", occurring simultaneous to 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 systematic kind of 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 sort of hierarchy of interpretation: from low-level perceptual actions to mid-level symbolic physical interpretation to high-level interpretations of discourse structure and information exchange.

Discourse Interpretation

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 parsing (e.g. keyword-spotting or 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 the MMD ontology itself together with the knowledge base's current model of discourse context (Ludwig, Bücher, & Görz 2002; Milward & 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.

Learning

The dynamic aspect of the ontology, together with knowledge about the meeting events and a temporally-grounded notion of knowledge assertion, allows understanding to adapt to new words, names, concepts or facts based on the history of the communicative context. As information is added to the knowledge base, furthering the specification of previously added partial information (whether by other components external to the meeting, or by explicit instruction by meeting participants), the understanding routines will automatically use them in subsequent interpretation. For example, the detection of a new face coupled with discussion about an object which is called "John" (and perhaps accompanied by a more explicit deictic pointing reference), provides reason to assert the previously unknown association of that name with the new person. As this and all other information is persistent, a more informed reinterpretation of utterances made both previous and subsequent to that assertion can be made.

In general, new entries can be initially loosely specified with subsequent experience filling in more detail (in terms of the ontology, moving from superclass to subclass), and thus allowing gradual learning over time. This process is of course facilitated by the direct integration of multimodal information (e.g. combining pointing gestures with new names when new objects are discussed), but also by the non-interactivity of the system during a meeting and its persistence between meetings: as there is no requirement to act on or respond to each human utterance immediately, understanding can be temporarily underspecified until resolved (or strengthened beyond a certain probabilistic threshold) by subsequent discourse.

Meeting Review

The temporal capabilities of *KronoBase*, together with its persistence between meetings, enable post-meeting interaction which can provide not only useful functionality but feedback to allow the system to learn further. We are developing a question-answering dialogue system *Meeting Reviewer* to allow a user to query information about the meeting 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.

References

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

Gruber, T. R. 1993. A translation approach to portable ontologies. *Knowledge Acquisition* 5(2):199–220.

Ludwig, B.; Bücher, K.; and Görz, G. 2002. Corega tabs: Mapping semantics onto pragmatics. In Görz, G.; Haarslev, V.; Lutz, C.; and Möller, R., eds., *Proceedings of the KI-2002 Workshop on Applications of Description Logics*.

Milward, D., and Beveridge, M. 2004. Ontologies and the structure of dialogue. In Ginzburg, J., and Vallduví, E., eds., *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, 69–77.