Learning how to learn: an adaptive dialogue agent for incrementally learning visually grounded word meanings

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 an optimised multi-modal dialogue agent for interactive learning of visually grounded word meanings from a human tutor, trained on real human-human tutoring data. Within a life-long interactive learning period, the agent, trained using Reinforcement Learning (RL), must be able to handle natural conversations with human users, and achieve good learning performance (i.e. accuracy) while minimising human effort in the learning process. We train and evaluate this system in interaction with a simulated human tutor, which is built on the BURCHAK corpus – a Human-Human Dialogue dataset for the visual learning task. The results show that: 1) The learned policy can coherently interact with the simulated user to achieve the goal of the task (i.e. learning visual attributes of objects, e.g. colour and shape); and 2) it finds a better trade-off between classifier accuracy and tutoring costs than hand-crafted rule-based policies, including ones with dynamic policies.

1 Introduction

As intelligent systems/robots are brought out of the laboratory and into the physical world, they must become capable of natural everyday conversation with their human users about their physical surroundings. Among other competencies, this involves the ability to learn and adapt mappings between words, phrases, and sentences in Natural Language (NL) and perceptual aspects of the external environment – this is widely known as *the grounding problem*.

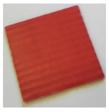
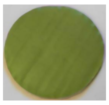
Image	Human-Human Dialogue
	T(utor): do you know this object? L(earner): a suzuli ... wait no ... sako wakaki? T: the color is right, but the shape is not. L: oh, okay, so? T: a burchak, burchak, sako burchak. L: cool, got it.
	L: what is this? T: en ... a aylana suzili. L: is aylana for color? T: no, it's a shape. L: so it is an suzili aylana, right? T: yes.

Figure 1: Human-Human Example Dialogues in the BURCHAK Corpus (Yu et al., 2017)
('sako' for 'red', 'burchak' for 'square', 'suzuli' for 'green', 'aylana' for 'circle', 'wakaki' for 'triangle')

The grounding problem can be categorised into two distinct, but interdependent types of problem: 1) agent as a second-language learner: the agent needs to learn to ground (map) NL symbols onto their existing perceptual and lexical knowledge (e.g. a dictionary of pre-trained classifiers) as in e.g. Silberer and Lapata (2014); Thomason et al. (2016); Kollar et al. (2013); Matuszek et al. (2014); and 2) the agent as a child: without any prior knowledge of perceptual categories, the agent must learn both the perceptual categories themselves and also how NL expressions map to these (Skocaj et al., 2016; Yu et al., 2016c). Here, we concentrate on the latter scenario, where a system learns to identify and describe visual attributes (colour and shape in this case) through interaction with human tutors, incrementally, over time.

Previous work has approached the grounding problem using a variety of resources and approaches, for instance, either using annotated visual datasets (Silberer and Lapata, 2014; Socher et al., 2014; Naim et al., 2015; Al-Omari et al., 2016; Tellex et al., 2014; Matuszek et al., 2012, 2014), or through interactions with other agents or real humans (Kollar et al., 2013; Tellex et al., 2013; Thomason et al., 2015, 2016; Skocaj et al., 2016; Yu et al., 2016c), where feedback from other

Proceedings of the First Workshop on Language Grounding for Robotics, pages 10–19,
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics

agents is used to learn new concepts.

However, most of these systems, which ground NL symbols through interaction have two common, important drawbacks: 1) in order to achieve better performance (i.e. high accuracy), these systems require a high level of human involvement – they always request feedback from human users, which might affect the quality of human answers and decrease the overall user experience in a lifelong learning task; 2) Most of these approaches are not built/trained based on real human-human conversations, and therefore can't handle them. Natural human dialogue is generally more messy than either machine-machine or human-machine dialogue, containing natural dialogue phenomena that are notoriously difficult to capture, e.g. *self- corrections, repetitions and restarts, pauses, fillers, interruptions, and continuations* (Purver et al., 2009; Hough, 2015). Furthermore, they often exhibit much more variation than in their synthetic counterparts (see dialogue examples in Fig. 1).

In order to cope with the first problem, recent prior work (Yu et al., 2016b,c) has built multimodal dialogue systems to investigate the effects of different dialogue strategies and capabilities on the overall learning performance. Their results have shown that, in order to achieve a good trade-off between learning performance and human involvement, the agent must be able to take initiative in dialogues, take into account uncertainty of its predictions, as well as cope with natural human conversation in the learning process. However, their systems are built based on hand-crafted, synthetic dialogue examples rather than real human-human dialogues.

