Grounding spoken interaction with real-time gaze in dynamic virtual environments

Matthew Crocker
Saarland University
crocker@coli.uni-saarland.de

ABSTRACT

Gaze is an important cue in visually situated dialog, grounding referring expressions to objects in the environment. We present a new technique which demonstrates that monitoring real-time listener gaze – and giving appropriate feedback – enhances reference resolution by the listener: In a 3D virtual environment, users followed directional instructions, including pressing a number of buttons that were identified using referring expression generated by the system (see GIVE; Koller et al., 2010). Gaze to the intended referent following a referring expression was taken as evidence of successful understanding and elicited positive feedback; by contrast, gaze to other objects triggered early negative feedback.

We compared this eye movement-based feedback strategy with two baseline systems, revealing that the eye-movement based feedback leads to significantly more successful button presses than the other two strategies. Our findings suggest that listener gaze immediately following a referring expression reliably indicates how a listener resolved the expression.

KEYWORDS : visually situated dialog, spoken interaction, referring expressions, eye-tracking

Proceedings of the First Workshop on Eye-tracking and Natural Language Processing, pages 1–4,
COLING 2012, Mumbai, December 2012.

Introduction

The interactive nature of dialogue entails that interlocutors are constantly anticipating what will be said next and speakers are monitoring the effects of their utterances on listeners. Gaze is an important cue in this task, providing listeners with information about the speaker's next referent (Hanna & Brennan, 2007) and offering speakers some indication about whether listeners correctly resolved their references (Clark & Krych, 2004). However, investigating listener gaze in response to spoken referring expression and, importantly, the benefit of listener gaze for the speaker, is non-trivial and requires a dynamic setting. Specifically, it requires s a shared task, a sufficiently complex environment, the systematic production of referring expressions and an appropriate reaction to listener gaze.

We present a new technique with which we successfully demonstrate that monitoring listener gaze and giving appropriate feedback enhances reference resolution by the listener. This technique employs a visually-situated, interactive natural language generation (NLG) system that exploits real-time user gaze. Users must following directional instructions, including pressing a number of buttons in the 3D environment that are identified using referring expression generated by the system, in order to find a trophy (see GIVE; Koller et al., 2010). Users' eye movements are remotely monitored for signs of referential success by mapping them to objects in the virtual environment. Gaze to the intended referent during or shortly after a referring expression is taken as evidence of successful understanding and elicits positive feedback; by contrast, gaze to other objects triggers negative feedback.

We compare this eye movement-based strategy of giving feedback with a system that generates feedback based on visibility of objects on the screen and the user's movements towards an object, as well as with a system that generates no such feedback. Performance measures reveal that the eye-movement based feedback leads to significantly more successful button presses than both the movement-based strategy and the no-feedback strategy. Further, confusion – as indicated by the overall number of requests for help – is significantly lower for eye movement-based feedback than for the two other strategies. This suggests that listener gaze between a referring expression and the intended button press indeed indicates how a listener resolved the

expression and that giving appropriate feedback can encourage or correct the listener for more efficient grounding of references.

Finally, user eye movements further reveal that the speaker's feedback to listener gaze (in contrast to movement-based feedback) generally increases looks towards all potential referents. Given that post-experiment questionnaires suggest that users did not take notice of being eye-tracked, we consider this to show that eye-movement based feedback implicitly increases visual attention to all potential targets. In conclusion, this study demonstrates that referential gaze findings from the visual world paradigm do appear to scale to dynamic and task-centered environments, and further suggest that listener gaze can be used in real-time to improve situated spoken language interaction.

References

Hanna, J. and Brennan, S. (2007) Speakers'eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57, 596-615.

Clark, H.H. and Krych, M.A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62-81.

Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., et al. (2010). The First Challenge on Generating Instructions in Virtual Environments. In E. Krahmer & M. Theune (Eds.), Empirical Methods in Natural Language Generation (pp. 337–361). Springer.