Refer-iTTS: A System for Referring in Spoken Installments to Objects in Real-World Images

Sina Zarrieß and **David Schlangen**
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
{sina.zarriess,david.schlangen}@uni-bielefeld.de

Commonly, the output of a referring expression generation system is written text which is, typically, presented to a human user as a one-shot expression. Consequently, the majority of existing REG systems interact with users in a very rigid and strictly turn-based fashion: only after the system has fully completed and delivered the result of the REG process, the user is able to read it and react accordingly. A lot of human referential communication, however, happens in situated interaction and via spoken language. Theoretically, it is well known that this change in modality fundamentally changes human production of referring expressions: Given the real-time constraints of situated interaction, a speaker often has to start uttering before she has found the optimal expression, but at the same time, she can observe the listener's reaction while speaking and extend, adapt, or correct her referring expressions accordingly (Clark and Wilkes-Gibbs, 1986; Clark and Krych, 2004). Practically, spoken and interactive REG has been rarely studied empirically or implemented in realistic systems, but see (DeVault et al., 2005; Staudte et al., 2012; Striegnitz et al., 2012; Fang et al., 2014).

We present Refer-iTTS, a system that is meant to support research on real-time spoken REG and builds upon recent approaches to REG from real-world images (Kazemzadeh et al., 2014; Zarrieß and Schlangen, 2016). We use the recently proposed words-as-classifiers (WAC) model for generation from low-level visual inputs and integrate it with InproTk (Baumann and Schlangen, 2012b), an open-source framework for incremental dialogue processing (http://wwwhomes.uni-bielefeld.de/dschlangen/inpro/). Importantly, InproTk features an incremental text-to-speech synthesis implementation (iTTS) (Baumann and Schlangen, 2012a) allowing for fine-grained, incremental manipulation of the audio signal (e.g. interruption, pausing, resumption, continuation).

We will show an interactive demonstration of the following set-up: the system presents an image with several objects in a visual scene on the screen and the user's task is to click on the object referred to. While generating and synthesizing the RE, the system continuously observes the non-verbal reactions of the user (i.e. her mouse movements) and adapts the generated utterances to these actions in an incremental fashion. At the same time, the system tries to be as cooperative as possible: if the user shows no reaction for a certain amount of time, the previous expression is expanded, i.e. the system splits its referring expression over several utterances, which is usually known as "reference in installments", cf. (Zarrieß and Schlangen, 2016).

Figure 1 illustrates the architecture of Refer-iTTS, which conceptually follows the framework of the Incremental Unit (IU) model (Schlangen and Skantze, 2009), and two example interactions. User actions and the system's generation and synthesis decisions happen concurrently, coordinated and monitored by an action manager (AM) module. Thus, besides decisions related to content planning and realization (e.g. attribute selection and ordering), a spoken installment-based REG system has to make a number of high-level decisions related to the delivery and timing of its own output. Using the Zarrieß and Schlangen (2016)'s generator, the system orders its

Proceedings of The 10th International Natural Language Generation conference, pages 72–73,
Santiago de Compostela, Spain, September 4-7 2017.

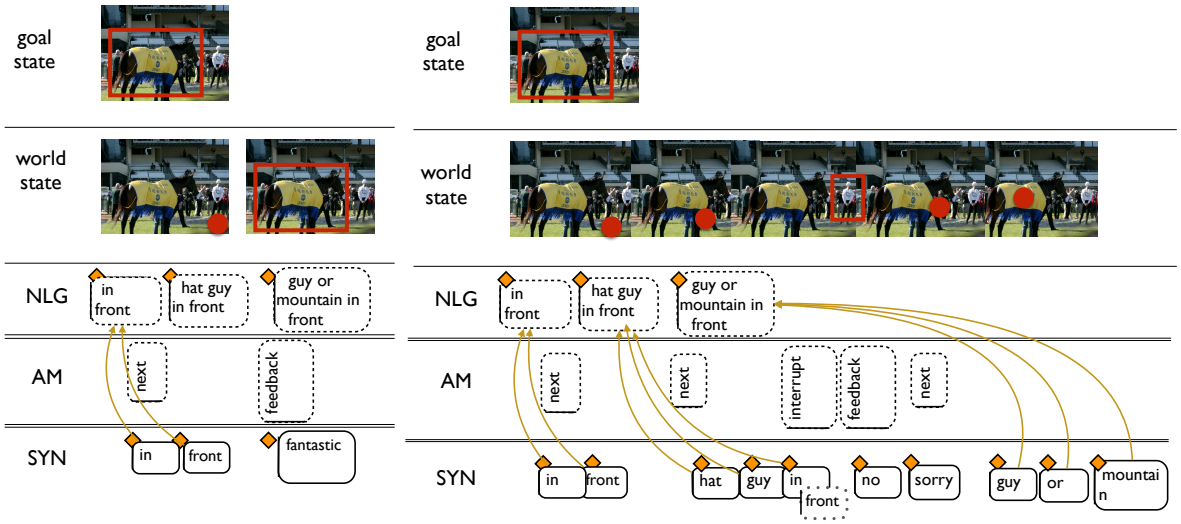

Figure 1: Two example interactions with Refer-iTTS, user actions (mouse movements and clicks are shown as red points and rectangles on the image), system decisions made by the NLG, AM (Action Manager) and SYN (synthesis) module are shown in rounded rectangles that correspond to incremental units (IUs) in InproTk, arrows between IUs indicate grounded-in links

installments according to its internal confidence, i.e. it first commits a phrase referring to the location, and then a phrase referring to the object's category. As shown in Figure 1, the Action Manager then decides when to initiate a new synthesis process for the next installment phrase (NEXT), when to interrupt the ongoing formulation of an installment phrase, e.g. in case the user clicks on an objects while the synthesis is speaking (INTERRUPT) or when to provide FEEDBACK that reacts to a user click or action. This architecture allows for highly dynamic interaction with a user.

References

Timo Baumann and David Schlangen. 2012a. INPRO_iSS: A Component for Just-In-Time Incremental Speech Synthesis. In *Proceedings of the ACL 2012 System Demonstrations*.

Timo Baumann and David Schlangen. 2012b. The InproTK 2012 release. In *Proceedings of the NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*, pages 29–32, Montreal, Canada.

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

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

David DeVault, Natalia Kariaeva, Anubha Kothari, Iris Oved, and Matthew Stone. 2005. An information-state approach to collaborative reference. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 1–4.

Rui Fang, Malcolm Doering, and Joyce Y. Chai. 2014. Collaborative Models for Referring Expression Generation in Situated Dialogue. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of EMNLP 2014*, pages 787–798, Doha, Qatar.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 710–718, Athens, Greece.

Maria Staudte, Alexander Koller, Konstantina Garoufi, and Matthew W Crocker. 2012. Using listener gaze to augment speech generation in a virtual 3d environment. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Kristina Striegnitz, Hendrik Buschmeier, and Stefan Kopp. 2012. Referring in installments: a corpus study of spoken object references in an interactive virtual environment. In *Proceedings of the 7th INLG Conference*, pages 12–16.

Sina Zarrieß and David Schlangen. 2016. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.