Learning how to learn: grounding word meanings through conversation with humans

Oliver Lemon, Arash Eshghi, Yanchao Yu

Interaction Lab, Heriot-Watt University, Edinburgh

Human-like machines must be able to communicate their perceptions, actions, and recommendations in human-like ways, in order to coordinate actions effectively with humans. This type of ‘explainable AI’ requires us to develop systems that can dynamically ground their communication in their perceptions in ways that align with human performance, and which can learn to use human language to communicate about the world around them – for example, in this paper: *learning how to interact with human tutors in order to learn about the visual attributes of different objects.*

To achieve this goal, we are combining deep semantic parsing of Natural Language with machine learning methods to create intelligent systems which can both understand and generate human language, grounded in the context of joint tasks [6]. At the Interaction Lab[1] we are working on this overall problem in several projects: BABBLE (ESPRC) - reinforcement learning of human-like (i.e. incremental) conversational skills from data; DILIGENT (EPSRC) - imitation learning for generation of human language; MuMMER (Horizon 2020) - machine learning for socially intelligent human-robot interaction.

We will demonstrate a new trainable dialogue system for interactive learning of perceptually grounded word meanings from human tutors [6]. The system integrates both symbolic and statistical AI methods:

- an incremental deep semantic parsing and generation framework (Dynamic Syntax and Type Theory with Records, DS-TTR),
- supervised learning of visual classifiers that are updated throughout the interactions and which ground the symbolic meaning representations that DS-TTR produces,
- Reinforcement Learning (RL) to derive an optimal dialogue strategy for agents which can learn from human tutors.

We have used this system to study the effects of different dialogue policies and capabilities on accuracy of learned meanings, learning rates, and efforts/costs to the human tutor [6]. This type of problem has been studied before as ‘symbol grounding’ for robots, but did not consider ‘learning how to learn’ – our novel contributions are in combining deep semantic and incremental parsing of Natural Language dialogues with both Reinforcement Learning (RL) of a language-learner’s conversational strategies and supervised learning of word meanings.

1 Motivation

Identifying, classifying, and talking about objects or events in the surrounding environment are key capabilities for intelligent, goal-driven systems that interact with other

[1] `http://www.macs.hw.ac.uk/InteractionLab`

agents and the external world in joint actions. To this end, there has recently been a surge of interest and significant progress made on a variety of related tasks, including generation of Natural Language (NL) descriptions of images, or identifying images based on NL descriptions, e.g. [1, 5, 2].

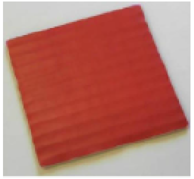
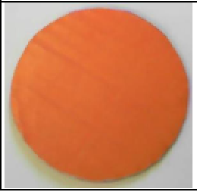
Dialogue Example		**Final Semantics in TTR**
T: what is this? S: a red circle? T: no, a red square. S: oh, okay.		$\left[\begin{array}{lll} x{=}o1 & : & e \\ p2 & : & red(x) \\ p3 & : & square(x) \end{array}\right]$
T: what can you see? S: something orange. T: what shape is it S: a square. T: no, it's a circle. S: uhu		$\left[\begin{array}{lll} x1{=}o2 & : & e \\ S{=}s & : & per \\ p & : & circle(x1) \\ p1 & : & orange(x1) \\ p2 & : & see(S,x1) \end{array}\right]$

Fig. 1. Example grounded semantics constructed in dialogues

Our goal is to build *interactive* systems that can learn grounded word meanings form humans, relating to their perceptions of real-world objects – this is different from previous work such as e.g. [4], that learns groundings from descriptions without any interaction, and more recent work using Deep Learning methods (e.g. [5]).

