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

Citation for published version:

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Peer reviewed version

Published In:
JerSem

General rights
Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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
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**


\(^2\)https://sites.google.com/site/hwinteractionlab/babble

\(^3\)http://mummer-project.eu/