In this paper, we extend this work to introduce an adaptive visual-attribute learning agent trained using Reinforcement Learning (RL). The agent, trained with a multi-objective policy, is capable not only of properly learning novel visual objects/attributes through interaction with human tutors, but also of efficiently minimising human involvement in the learning process. It can achieve equivalent/comparable learning performance (i.e. accuracy) to a fully-supervised system, but with less tutoring effort. The dialogue control policy is trained on the BURCHAK Human-Human Dialogue dataset (Yu et al., 2017), consisting of conversations between a human 'tutor' and a human 'learner' on a visual attribute learning task. The dataset includes a wide range of natural, *incremental* dialogue phenomena (such as overlapping turns, self-correction, repetition, fillers, and continuations), as well as considerable variation in the dialogue strategies used by the tutors and the learners.

Here we compare the new optimised learning agent to rule-based agents with and without adaptive confidence thresholds (see section 3.2.1). The results show that the RL-based learning agent outperforms the rule-based systems by finding a better trade-off between learning performance and the tutoring effort/cost.

2 Related Work

In this section, we review some of the work that has addressed the language grounding problem generally. The problem of grounding NL in perception has received very considerable attention in the computational literature recently. On the one hand, there is work that only addresses the grounding problem implicitly/indirectly: in this category of work is the large literature on image and video captioning systems that learn to associate an image or video with NL descriptions (Silberer and Lapata, 2014; Bruni et al., 2014; Socher et al., 2014; Naim et al., 2015; Al-Omari et al., 2016). This line of work uses various forms of neural modeling to discover the association between information from multiple modalities. This often works by projecting vector representations from the different modalities (e.g. vision and language) into the same space in order to retrieve one from the other. Importantly, these models are holistic in that they learn to use NL symbols in specific tasks without any explicit encoding of the symbol-perception link, so that this relationship remains implicit and indirect.

On the other hand, other models assume a much more explicit connection between symbols (either words or predicate symbols of some logical language) and perceptions (Kennington and Schlangen, 2015; Yu et al., 2016c; Skocaj et al., 2016; Dobnik et al., 2014; Matuszek et al., 2014). In this line of work, representations are both compositional and transparent, with their constituent atomic parts grounded individually in perceptual classifiers. Our work in this paper is in the spirit of the latter.

Another dimension along which work on grounding can be compared is whether groundings are learned offline (e.g. from images or videos annotated with descriptions or definite reference expressions as in (Kennington and Schlangen, 2015; Socher et al., 2014)) or from live interaction as in, e.g. (Skocaj et al., 2016; Yu et al., 2015, 2016c; Das et al., 2017, 2016; de Vries et al., 2016; Thomason et al., 2015, 2016; Tellex et al., 2013). The latter, which we do here, is clearly more appropriate for multimodal systems or robots that are expected to continuously, and incrementally learn from the environment and their users.

Multi-modal, interactive systems that involve grounded language are either: (1) *rule-based* as in e.g. Skocaj et al. (2016); Yu et al. (2016b); Thomason et al. (2015, 2016); Tellex et al. (2013); Schlangen (2016): in such systems, the dialogue control policy is hand-crafted, and therefore these systems are *static*, cannot adapt, and are less robust; or (2) *optimised* as in e.g. Yu et al. (2016c); Mohan et al. (2012); Whitney et al. (fcmng); Das et al. (2017): in contrast such systems are learned from data, and live interaction with their users; they can thus *adapt* their behaviour dynamically not only to particular dialogue histories, but also to the specific information they have in another modality (e.g. a particular image or video).

Ideally, such interactive systems ought to be able to handle natural, spontaneous human dialogue. However, most work on interactive language grounding learn their systems from synthetic, hand-made dialogues or simulations which lack both in variation and the kinds of dialogue phenomena that occur in everyday conversation; they thus lead to systems which are not robust and cannot handle everyday conversation (Yu et al., 2016c; Skocaj et al., 2016; Yu et al., 2016a). In this paper, we try to change this by training an adaptive learning agent from *human-human dialogues in a visual attribute learning task.*

Given the above, what we achieve here is: we have trained an adaptive attribute-learning dialogue policy from realistic human-human conversations that learns to optimise the trade-off between a learning/grounding performance (*Accuracy*) and costs form human tutors,in effect doing a form of active learning.

3 Learning How to Learn Visual Attributes: an Adaptive Dialogue Agent

We build a multimodal and teachable system that supports a visual attribute (e.g. colour and shape) learning process through natural conversational interaction with human tutors (see Fig. 1 for example dialogues), where the tutor and the learner interactively exchange information about the visual attributes of an object they can both see. Here we use Reinforcement Learning for policy optimisation for the learner side (see below Section 3.2). The tutor side is simulated in a data-driven fashion using human-human dialogue data (see below, Sections 4 & 5.2).

