An Incremental Dialogue System for Learning Visually Grounded Language (demonstration system)

Yu, Yanchao; Eshghi, Arash; Lemon, Oliver

Published in:
JerSem

Publication date:
2016

Document Version
Peer reviewed version

An Incremental Dialogue System for Learning Visually Grounded Language (demonstration system)

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

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

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

Abstract

We present a multi-modal dialogue system for interactive learning of perceptually grounded word meanings from a human tutor. 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)) 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).

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 (Yu et al., 2015)) which ground the simple types (atoms) in the system (e.g. 'red', 'square'), and composing their output to construct the 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 object 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). Conversely, the system can generate questions to the tutor (Yu et al., 2016b) 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 screenshot.

2 Learning via Incremental Dialogue

Interaction with a human tutor enables systems to take initiative to seek the particular information they need by e.g. asking questions with the highest information gain (see e.g. (Skocaj et al., 2011), and Fig. 1). For example, a robot could ask questions to learn the colour of a "square" or to request to be presented with more "red" things to improve performance. Furthermore, such systems could allow for meaning negotiation in the form of clarification interactions with the tutor.

Dialogue with the tutor continuously provides semantic information about objects in the visual scene which is then fed to online classifiers in the form of training instances. Conversely, the system can utilise the DS-TTR grammar and its existing knowledge about the world, encoded in its classifiers, to make reference to and formulate questions about the different attributes of objects identified in the visual scene.

We will show an interactive demonstration of this system, illustrating how questions, answers

[1]Download at sourceforge.net/projects/dylan/

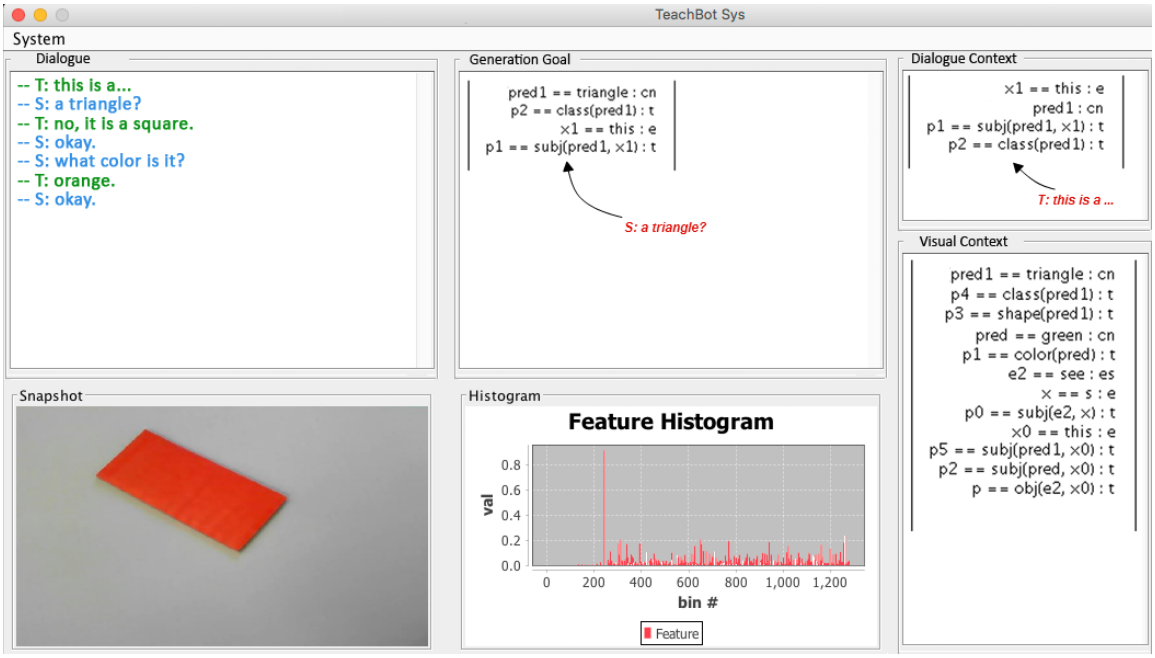

Figure 1: Incremental, visually grounded dialogue in the Concept Learning System. T= tutor, S=system

and object descriptions are derived and generated incrementally by the Concept Learner in real-time. Work in progress addresses: (1) optimising the Learner dialogue strategy (Yu et al., 2016a); (2) data-driven, incremental dialogue management at the lexical level.

Acknowledgements

This research is supported by the EPSRC, under grant number EP/M01553X/1 (BABBLE project[2]), and by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688147 (MuMMER [3]).

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 *Proc. CSLP*.

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: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349.

[2]https://sites.google.com/site/hwinteractionlab/babble

[3]http://mummer-project.eu/

A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. 2015. Feedback in conversation as incremental semantic update. In *Proc. IWCS*.

Julian Hough and Matthew Purver. 2014. Probabilistic type theory for incremental dialogue processing. In *EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 80–88.

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*.

Danijel Skocaj, Matej Kristan, Alen Vrecko, Marko Mahnic, Miroslav Janícek, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. 2011. A system for interactive learning in dialogue with a tutor. In *IROS*, pages 3387–3394.

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

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016a. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *(under review)*.

Yanchao Yu, Oliver Lemon, and Arash Eshghi. 2016b. Comparing dialogue strategies for learning grounded language from human tutors. In *Proceedings of SEMDIAL*.