Teaching Agents Through Correction

Mattias Appelgren
University of Edinburgh
Edinburgh
M.R.Appelgren@sms.ed.ac.uk

ABSTRACT

We motivate and describe a novel task which is modelled on interactions between apprentices and expert teachers. In the task an agent must learn to build towers constrained by rules. The teacher provides verbal corrective feedback from which the agent learns. The agent starts out unaware of the constraints as well as the domain concepts in which the constraints are expressed. Therefore an agent that takes advantage of the linguistic evidence must learn the denotations of neologisms and adapt its conceptualisation of the planning domain to incorporate those denotations. We show that an agent which does utilise linguistic evidence outperforms a strong baseline which does not.

KEYWORDS

human-robot interaction; interactive learning; knowledge representation and reasoning

ACM Reference Format:
Mattias Appelgren. 2019. Teaching Agents Through Correction. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

It has long been a goal of Artificial Intelligence research to be able to interact with agents through natural language [15]. One advantage that could be gained through such interaction is the ability to teach new tasks and concepts, a goal of Interactive Task Learning (ITL) [8]. In ITL we seek to teach agents the requirements of a previously unknown task, rather than how to perform a known task optimally. For example, an agent might learn the rules of a game [5] rather than how to play optimally once the rules are known [13].

The mode of teaching is through interaction, which might replace a programmer explicitly detailing the parameters of the task, for example through a reward function. Although several modes of interaction are possible, such as physical demonstration [3], we will focus on language. Language allows us to create concepts for things like objects and actions. These concepts are extremely powerful since they allow for more efficient and effective communication. For example, saying "give me a disk of bread covered with tomato sauce, cheese, and pepperoni baked in the oven until the cheese melts and the bread has browned" is a lot less efficient than "give me a pepperoni pizza". Teaching such concepts to an agent gives the added benefit that a non technical human may be able to interact with the agent in a meaningful manner, since they share a conceptualisation of the world. However, these benefits do require that the agent learn to associate the concepts with objects in the real world [10].

To teach an agent through interaction an interactive dialogue is held between agent and teacher [7, 12]. The agent and teacher may make both verbal and non-verbal moves in the dialogue, such as issuing instructions, asking and answering questions, giving definitions, or describing objects or situations [1, 4, 6, 7, 12, 14]. However, there are many other ways in which people interact [9] that have not been well studied for ITL scenarios.

The goal of my thesis is to explore the use of a broader set of dialogue moves for teaching agents a new task. In this extended abstract I will briefly present the work I have done on learning from correction (for details see full paper [2]), consisting of a new task where an agent must learn to build towers out of coloured blocks in such a way that they comply to a number of constraints as well as a proof of concept agent that solves this task. These constraints and the words used to describe the block are unknown to the agent. The agent must learn to build rule compliant towers from a teacher's feedback.

Correction is a useful mode of interaction as it can tell the agent about it's own knowledge as well as how the teacher sees the world. For example, if a teacher says "no, don't put red blocks on blue blocks", this tells the agent that putting red blocks on blue blocks should be avoided. However, assuming it was given in response to an action a, then the agent would also be able to infer that a resulted in a bad state where "don't put red blocks on blue blocks" is violated. This is based on the fact that correction should only be said if it draws to light an inconsistency between the corrected move and what is said in the correction [9]. Thus, this interaction can both teach the agent about a constraint but also give instances of "red blocks" and "blue blocks" for the purpose of learning to recognise these concepts in the future.

We present our agent and show that it outperforms two baselines that do not attempt to use the full language, instead just learning from "no". In the process the agent learns the actual rules constraining the problem as well as colour words which are shared with the teacher.

2 METHODOLOGY

Our task requires an agent to build towers out of blocks such that they comply with a set of rules along the lines of

$$\forall x.red(x) \rightarrow \exists y.on(x, y) \wedge blue(y) \quad (1)$$

thus constraining where the blocks can be placed given their colour. The agent must place all available blocks into the tower in each scenario. It begins aware that constraints of this form exist, but does not know what they are, how many there are, what colour terms exist, nor how to recgonise colour terms given RGB values.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

These things must be learned through interaction. In addition to this, the corrective utterances may be ambiguous as to what rule is implied by that utterance.