3.1 Overall System Architecture

The system architecture loosely follows that of Yu et al. (2016c), and employs two core modules:

Vision Module produces visual attribute predictions, using two base feature categories, i.e. the HSV colour space for colour attributes, and a 'bag of visual words' (i.e. PHOW descriptors) for the object shapes/class. It consists of a set of binary classifiers - Logistic Regression SVM classifiers with Stochastic Gradient Descent (SGD) (Zhang, 2004) – to incrementally learn attribute predictions. The visual classifiers ground visual attribute words such as 'red', 'circle' etc. that appear as parameters of the Dialogue Acts used in the system.

Dialogue Module that implements a dialogue system with a classical architecture, composed of Dialogue Management (DM), Natural Language Understanding (NLU) and Generation (NLG) components. The components interact via Dialogue Act representations (e.g. `inform(color=red),ask(shape)`). It is these action representations that are grounded in the visual classifiers that reside in the vision module. The DM relies on an adaptive policy that is learned using RL. The policy is trained to: 1) handle natural interactions with humans and to produce coherent dialogues; and 2) optimise the trade-off between accuracy of visual classifiers and the cost of the dialogue to the tutor.

3.2 Adaptive Learning Agent with Hierarchical MDP

Given the visual attribute learning task, the smart agent must learn novel visual objects/attributes as accurately as possible through natural interactions with real humans, but meanwhile it should attempt to minimise the human involvement as much as possible in this life-long learning process. We formulate this interactive learning task into two sub-tasks, which are trained using Reinforcement Learning with a hierarchical Markov

Decision Process (MDP), consisting of two interdependent MDPs (sections 3.2.1 and 3.2.2):

3.2.1 Adaptive Confidence Threshold

Following previous work (Yu et al., 2016c), we also here use a positive confidence threshold: this is a threshold which determines when the agent believes its own predictions. This threshold plays an essential role in achieving the trade-off between the learning performance and the tutoring cost, since the agent's behaviour, e.g. whether to seek feedback from the tutor, is dependent on this threshold. A form of *active learning* is taking place: the learner only asks a question about an attribute if it isn't confident enough already about that attribute.

Here, we learn an adaptive strategy that aims at maximising the overall learning performance simultaneously, by properly adjusting the positive confidence threshold in the range of 0.65 to 0.95. We train the optimization using a RL library – Burlap (MacGlashan, 2015) as follows, in detail:

State Space The adaptive-threshold MDP initialises a 3-dimensional state space defined by $Num_{Instance}$, $Threshold_{cur}$, and $deltaAcc$, where $Num_{Instance}$ represents how many visual objects/images have been seen (the number of instances will be clustered into 50 bins, each bin contains 10 visual instances); $Threshold_{cur}$ represents the positive threshold the agent is currently applying; and $deltaAcc$ represents, after seeing each 10 instances, whether the classifier accuracy increases, decreases or keep constant comparing to the previous bin. The $deltaAcc$ is configured into three levels, (see Eq.1)

$$deltaAcc = \begin{cases} 1, & \text{if } \Delta Acc > 0 \\ 0, & \text{else if } \Delta Acc = 0 \\ -1, & \text{otherwise} \end{cases} \tag{1}$$

Action Selection the actions were either to increase or decrease the confidence threshold by 0.05, or keep it the same.

Reward signal The reward function for the learning tasks is given by a local function R_{local}. This local reward signal was directly proportional to the agents delta accuracy over the previous Learning Step (10 training instances, see above). The single training episode will be terminated once the agent goes through 500 instances.

3.2.2 Natural Interaction

The second sub-task aims at learning an optimised dialogue strategy that allows the system to achieve the learning task (i.e. learn new visual attributes) through natural, human-like conversations.

State Space The dialogue agent initialises a 4-dimensional state space defined by (C_{state}, S_{state}, $preDAts$, $preContext$), where C_{state} and S_{state} are the status of visual predictions for the colour and shape attributes respectively (where the status is determined by the prediction score ($conf.$) and the adaptive confidence threshold ($posThd.$) described above (see Eq.2)), the $preDAts$ represents the previous dialogue actions from the tutor response, and the $preContext$ represents which attribute categories (e.g. colour, shape or both) were talked about in the context history.

$$State = \begin{cases} 2, & \text{if } conf. \geq posThd \\ 1, & \text{else if } 0.5 < conf. < posThd. \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

i.e. C_{state} or S_{state} will be updated to 2 also when the related knowledge has been provided by the tutor.

