Workshop on Mental model ascription by language-enabled intelligent agents, July 31, 2013, Berlin, Germany.

Abduction of Mental States with a Formal Theory of Commonsense Psychology

Andrew S. Gordon, Jerry R. Hobbs, Katya Ovchinnikova,
Melissa Roemmele, and Louis-Philippe Morency
University of Southern California

Successful communication and collaboration between humans and intelligent agents of the future will require a robust ability to algorithmically infer the subjective mental states of the human participants. As in human to human interaction, the central concerns of plans, goals, emotions, and beliefs of another must inferred from a mix of explicit and implicit evidence in language, along with contextual and behavioral cues. We propose that this cognitive ability of mental model ascription is best conceived as a process of abduction, where a hypothetical explanation is inferred to account for observable evidence. In this approach, speech and other behavior of a person are observables that require explanation, where the challenge is to find a theoretical explanation that requires the fewest assumptions. Recent advances in abduction-based language processing [1] have led to efficient implementations of Hobbs's conception of weighted-abduction [2], where textual inputs (observations) are explained by searching a knowledgebase of logical axioms for the least-cost proof, with cost incurred when assumptions are asserted.

For mental model ascription, the knowledgebase of axioms used to explain the observable behavior of others would constitute a Theory of Mind, a set of inference rules that encode a commonsense understanding of human psychology. In our own work [3], we attempt to formalize a large-coverage theory of commonsense psychology in first-order predicate logic. Our formalization efforts have been organized around 30 specific content theories of various mental states and processes, including those related to plans, goals, emotions, beliefs, decisions, explanations, and expectations. We hypothesize that the contents of these formal theories are sufficiently rich to serve as a theoretical foundation for mental model ascription, and are now working to integrate these theories into an abduction-based interpretation system.

To explore this hypothesis, we have chosen to focus our initial efforts not on the interpretation of language evidence, but rather on motion and gesture observations. To further simplify the task of recognizing low-level motion and gesture actions, we are building a system that ascribes mental models to abstract shapes moving around an empty field of view in the style of the stimulus used in Heider and Simmel's classic experiment on intention perception [4]. In this ongoing project, our aim is to produce a computational model of mental state attribution from the observable actions of others, and build a foundation for a broader model that incorporates additional evidence including language.

[1] Ovchinnikova, E. (2012) Integration of World Knowledge for Natural Language Understanding, Atlantis Press, Springer.
[2] Hobbs, J., Stickel, M., Appelt, D., and Martin, P. (1993) Interpretation as Abduction, Artificial Intelligence 63(1-2):69-142.
[3] Gordon, A. and Hobbs, J. (2004) Formalizations of Commonsense Psychology. AI Magazine 25(4):49-62.
[4] Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. American Journal of Psychology, 13, 1944.