Workshop on the Interaction of Vision and Language in Cross-Modal Comprehension

Presentation 3

A Constraint-based Model for the Integration of Visual Scene Context into Sentence Processing
by Patrick McCrae

While cross-modal interactions at sensory level have been studied quite extensively over the past decades, it is only in relatively recent years that systematic research into the cross-modal interactions of the higher cognitive faculties is being undertaken. In the latter line of research, the interplay between vision and language has received particular attention, not least because of its potential to provide deeper insights into the nature of the underlying mechanisms of sentence processing and visual perception as such.

In an early application of the *Visual World Paradigm* Tanenhaus et al. were able to show in 1995 that sentence processing progresses incrementally and that eye fixations on visual scene elements directly correlate with the temporal progress of online sentence processing (Tanenhaus et al., 1995). More importantly, they concluded that *referentially relevant non-linguistic information immediately affects the manner in which the linguistic input is initially structured.* It is precisely this kind of interaction between visual context and language processing that has inspired and continues to inspire a broad range of investigations into the language-vision interface.

More recently, (Knoeferle, 2005) has studied the relative strength of the influence of immediate visual scene context upon sentence processing. Based on the evaluation of anticipatory eye movements in the presence of controlled visual scene contexts she concludes that the influence of immediate visual scene context upon language processing prevails over that of world knowledge.

Despite an increasing body of behavioural studies examining the language-vision interface, extremely few attempts to model these effects computationally have been reported at the time of writing. (Mayberry et al., 2006) provide a connectionist model for the interplay of scene, utterance and world knowledge. Their simple recurrent network implementation parses incrementally and models the influence of contextual information by feeding knowledge from step *n* into step *n+1* via a feedback loop. Particularly interesting in their approach is the capability to model the interaction between language and vision as a bidirectional process by incorporating the influence of linguistic processing on visual attention as a top-down effect. (Knoeferle & Crocker, 2006) refer to this bidirectionality in the interaction between language and vision as *Coordinated Interplay Account*.

The key challenge with artificial neural network approaches, however, is that the mechanistic details which lead to the actual output often remain obscure as they emerge from the parallel interactions of a large number of neurons in the hidden layers. For the modelling efforts at the language-vision interface it would therefore be desirable to have a mechanistically more transparent formalism at hand.

Based on the assumption that the interaction between visual modality and language can be modelled in terms of interactions of their underlying representations (McCrae, 2007) proposes such an alternative, mechanistically more transparent modelling approach: Visual context integration is achieved by incorporating visual scene knowledge into the processing stages of a constraint-based parser system. In the proposed context integration architecture visual scene information is encoded in thematic-role based event representations. Knowledge from these representations is subsequently integrated into the thematic processing stages of a weighted constraint dependency parser at constraint level. Due to a strong coupling between semantic and syntactic processing levels in the parser implementation, visual scene information can have an influence on the syntactic processing as well. For a given grammar the resultant syntax structure thus represents the global optimum with respect to syntactic, semantic and contextual constraint satisfaction.

This workshop slot is dedicated to the report of recent findings from the modelling of contextual influences upon the processing of structurally ambiguous sentences employing McCrae's context integration architecture. Discussion focus shall be on the assessment to what extent the assumptions underlying the proposed constraint-based model are cognitively adequate. Specifically, this workshop slot aims to address the following range of questions:

- How is sensory information about the visual world represented and structured in the cognitive system to support the understanding of natural language?
- What is the nature of the representations and processes involved in the interaction of vision and language?
- Which aspects of visual scene information need to be encoded representationally in an adequate model of the language-vision interface?
- To what extent is the proposed constraint-based model cognitively adequate for modelling the interactions at the language-vision interface?

References

Ferreira, F. & Tanenhaus, M. K. (2007). Introduction to the special issue on language–vision interactions. Journal of Memory and Language (57), pp 455 – 459.

Henderson, J. M. & Ferreira, F. (2004). Scene Perception for Psycholinguists. In: *The Interface of Language, Vision and Action: Eye movements and the visual world.* New York: Psychology Press, pp 1 – 58.

Knoeferle, P. S. (2005). The Role of Visual Scenes in Spoken Language Comprehension: Evidence from Eye-Tracking (PhD Thesis). Saarbrücken: Universität des Saarlandes.

Knoeferle, P. S. & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. Cognitive Science (30), pp 481 – 529.

Mayberry, M. R., Crocker, M. W., Knoeferle, P. S. (2006). A Connectionist Model of the Coordinated Interplay of Scene, Utterance, and World Knowledge. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Vancouver, Canada, July 2006.

McCrae, P. (2007). Integrating Cross-Modal Context for PP Attachment Disambiguation. *Proceedings of Third International Conference on Natural Computation*, Haikou, China. 24 - 27 August 2007. Vol. 3, pp 292 – 296. IEEE.

Tanenhaus, M., Spivey-Knowlton, M. J., Eberhard, K. M. et al. (1995). Integration of visual and linguistic information in spoken language comprehension. SCIENCE (Volume 268), 16 June 1995, pp 1632 – 1634.