Action Selection The actions were chosen based on the statistics of the dialog action frequency occurred from the BURCHAK corpus, including *question-asking(for WH questions or polar questions)*, *inform*, *acknowledgment*, as well as *listening*. These actions can be applied for either specific single attribute or both. The action of *inform* can be separated into two sub-actions according to whether the prediction score is greater than 0.5 (i.e. *polar question*) or not (i.e. *doNotKnow*).

Reward signal The reward function for the learning tasks is given by a global function R_{global} (see Eq.3). The dialogue will be terminated when both colour and shape knowledge are either taught by human tutors or known with high confidence scores.

$$R_{global} = 10 - C_{ost} - penal.; \tag{3}$$

where C_{ost} represents the cumulative cost by the tutor (see more details about this setup in Section 5.1) in a single dialogue, and $penal.$ penalizes all performed actions which cannot respond to the user properly.

Dialogue Capability	Speaker	Annotation Tag
Listen	Tutor/Learner	Listen()
Inform	Tutor/Leaner	Inform(colour:sako&shape:burchak)
Question_asking	Tutor/Leaner	Ask(colour), Ask(shape), Ask(colour&shape)
Question-answering	Tutor/Leaner	Inform(colour:sako), Polar(shape:burchak)
Acknowledgement	Tutor/Learner	Ack(), Ack(colour)
Rejection	Tutor	Reject(), Reject(shape)
Focus	Tutor	Focus(colour), Focus(shape)
Clarification	Tutor	CLr()
Clarification-request	Learner	CLrRequest()
Help-offer	Tutor	Help()
Help-request	Learner	HelpRequest()
Checking	Tutor	Check()
Repetition-request	Tutor	Repeat()
Retry-request	Tutor	Retry()

Table 1: List of Dialogue Capabilities/Actions and Corresponding Annotations in the Corpus

i.e. we applied the **SARSA algorithm** (Sutton and Barto, 1998) for learning the multi-MDP learning agent with each episode defined as a complete dialogue for an object. It was configured with a $\xi-$Greedy exploration rate of 0.2 and a discount factor of 1.

4 Human-Human Dialogue Corpus: BURCHAK

BURCHAK (Yu et al., 2017) is a freely available Human-Human Dialogue dataset consisting of 177 dialogues between real human users on the task of interactively learning visual attributes.

The DiET experimental toolkit These dialogue were collected using a new *incremental variation* of the DiET chat-tool developed by (Healey et al., 2003; Mills and Healey, submitted), which allows two or more participants to communicate in a shared chat window. It supports live, fine-grained and highly local experimental manipulations of ongoing human-human conversation (see e.g. (Eshghi and Healey, 2015)). The chat-tool is designed to support, elicit, and record at a fine-grained level, dialogues that resemble face-to-face dialogue in that turns are: (1) constructed and displayed incrementally as they are typed; (2) transient; (3) potentially overlapping; (4) not editable, i.e. deletion is not permitted.

Task The learning/tutoring task given to the participants involves a pair of participants who talk about visual attributes (e.g. colour and shape) through a series of visual objects. The overall goal of this task is for the learner to discover groundings between visual attribute words and aspects in the physical world through interaction. However, since humans have already known all groundings, such as “red” and “square”, the task is assumed in a second-language learning scenario, where each visual attribute, instead of standard English words, is assigned to a new unknown word in a made-up language (see examples in Fig. 1). (see more details in (Yu et al., 2017))

Dialogue Phenomena As the chat-tool is designed to resemble face-to-face dialogue, the most important challenge of this BURCHAK is that it refers to a wide range of natural, incremental dialogue phenomena, such as overlapping, self-correction and repetition, filler as well as continuation (Fig. 1). On the other hand, BURCHAK, which focuses on the visual attribute learning task, offers a list of interesting task-oriented dialogue strategies (e.g. initiative, context-dependency and knowledge-acquisition) and capabilities, such as inform, question-asking and answering, listen (no act), as well as acknowledgement and rejection. Each dialogue action contains a huge variations in the realistic conversation. All dialogue actions are tagged in the dataset (as shown in Table 1).

i.e. we have trained and evaluated the optimised learning agents on the cleaned-up version of this corpus, in which spelling mistakes, emoticons, as well as some snippets of conversations where the participant misunderstood the task have been corrected or removed.

