Incremental Generation of Visually Grounded Language in Dialogue (demonstration system)

Arash Eshghi
Interaction Lab
Heriot-Watt University
a.eshghi@hw.ac.uk

Yanchao Yu
Interaction Lab
Heriot-Watt University
y.yu@hw.ac.uk

Oliver Lemon
Interaction Lab
Heriot-Watt University
o.lemon@hw.ac.uk

We present a multi-modal dialogue system for interactive learning of perceptually grounded word meanings from a human tutor (Yu et al.,). The system integrates an incremental, semantic, and bi-directional grammar framework – Dynamic Syntax and Type Theory with Records (DS-TTR[1], (Eshghi et al., 2012; Kempson et al., 2001)) – with a set of visual classifiers that are learned throughout the interaction and which ground the semantic/contextual representations that it produces (c.f. Kennington & Schlangen (2015) where words, rather than semantic atoms, are grounded in visual classifiers). Our approach extends Dobnik et al. (2012) in integrating perception (vision in this case) and language within a single formal system: Type Theory with Records (TTR (Cooper, 2005)). The combination of deep semantic representations in TTR with an incremental grammar (Dynamic Syntax) allows for complex multi-turn dialogues to be parsed and generated (Eshghi et al., 2015). These include clarification interaction, corrections, ellipsis and utterance continuations (see e.g. the dialogue in Fig. 1).

Architecture: the system is made up of two key components – a **Vision system** and the **DS-TTR parser/generator**. The Vision system classifies a (visual) situation, i.e. deems it to be of a particular type, expressed as a TTR Record Type (RT) (see Fig. 1). This is done by deploying a set of binary attribute classifiers (Logistic Regression SVMs with Stochastic Gradient Descent, see Yu et al. (2015)) which ground the simple types (atoms) in the system (e.g. 'red', 'square'), and composing their output to construct the more complex, total type of the visual scene. This representation then acts not only as (1) the non-linguistic context of the dialogue for DS-TTR, for the resolution of e.g. definite references and indexicals (see Hough & Purver (2014)); but also (2) the logical database from which answers to questions about the objects' attributes are generated. Questions are parsed and their logical representation acts directly as a query on the non-linguistic/visual context to retrieve an answer (via *type checking* in TTR, itself done via *unification*, see Fig. 1 for a simple example). Conversely, the system can generate questions to the tutor about the attributes of objects based on the entropy of the classifiers that ground the semantic concepts, e.g. those for colour and shape. The tutor's answer then acts as a training instance for the classifiers (basic, atomic types) involved - see Fig. 1 for a snapshot of the current system.

Incremental Generation in Context: Generation (surface realisation) in DS-TTR follows exactly the same dynamics as parsing except for an additional *subsumption check* after every word against some *goal concept/context* (Purver et al., 2014). Generation is therefore just as incremental and contextual as parsing (Eshghi et al., 2015). This allows for the *generation of acceptances, elliptical utterances, short answers, and corrections, as well as continuations*. Here, it is the dialogue manager that constructs the goal concept from the semantic analysis of the visual scene, and sends it the the grammar for surface realisation – whether this is the semantics of a question, an answer, or an object description (see the system responses in Fig. 1).

[1]Downloadable from: http://sourceforge.net/projects/dylan/

Proceedings of The 9th International Natural Language Generation conference, pages 109–110,
Edinburgh, UK, September 5-8 2016.

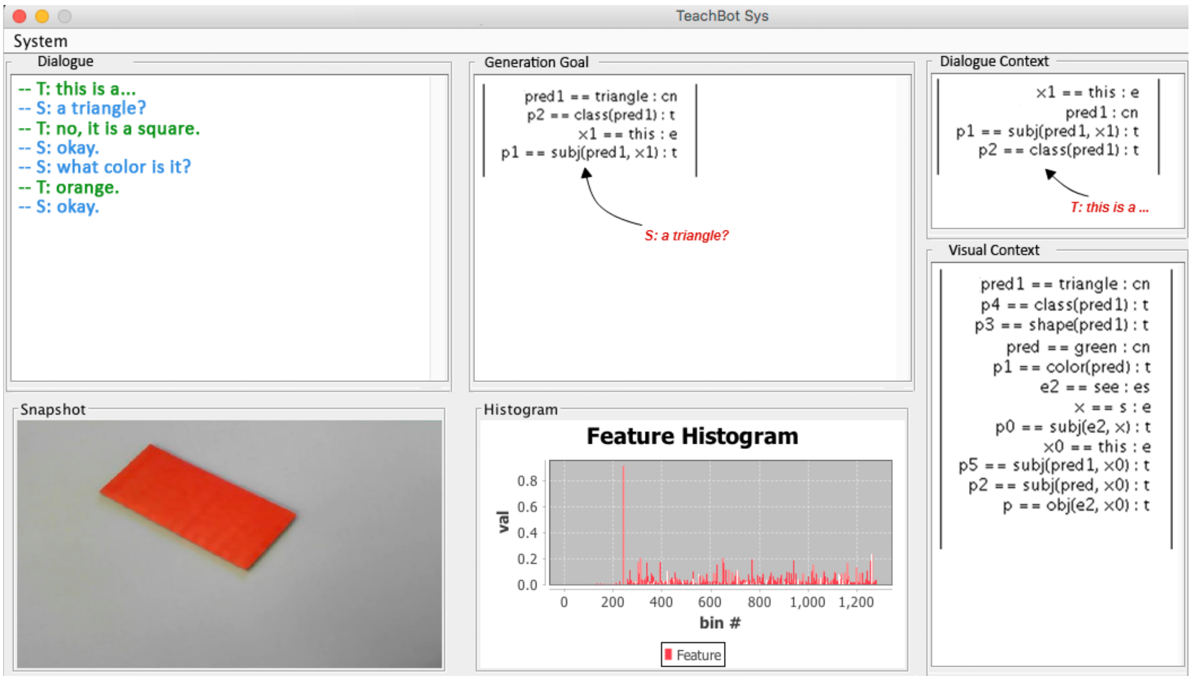

Figure 1: Incremental, visually grounded NLG in the Concept Learning System. T= tutor, S=system (screenshot)

We will show an interactive demonstration of this system at the conference, illustrating how questions, answers and object descriptions are derived and generated incrementally in real-time (Yu et al.,). Work in progress addresses: (1) more complex dialogues; (2) data-driven, incremental dialogue management at the lexical level; (3) integrating the existing DS-TTR model of incremental definite reference generation within the implemented system.

References

Robin Cooper. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2012. Modelling language, action, and perception in type theory with records. In *Proceedings of the 7th International Workshop on Constraint Solving and Language Processing (CSLPÄô12)*, pages 51–63.

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation*, volume 19, pages 325–349.

A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK. Association for Computational Linguisitics.

Julian Hough and Matthew Purver. 2014. Probabilistic type theory for incremental dialogue processing. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 80–88, Gothenburg, Sweden, April. Association for Computational Linguistics.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proc. ACL-IJCNLP*.

Matthew Purver, Julian Hough, and Eleni Gregoromichelaki. 2014. Dialogue and compound contributions. In S. Bangalore and A. Stent, editors, *Natural Language Generation in Interactive Systems*, pages 63–92. Cambridge University Press, June.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *Proc. SIGDIAL 2016*.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2015. Comparing attribute classifiers for interactive language grounding. In *Proceedings of ENMLP workshop on Vision and Language*.