Contextual Information Sharing in Natural Language and Gesture Crossmodal Integration for Aged People Assistive Home Care Application

Olga Vybornova (vybornova@tele.ucl.ac.be)
Monica Gemo (gemo@tele.ucl.ac.be)
Ronald Moncarey (moncarey@tele.ucl.ac.be)
Benoit Macq (macq@tele.ucl.ac.be)
UCL-TELE, Universite Catholique de Louvain,
Batiment Stevin, Place du Levant, 2, B-1348, Louvain-la-Neuve, Belgium

Keywords: multimodal assistive interface, domain ontology, user profile, context awareness, semantic representations, multimodal fusion.

Application overview

We present a method for knowledge-based context sharing within multimodal high level fusion integrating data obtained from spoken input and visual scene analysis. The goal of our work is to develop a multimodal interface as an "intelligent diary" to proactively assist elderly people living alone at home to perform their daily activities, to prolong their safe and secure personal autonomy, to support their active ageing and social cohesion.

To provide natural interaction with the user(s) a system must be able to comprehend the fully coordinated mind-and-body behavior, to handle semantic-level input data fusion, i.e. to combine information arriving simultaneously from different modalities into one or several unified and coherent representations of the user's intention. Our context-aware user-centered application should accept spontaneous multimodal input – English speech, 3D gestures (pointing, iconic, possibly metaphoric) and user's physical action. In the near future we plan to add in the research also eye gaze tracking modality to facilitate capturing salient objects in the scene. Thus we have a restricted domain to work with, but we deal with unrestricted natural human behavior – spontaneous spoken input and gesture. At present we are implementing multi-stage crossmodal fusion that is seen promising from the point of view of reference ambiguity resolution before the final fusion. It is exactly crossmodal fusion that helps us cope with problems of speech recognition for elderly people caused by age-related decline of language production ability (for instance, difficulties in retrieving appropriate (familiar) words or tip-of-the-tongue (TOT) states when a person produces one or more incorrect sounds in a word (Burke & Shafto, 2004) because information from other modalities refines the language analysis at the early stage of recognition.

Method

Everything that is said or done is meaningful only in a particular context. To accomplish the task of semantic fusion we should take into account the information obtained at least in the following three types of context (Chai, Pan and Zhou, 2005): (i) domain context (meaning personalized prior knowledge of the domain, semantic frames with predefined action patterns, adaptive user profile, situation modeling, a priori developed and dynamically updated ontology defining subjects, objects, activities and relations between them for a particular person). (ii) conversational context (derived from natural language semantic analysis); (iii) visual context (capturing the user's gesture/action in the observation scene and allowing eye gaze tracking to enable salience models while activity monitoring).

To derive contextual information from spoken input we extract natural language semantic representations (discourse representation structures) (Bos, 2005) and map them onto the restricted domain ontology. This information is then processed together with visual scene input for multimodal reference resolution. The ontology allows contextual information sharing within the domain and serves as a metamodel for Bayesian networks used to analyze and combine the modalities of interest. With the help of non-deterministic weighting of multimodal data streams we obtain robust contextual fusion to recognize the user's intentions, to predict behavior, to provide reliable interpretation and to reason about the cognitive status of the person.

Acknowledgments

This work is supported by the European Commission FP6 Network of Excellence SIMILAR, project # FP6-507609 (http://www.similar.cc).

References

Bos J. Towards wide-coverage semantic interpretation. (2005). *Proceedings of IWCS-6.*

Burke D. and Shafto M. (2004). Aging and language production, *Current Directions in Psychological Science, 13.*

Chai J., Pan S. and Zhou M. (2005). MIND: A context-based multimodal interpretation framework, *Kluwer Academic Publishers*.

Pfleger N. and Alexandersson J. (2006) Towards resolving referring expressions by implicitly activated referents in practical dialogue systems, *Proc. of the Workshop on the Semantics and Pragmatics of Dialogue.*

Pollack M. (2005) Intelligent Technology for an Aging Population: The Use of AI to Assist Elders with Cognitive Impairment, *AI Magazine, 26(2):9-242005.*