5 Experiment Setup

In this section, we follow previous work (Yu et al., 2016c) to compare the trained RL-based learning agent with a rule-based system with the best performance (i.e. an agent which takes the initiative in dialogues, takes into account its changing confidence about its predictions, and is also able to process natural, human-like dialogues) from previous work. Instead of using hand-crafted dialogue examples as before, both the RL-based system and the rule-based system are trained/developed against a simulated user, itself trained from the BURCHAK dialogue data set as above. For learning simple visual attributes (e.g. *"red"* and *"square"*), we use the same hand-made visual object dataset from Yu et al. (2016c).

In order to further investigate the effects of the optimised adaptive confidence threshold on the learning performance, we build the rule-based system under three different settings, i.e. with a constant threshold (0.95) (see *blue* curve in Fig. 2), with a hand-crafted adaptive threshold which drops by 0.05 after each 10 instances (*grey* curve in Fig. 2), and with a hand-crafted adaptive threshold which drops by 0.01 after each 10 instances (*orange* curve in Fig. 2).

5.1 Evaluation Metrics

To compare the optimised and the rule-based learning agents, and also further investigate how the adaptive threshold affect the learning process, we follows the evaluate metrics from the previous work (see (Yu et al., 2016c)) considering both the cost to the tutor and the accuracy of the learned meanings, i.e. the classifiers that ground our colour and shape concepts.

Cost The cost measure reflects the effort needed by a human tutor in interacting with the system. Skocaj et. al. (2009) point out that a comprehensive teachable system should learn as autonomously as possible, rather than involving the human tutor too frequently. There are several possible costs that the tutor might incur: C_{inf} refers to the cost $(i.e.\ 5\, points)$ of the tutor providing information on a single attribute concept (e.g. "this is red" or "this is a square"); C_{ack} is the cost $(i.e.\ 0.5)$ for a simple confirmation (like "yes", "right") or rejection (such as "no"); C_{crt} is the cost of correction for a single concept (e.g. "no, it is blue" or "no, it is a circle"). We associate a higher cost $(i.e.\ 5)$ with correction of statements than that of polar questions. This is to penalise the learning agent when it confidently makes a false statement – thereby incorporating an aspect of trust in the metric (humans will not trust systems which confidently make false statements).

i.e. differently to the previous evaluation metrics, we do not take into account the costs of parsing and producing utterances

Learning Performance As mentioned above, an efficient learner dialogue policy should consider both classification accuracy and tutor effort (Cost). We thus define an integrated measure – the *Overall Performance Ratio* (R_{perf}) – that we use to compare the learner's overall performance across the different conditions:

$$R_{perf} = \frac{\Delta Acc}{C_{tutor}}$$

i.e. the increase in accuracy per unit of the cost, or equivalently the gradient of the curve in Fig. 2c. We seek dialogue strategies that maximise this.

5.2 User Simulation

In order to train and evaluate these learning agents, we build an user simulation using a generic n-gram framework (see (Yu et al., 2017)) on the BURCHAK corpus. This user framework takes as input the sequence of N most recent words in the dialogue, as well as some optional additional conditions, and then outputs the next user response on multiple levels as required, e.g. full utterance, a sequence of dialogue actions, or even a sequence of single word outputs for incremental dialogue. Differently to other existing user simulations, this framework aims at not only resembling user strategies and capabilities in realistic conversations, but also at simulating incremental dialogue phenomena, e.g. self-repair and repetition, and pauses, as well as fillers. In this paper, we created an action-based user model that predict the next user response in a sequence of dialogue actions. The simulator then produces a full utterance by following the statistics of utterance templates for each predicted action.

5.3 Results

Table 2 shows example interactions between the learned RL agent and the simulated tutor on the learning task. The dialogue agent learned to take the initiative and constantly produces coherent conversations through the learning process.

Dialogue Example (a)
T: what is this object called?
L: a red square?
T: the shape is correct, but the colour is wrong.
L: so what colour is this?
T: green.
L: okay, get it.
Dialogue Example (b)
L: blue?
T: yes, blue is for the colour. and shape?
L: sorry, i don't know the shape.
T: the shape is circle.
L: okay, got it.

Table 2: User Simulation Examples for *(a) Tutor takes the initiative (b) Learner takes the initiative*

Fig. 2a and 2b plot the progression of average Accuracy and (cumulative) Tutoring Cost for each of the 4 learning agents in our experiment, as the system interacts over time with the tutor about each of the 500 training instances.

As noted in passing, the vertical axes in these graphs are based on averages across the 20 folds - recall that for Accuracy the system was tested, in each fold, at every learning step, i.e. after every 10 training instances.

