A Connectionist Model of Covert Attention in Situated Language Comprehension

Dr. Marshall Mayberry, University of Saarland

1. November 2007, Raum 0.76 im Haus 24 von 15-17 Uhr

Over the past decade, the visual world paradigm has accumulated considerable evidence that shows the closely time-locked coordination of incremental language comprehension with attention to relevant objects and events in visual contexts. The evidence has also revealed that comprehension is anticipatory as revealed by attention to objects in a scene before they are mentioned. The interaction of language and scene is further marked by the rapid and seamless integration of, and adaptation to, diverse information sources in both the utterance and visual scene. These sources can interact dynamically, both complementing and, at times, conflicting with each other. Based on these hallmark properties of situated language comprehension: incrementality, anticipation, multimodal integration of diverse information sources, adaptation to available information, and coordination between language and visual context, Knoeferle and Crocker (2006) proposed the coordinated interplay account (CIA) to explain the seamless and rapid integration of linguistic and non-linguistic information during situated human language comprehension. The CIA posits that initially the unfolding utterance guides attention in the visual scene to establish reference to objects and events. The attended information in the scene then rapidly influences comprehension of the utterance, allowing the anticipation of upcoming role-fillers based on either the events and affordances of the scene (Knoeferle et al., 2005), or stored selectional and stereotypical relations (Kamide et al., 2003). Additionally, Knoeferle and Crocker (2006) provide evidence that people easily use available information sources, notably preferring information from depicted events when they conflict with stored knowledge.

Despite the growing evidence for these hallmark properties of situated language comprehension, no explicit model of the mechanisms which underlie utterance-mediated attention or the influence of scene information on comprehension has been developed. Accordingly, we present a model, CIANet, of these results using a recurrent neural network that processes a sentence incrementally into a case-role interpretation and that accepts an optional visual scene as additional, multimodal context. The network is able to adaptively exploit whatever information sources are available as each new word is encountered in context. In contrast with previous efforts to model the use of the scene (Mayberry et al., 2005), the model incorporates an explicit attentional mechanism to directly instantiate the CIA. It does so by learning to attend to the event in the scene that is most relevant to the utterance it is processing.

The agents, actions, and patients of the two events in the scene, when present, are propagated directly to the network's hidden layer through shared weights, and the attentional mechanism learns to dynamically shift the network's attention to the relevant event as the utterance is processed. The attentional mechanism is realized as a vector of gating units that uses implicit inhibition to competitively select the most relevant event in the scene through multiplicative connections to each event's constituents, making the model a recurrent sigma-pi network (Rumelhart et al., 1986). The attended event thus has more influence on the developing interpretation, allowing the model to anticipate upcoming role fillers, whether this anticipation is based on stereotypical associations or derived from depicted actions and their associated thematic roles in the depicted event. This mechanism both improves the network's overall performance and provides a quantitative model of human attentional behavior during situated comprehension, instantiating the central claims of the coordinated interplay account. In addition to incremental processing, the integration of multimodal information, the adaptive use of information from the scene, and the anticipation of upcoming role fillers, the network also demonstrates the tight temporal coordination of language and scene processing using attention as proposed by the CIA.

To show that it can go beyond fitting data and actually predict observed behavior, the network is only trained on complete interpretations for sentences in which either the scene provided relevant event information, or stereotypical information was available, but never both. Despite this, the model nevertheless exhibits the observed preferred reliance on depicted actions over stereotypical knowledge when tested on sentences with conflicting information sources, even when these sources were potentially equally informative.

Despite its ability to model the use of attention to relevant events in the scene, when present, and use stereotypical knowledge even when the scene is absent, CIANet is still a model of covert attention. It will identify and activate the event most relevant to an unfolding utterance, while suppressing the other, less relevant, event. We conclude the talk with a discussion of a generalized model that we are developing that shows promise in modelling overt attention at the level of event constituents, rather than just the event itself. The innovation of this model is the incorporation of previous work on semantically self-organized maps to allow the representation of flat semantic representations of acyclic directed graphs that require the implementation of context-sensitive pointers that capture dependencies between semantic frames and their argument fillers. The resulting system is accordingly more transparent that CIANet, and is not limited to two events, but can represent more typical scenes that involve only objects (such as distractors), or where the event information is implicit in the relationships among objects.