To learn the agent constructs the probability of a particular instance of correction

$$P(Corr(a_i, u_i), X, Z) \tag{2}$$

where $Corr(a_i, u_i)$ represents that action a_i is corrected with utterance u_i. X is made up of visible features, namely the RGB value of relevant objects (referred to as $F(x)$). Z is made up of hidden variables relevant to the correction. In particular, these include whether a particular rule is part of the goal: $r_i \in G$, and whether a particular colour term can be used to describe an object: e.g. $Red(x)$.

To perform planning the agent must learn what rules are in the goal and it must learn to predict whether a particular object belongs to a relevant colour category. To predict the colour we create grounding models which estimate

$$P(Red(x)|F(x)) \propto P(F(x)|Red(x))P(Red(x)) \tag{3}$$

and equivalent for other colours. These are binary classifiers, answering the question "is object x red or not" rather than "is object x red, blue, or green". This allows more flexibility since we do not initially know what colour terms exist and it is possible that more than one colour term applies to an object (e.g. red and maroon).

With this in mind the correction model is used to estimate two things. First, to estimate how likely it is that a rule is part of the goal, given the evidence

$$P(r_i \in G|Corr(a_i, u_i), X) \tag{4}$$

which is used to update the agents goal representation.

Second, to estimate how likely it is that each relevant object can be described by a relevant colour

$$P(Red(x)|Corr(a_i, u_i), X) \tag{5}$$

Since the agent does not have access to labelled data for recognising colour terms these probabilities are used in an Expectation Maximisation like fashion to update the parameters of the sensor models. Every time the agent is corrected it updates its belief about the goal and estimates the most likely setting of the colour variables. It then updates the sensor models and use these as priors when calculating the probability of the next correction. If the agent is extremely unsure it may ask a clarification question such as: "is the top object red?" which the teacher answers with yes or no. This happens especially at the beginning of training when the agent has no information about any colour terms.

Repeated trial and error, with the teacher's correction, allow the agent to acquire the necessary colour terms and learn the constraints.

3 RESULTS

To test the agent's capabilities we ran it on a number of different planning problems, defined by what rules constrained the towers. Each problem varied on both what rules were included as well as the number of rules. We compared the agent to two baselines, a naive baseline which did not attempt to learn at all (it simply avoided performing a corrected action twice) and a more clever baseline which attempted to learn to avoid situations similar to previously corrected states but did not attempt to learn the colour terms or the true underlying rules.

The clever baseline outperformed the naive one, showing that this agent does learn. However, our language aware agent outperformed both agents consistently over all tested planning problems. Further in cases where more than one rule contained the same colour term the language aware agent performed significantly better, as it could generalise these overlapping terms.

4 DISCUSSION

The agent described here serves as a proof of concept for learning from correction. However, there are several limitations which would need to be overcome to deploy the system in the wild.

First, the visual processing system is simple, only dealing with colours of simulated blocks. Partially this comes down to a lack of maturity in the research project, as a more complicated model could replace the current system, however, it is also a deliberate choice since our system is largely bounded by the speed of concept acquisition. It is untenable to have a teacher give thousands of instances of feedback, which would be required by many current vision systems. Instead, we work on simple vision with an update function simple enough to allow a broad range of vision systems to be integrated when the time comes.

We are more concerned with the restrictions placed on the language. Our system makes strong assumptions both about what is said and how it is said. The language content we wish to support is simple enough that current semantic parsing technology should be sufficient, and there is work on updating parsers online [11, 16]. The more significant problem for us are assumptions about the type of moves the teacher makes. We assume strict, orderly interaction, whereas real interaction is messy and ambiguous. We assume the agent knows that type of move has been made, but this will not be obvious and must be predicted by the agent.

Our current direction for future work is to expand what interactions the agent supports, allowing the teacher to make reference to previous moves. Further, we will expand the set of rules the agent can learn and the diversity of how these things are expressed.