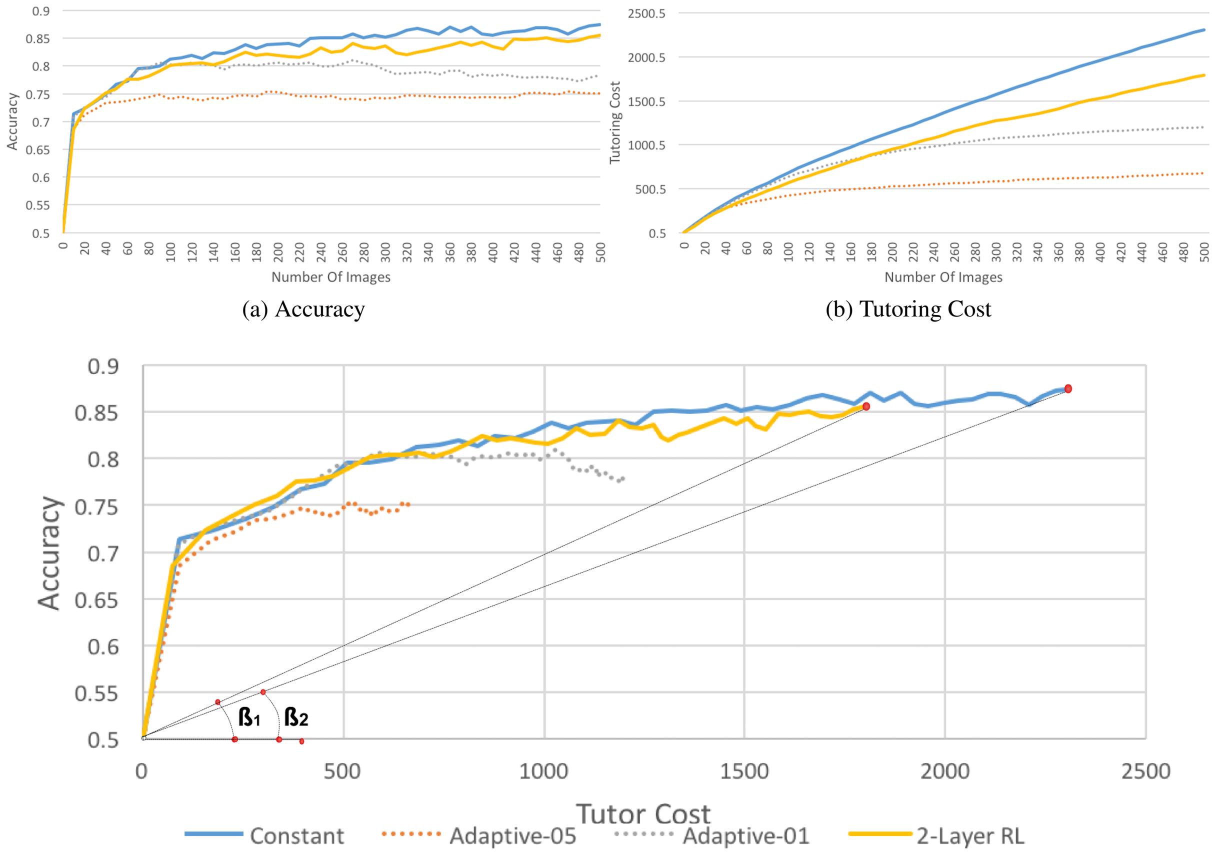
Fig. 2c, on the other hand, plots Accuracy against Tutoring Cost directly. Note that it is to be expected that the curves should not terminate in the same place on the x-axis since the different conditions incur different total costs for the tutor across the 500 training instances. The gradient of this curve corresponds to *increase in Accuracy per unit of the Tutoring Cost*. It is the gradient of the line drawn from the beginning to the end of each curve ($tan(\beta)$ on Fig. 2c) that constitutes our main evaluation measure of the system's overall performance in each condition, and it is this measure for which we report statistical significance results: there are significant differences in accuracy between the RL-based policy and two rule-based policies with the hand-crafted threshold ($p < 0.01$ for both). The RL-based policy shows significantly less tutoring cost than the rule-based system with a constant threshold ($p < 0.01$). The mean gradient of the yellow, RL curve is actually slightly higher than the constant-threshold policy blue curve - discussed below.

