An Alignment-based model of scan patterns during Visual World experiments

Michal Dziemianko & Frank Keller (University of Edinburgh)

m.dziemianko@sms.ed.ac.uk

Cross-modal perception; Visual attention; Visual world paradigm; Eye-tracking; Computational modeling; English

Linguistic processing in the context of a visual scene triggers characteristic eye-movement responses, with fixated objects reflecting, for instance, syntactic disambiguation or semantic anticipation (see e.g. Tanenhaus et al., 1995). However, existing models based on (VWP) (e.g., Mayberry et al., 2005; Roy and Mukherjee, 2005) focus on modelling phenomena such as anticipation, i.e., predicting the next word given the linguistic and visual context, rather than capturing eye-movements directly.

We present a computational model of eye-movements on a visual scene during the interpretation of a spoken sentence. The key modelling insight is that this is an alignment problem, in which objects in the scene have to be aligned with phrases in the sentence. Alignment is well-studied in computational linguistics, and a range of relevant algorithms exist. Our model consists of two stages: A Hidden Markov Model (HMM) aligns objects with phrases based on semantic roles, i.e., predicts which objects are fixated when a phrase is processed. For example, in (1), our model predicts in which order the scene objects corresponding to agent, location, target are fixated.

(1) [The boy]AGENT [will move]PREDICATE [the ball]PATIENT [in the bin]LOCATION [on the table]TARGET

A simple HMM that takes a sequence of semantic roles as input can reliably align the corresponding phrases with correct scene objects. This confirms a basic finding of VWP experiments: objects are fixated when or shortly after they are mentioned. This stage of our model is conceptually similar to Mayberry et al. (2005), who use a simple recurrent network to align words and objects. However, when syntactic or visual ambiguity are introduced (i.e., several objects can correspond to a semantic role), the HMM is prone to errors caused by variability in the data, and predicts either the correct object or its direct competitor at a given time frame, depending on the number of fixations they receive during training. We therefore extend our model to predict the probabilities with which scene objects are fixated. We apply Monte-Carlo Markov Chain (MCMC) sampling to these probabilities in order to generate sequences of fixations. This predicts human-like scan paths during VWP experiments, which models such as that of Mayberry et al. (2005) are not able to do.

We evaluate our model on three VWP datasets (Coco, 2011). The sentences include syntactic ambiguity and the scenes are referentially ambiguous, giving rise to competition between target objects and their competitors, which we can capture using the MCMC approach. The results - average similarity calculated using the Needleman-Wunsch algorithm (see e.g. Cristino et al., 2010) – summarized in the table below, indicate that MCMC sampling improves the performance considerably over an HMM baseline.

We introduced a model that predicts scan paths in VWP experiments, even in the face of syntactic and referential ambiguity. Future work includes modelling the fixation dynamics within phrases: e.g., the amount of fixations on the target increases after the onset of the noun.

Model	HMM alignment	MCMC alignment	Subject agreement
Needleman-Wunch distance	0.97 ± 0.0015	0.29 ± 0.0012	0.25 ± 0.0060

Table 1: Results for the prediction of sequences of fixated objects. Lower distance is better.

References

Coco, M. (2011). Coordination of Vision and Language in Cross-Modal Referential Processing. PhD thesis, School of Informatics (ILCC), University of Edinburgh.

Cristino, F., Mathot, S., Theeuwes, J., and Gilchrist, I. (2010). Scanmatch: A novel method for comparing fixation sequences. Behaviour Research Methods, 42:692–700.

Mayberry, M., Crocker, M., and Knoeferle, P. (2005). A connectionist model of sentence comprehension in visual worlds. In Proceedings of the 27th Annual Conference of the Cognitive Science Society.

Roy, D. and Mukherjee, N. (2005). Towards situated speech understanding: visual context priming of language models. Computer Speech & Language, 2(19):227–248.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217):1632–1634.