5 CONCLUSION

We present a new task which requires an agent to learn from a teacher's corrections. The agent learns to recognise colours and learns rules that constrain the problem. The agent is tested against two baselines and outperforms them both. Due to simplifying assumptions the agent is severely limited, but serves as a proof of concept, which will be expanded in future work.

ACKNOWLEDGMENTS

I would like to thank EPSRC for funding my PhD, Alex Lascarides for patient supervision and guidance, and Ram Ramamoorthy and Yordan Hristrov for helpful discussions.

REFERENCES

[1] Muhannad Al-Omari, Paul Duckworth, Majd Hawasly, David C. Hogg, and Anthony G. Cohn. 2017. Natural Language Grounding and Grammar Induction for Robotic Manipulation Commands. In *RoboNLP@ACL*.

[2] Mattias Appelgren and Alex Lascarides. 2019. Learning Plans by Acquiring Grounded Linguistic Meanings from Corrections. In *In Proc. of the 18th Inter-*

national Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13âĂŞ17, 2019, IFAAMAS. 9 pages.

[3] Brenna Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. 57 (05 2009), 469–483.

[4] Maxwell Forbes, Rajesh P. N. Rao, Luke Zettlemoyer, and Maya Cakmak. 2015. Robot Programming by Demonstration with situated spatial language understanding. In *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*. 2014–2020. https://doi.org/10.1109/ICRA.2015.7139462

[5] Thomas R. Hinrichs and Kenneth D. Forbus. 2014. X Goes First: Teaching Simple Games through Multimodal Interaction.

[6] Siddharth Karamcheti, Edward C. Williams, Dilip Arumugam, Mina Rhee, Nakul Gopalan, Lawson L. S. Wong, and Stefanie Tellex. 2017. A Tale of Two DRAGGNs: A Hybrid Approach for Interpreting Action-Oriented and Goal-Oriented Instructions. In *Proceedings of the First Workshop on Language Grounding for Robotics, RoboNLP@ACL 2017, Vancouver, Canada, August 3, 2017*, Mohit Bansal, Cynthia Matuszek, Jacob Andreas, Yoav Artzi, and Yonatan Bisk (Eds.). Association for Computational Linguistics, 67–75. https://aclanthology.info/papers/W17-2809/w17-2809

[7] Evan A. Krause, Michael Zillich, Thomas Emrys Williams, and Matthias Scheutz. 2014. Learning to Recognize Novel Objects in One Shot through Human-Robot Interactions in Natural Language Dialogues. In *AAAI*.

[8] John E. Laird, Kevin A. Gluck, John R. Anderson, Kenneth D. Forbus, Odest Chadwicke Jenkins, Christian Lebiere, Dario D. Salvucci, Matthias Scheutz, Andrea Lockerd Thomaz, J. Gregory Trafton, Robert E. Wray, Shiwali Mohan, and James R. Kirk. 2017. Interactive Task Learning. *IEEE Intelligent Systems* 32 (2017), 6–21.

[9] Alex Lascarides and Nicholas Asher. 2009. Agreement, Disputes and Commitments in Dialogue. *J. Semantics* 26, 2 (2009), 109–158. https://doi.org/10.1093/jos/ffn013

[10] Cynthia Matuszek. 2018. Grounded Language Learning: Where Robotics and NLP Meet. In *IJCAI*.

[11] Matthias Scheutz, Evan A. Krause, Bradley Oosterveld, Tyler M. Frasca, and Robert Platt Jr. 2017. Spoken Instruction-Based One-Shot Object and Action Learning in a Cognitive Robotic Architecture. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*. 1378–1386. http://dl.acm.org/citation.cfm?id=3091315

[12] Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Yue Chai, and Ning Xi. 2014. Back to the Blocks World: Learning New Actions through Situated Human-Robot Dialogue. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*. 89–97. http://aclweb.org/anthology/W/W14/W14-4313.pdf

[13] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362 (2018), 1140–1144.

[14] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. 2011. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*. http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3623

[15] Terry Winograd. 1972. Understanding natural language. *Cognitive psychology* 3, 1 (1972), 1–191.

[16] Luke S. Zettlemoyer and Michael Collins. 2007. Online Learning of Relaxed CCG Grammars for Parsing to Logical Form. In *EMNLP-CoNLL*.