5.4 Discussion

Accuracy As can be seen in Fig. 2a, the rule-based system with a constant threshold (0.95) shows the fastest increase in accuracy and finally reaches around 0.87 at the end of the learning process (i.e. after seeing 500 instances) – the blue curve. Both systems with a hand-crafted adaptive threshold, with an incremental decrease of 0.01 (grey curve) and 0.05 (orange curve), have shown an unexpected trend in accuracy across 500 instances, where the orange curve flattens out at about 0.76 after seeing only 50 instances, and the grey curve shows a good increase in the beginning but later drops down to about 0.77 after 150 instances. This is because the thresholds were decreased too fast, so that the agent cannot hear enough feedback (i.e. corrective attribute labels) from tutors to improve its predictions. In contrast to this, the optimised RL-based agent achieves much better accuracy (i.e. about 0.85) by the end of the experiment.

Tutoring Cost As mentioned above, there is a form of *active learning* taking place in the experiment: the agent can only hear feedback from the tutor if it is not confident enough about its own predictions. This also explains the slight decrease in the gradients of the curves (i.e. the cumulative cost for the tutor) (see Fig. 2b) as the agent is exposed to more and more training instances: its subjective confidence about its own predictions increases over time, and thus there is progressively less need for tutoring. In detail, the tutoring cost progresses much more slowly while the system was applying a hand-crafted adaptive threshold (i.e. incrementally decreases by either 0.01 or 0.05 after each bin). This is still because there were not interactions taking place at all once the threshold is lower than a certain value (for instance, 0.65), where the agent might be highly confident on all its predictions. In contrast, the RL-based agent shows a faster progress in the cumulative tutoring cost, but achieves higher accuracy.

Overall Performance Here, we only compare the gradients of the curves between the optimised learning agent (yellow curve) and the rule-based system with a constant threshold (blue curve) in Fig. 2c, because others with the incremental decreased threshold cannot achieve an acceptable learning performance. The agent with an adaptive threshold (yellow) achieves slightly better overall gradient ($tan(\beta_1)$) than the rule-based system



(a) Accuracy

(b) Tutoring Cost

(c) Overall Performance

Figure 2: Evolution of Learning Performance

($tan(\beta_2)$), it achieves a comparable accuracy and does it faster. We therefore conclude that the optimised learning agent, which finds a better trade-off between the learning accuracy and the tutoring cost, is more desirable.

6 Conclusion & Future Work

We have introduced a multi-modal learning agent that can incrementally learn grounded word meanings through interaction with human tutors over time, and deploys an *adaptive* dialogue policy (optimised using Reinforcement Learning). We applied a human-human dialogue dataset (i.e. BURCHAK) to train and evaluate the optimised learning agent. We evaluated the system by comparing it to a rule-based system, and results show that: 1) the optimised policy has learned to coherently interact with the simulated user to learn visual attributes of an object (e.g. colour and shape); 2) it achieves comparable learning performance to a rule-based systems, but with less tutoring effort needed from humans.

Ongoing work further applies Reinforcement Learning at the word level to learn a complete, incremental dialogue policy, i.e. which chooses system output at the lexical level (Eshghi and Lemon, 2014; Kalatzis et al., 2016). In addition, instead of acquiring visual concepts for toy objects (i.e. with simple colour and shape), the system has recently been extended to interactively learn about real object classes (e.g. `shampoo`, `apple`). The latest system integrates with a *Self-Organizing Incremental Neural Network* and a deep *Convolutional Neural Network* to learn object classes through interaction with humans incrementally, over time.

Acknowledgements

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

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

[2] `http://mummer-project.eu/`

References

Muhannad Al-Omari, Eris Chinellato, Yiannis Gatsoulis, David C. Hogg, and Anthony G. Cohn. 2016. Unsupervised grounding of textual descriptions of object features and actions in video. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016.*. pages 505–508.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)* 49(1–47).

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *CoRR* abs/1611.08669.

Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. *CoRR* abs/1703.06585.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2016. Guesswhat?! visual object discovery through multi-modal dialogue. *CoRR* abs/1611.08481.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2014. Type theory with records: a general framework for modelling spatial language. In *Proceedings of The Second Workshop on Action, Perception and Language (APL'2)*.

Arash Eshghi and Patrick G. T. Healey. 2015. Collective contexts in conversation: Grounding by proxy. *Cognitive Science* pages 1–26.

Arash Eshghi and Oliver Lemon. 2014. How domain-general can we be? learning incremental dialogue systems without dialogue acts. In *Proceedings of SemDial*.

P. G. T. Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. Boston, Massachusetts.

Julian Hough. 2015. *Modelling Incremental Self-Repair Processing in Dialogue*. Ph.D. thesis, Queen Mary University of London.

Dimitrios Kalatzis, Arash Eshghi, and Oliver Lemon. 2016. Bootstrapping incremental dialogue systems: using linguistic knowledge to learn from minimal data. *CoRR* abs/1612.00347. http://arxiv.org/abs/1612.00347.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. pages 292–301.