Most of these systems rely on training data of high quantity with no possibility of online error correction. Furthermore, they are unsuitable for robots and multimodal systems that need to continuously, and incrementally learn from the environment, and may encounter objects they haven't seen in training data. These limitations can be alleviated if systems can learn concepts, as and when needed, from situated dialogue with humans. Interaction with a human tutor also enables systems to take initiative and seek the particular information they need or lack by e.g. asking questions with the highest information gain. For example, a robot could ask questions to learn the colour of an object or to request to be presented with more "red" things to improve its performance on the concept (see e.g. Fig. 1). Furthermore, such systems allow for meaning negotiation in the form of clarification interactions with the tutor.

This setting means that the system must be *trainable from little data, compositional, adaptive, and able to handle natural human dialogue* – for instance so that it can learn visual concepts suitable for specific tasks/domains, or even those specific to a particular user. Interactive systems that learn continuously, and over the long run from humans need to do so *incrementally*, *quickly*, and *with minimal effort/cost to human tutors*.

In [6] we presented an implemented dialogue system (see Fig. 2) that integrates an incremental semantic grammar framework, suitable for dialogue processing – Dynamic Syntax and Type Theory with Records (DS-TTR[2] [3]) with visual classifiers which

[2] Download from `https://bitbucket.org/dylandialoguesystem/`

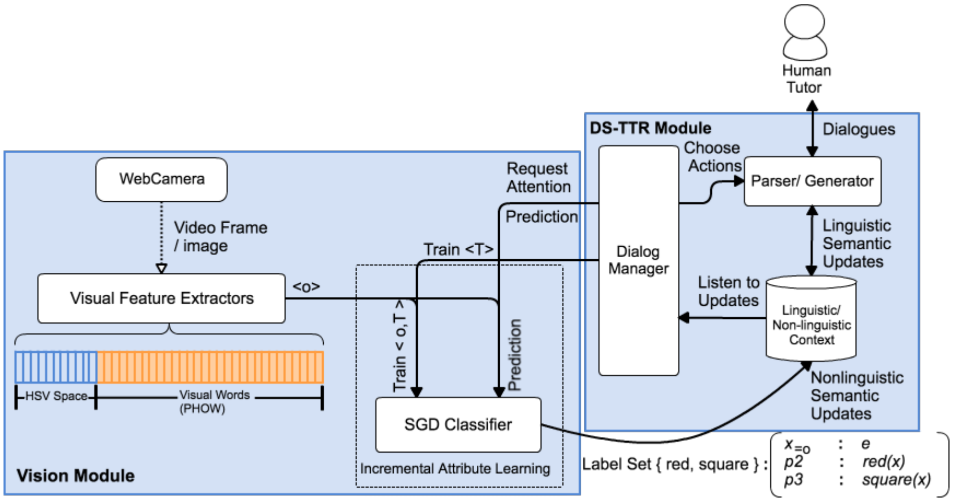


Fig. 2. Architecture of the teachable system [6]

are learned through the interaction, and which provide perceptual grounding for the semantic atoms in the representations (Record Types in TTR) produced by the parser.

We have trained this system (using RL methods) in interaction with simulated human tutors to test hypotheses about how the accuracy of learned meanings, learning rates, and the overall cost/effort for the tutors are affected by different dialogue policies and capabilities: (1) who takes *initiative* in the dialogues; (2) the agent's ability to utilise their level of *uncertainty* about an object's attributes; and (3) their ability to process *elliptical as well as incrementally constructed dialogue turns*. The results show [6] that differences along these dimensions have significant impact both on the accuracy of the grounded word meanings that are learned, and the effort required by the tutors.

References

1. Bruni, E., Tran, N.K., Baroni, M.: Multimodal distributional semantics. J. Artif. Intell. Res.(JAIR) 49(1–47) (2014)
2. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR (2009)
3. Kempson, R., Meyer-Viol, W., Gabbay, D.: Dynamic Syntax: The Flow of Language Understanding. Blackwell (2001)
4. Roy, D.: A trainable visually-grounded spoken language generation system. In: Proceedings of the International Conference of Spoken Language Processing (2002)
5. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics 2, 207–218 (2014)
6. Yu, Y., Eshghi, A., Lemon, O.: Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In: Proceedings of SIGDIAL (2016)