Thomas Kollar, Jayant Krishnamurthy, and Grant Strimel. 2013. Toward interactive grounded language acqusition. In *Robotics: Science and Systems*.

James MacGlashan. 2015. Burlap http://burlap.cs.brown.edu/.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*. pages 2556–2563.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In *Proc. of the 2012 International Conference on Machine Learning*. Edinburgh, Scotland.

Gregory J. Mills and Patrick G. T. Healey. submitted. The Dialogue Experimentation toolkit. *xx* (?).

Shiwali Mohan, Aaron Mininger, James Kirk, and John E. Laird. 2012. Learning grounded language through situated interactive instruction. In *Robots Learning Interactively from Human Teachers, Papers from the 2012 AAAI Fall Symposium, Arlington, Virginia, USA, November 2-4, 2012*.

Iftekhar Naim, Young Chol Song, Qiguang Liu, Liang Huang, Henry A. Kautz, Jiebo Luo, and Daniel Gildea. 2015. Discriminative unsupervised alignment of natural language instructions with corresponding video segments. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. pages 164–174.

Matthew Purver, Raquel Fernández, Matthew Frampton, and Stanley Peters. 2009. Cascaded lexicalised classifiers for second-person reference resolution. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*. Association for Computational Linguistics, London, UK, pages 306–309. http://www.dcs.qmul.ac.uk/ mpurver/papers/purver-et-al09sigdial-you.pdf.

David Schlangen. 2016. Grounding, justification, adaptation: Towards machines that mean what they say. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers). Association for Computational Linguistics, Baltimore, Maryland, volume 1, pages 721–732.

Danijel Skocaj, Alen Vrecko, Marko Mahnic, Miroslav Janícek, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Jeremy L. Wyatt, Thomas Keller, Kai Zhou, Michael Zillich, and Matej Kristan. 2016. An integrated system for interactive continuous learning of categorical knowledge. *J. Exp. Theor. Artif. Intell.* 28(5):823–848. https://doi.org/10.1080/0952813X.2015.1132268.

Danijel Skočaj, Matej Kristan, and Aleš Leonardis. 2009. Formalization of different learning strategies in a continuous learning framework. In *Proceedings of the Ninth International Conference on Epigenetic Robotics; Modeling Cognitive Development in Robotic Systems*. Lund University Cognitive Studies, pages 153–160.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2:207–218.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: an Introduction*. MIT Press.

Stefanie Tellex, Pratiksha Thaker, Joshua Mason Joseph, and Nicholas Roy. 2014. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning* 94(2):151–167. https://doi.org/10.1007/s10994-013-5383-2.

Stefanie Tellex, Pratiksha Thakerll, Robin Deitsl, Dimitar Simeonovl, Thomas Kollar, and Nicholas Royl. 2013. Toward information theoretic human-robot dialog. *Robotics: Science and Systems* page 409.

Jesse Thomason, Jivko Sinapov, Maxwell Sevtlik, Peter Stone, and Raymond J. Mooney. 2016. Learning multi-modal grounded linguistic semantics by playing ”i spy”. In *To Appear: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-16, New York City, USA, July 9-15, 2016*.

Jesse Thomason, Shiqi Zhang, Raymond J. Mooney, and Peter Stone. 2015. Learning to interpret natural language commands through human-robot dialog. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. pages 1923–1929.

David Whitney, Eric Rosen, James MacGlashan, and Lawson and Tellex Stefanie L.S. Wong. fcmng. Reducing errors in object-fetching interactions through social feedback. In *Proceedings of the IEEE International Conference on Robotics and Automation ICRA 2017, May 29 – June 3, 2017, Marina Bay Sands, Singapore*.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2015. Comparing attribute classifiers for interactive language grounding. In *Proceedings of the Fourth Workshop on Vision and Language*. Association for Computational Linguistics, Lisbon, Portugal, pages 60–69. http://aclweb.org/anthology/W15-2811.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016a. Comparing dialogue strategies for learning grounded language from human tutors. In *Proceedings of Semdial 2016 (JerSem)*. New Jersey.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016b. Interactively learning visually grounded word meanings from a human tutor. In *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016c. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*. pages 339–349.

Yanchao Yu, Arash Eshghi, Gregory Mills, and Oliver Lemon. 2017. *Proceedings of the Sixth Workshop on Vision and Language*, Association for Computational Linguistics, chapter The BURCHAK corpus: a Challenge Data Set for Interactive Learning of Visually Grounded Word Meanings, pages 1–10. http://aclweb.org/anthology/W17-2001.

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, page 116.