ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1640

Transfer of Reinforcement Learning Negotiation Policies: From Bilateral to Multilateral Scenarios

Ioannis Efstathiou and **Oliver Lemon**[1]

Abstract. Trading and negotiation dialogue capabilities have been identified as important in a variety of AI application areas. In prior work, it was shown how Reinforcement Learning (RL) agents in bilateral negotiations can learn to use manipulation in dialogue to deceive adversaries in non-cooperative trading games. In this paper we show that such trained policies can also be used effectively for multilateral negotiations, and can even outperform those which are trained in these multilateral environments. Ultimately, it is shown that training in simple bilateral environments (e.g. a generic version of "Catan") may suffice for complex multilateral non-cooperative trading scenarios (e.g. the full version of Catan).

1 Introduction

Work on automated conversational systems has previously been focused on cooperative dialogue, where a dialogue system's core goal is to assist humans in their tasks such as finding a restaurant [13]. However, non-cooperative dialogues, where an agent may act to satisfy its own goals, are also of practical and theoretical interest [6]. It may be useful for a dialogue agent not to be fully cooperative when trying to gather information from a human, or when trying to persuade, or in the area of believable characters in video games and educational simulations [6]. Another area in which non-cooperative dialogue behaviour is desirable is in negotiation [12]. Recently, Reinforcement Learning (RL) methods have been applied in order to optimise *cooperative* dialogue management, where the decision of the next dialogue move to make in a conversation is in focus, in order to maximise an agent's overall long-term expected utility [13, 9, 10]. Those methodologies used RL with reward functions that give positive feedback to the agent only when it meets the user's goals. This work has shown that robust and efficient dialogue management strategies can be learned, but until [3], has only addressed cooperative dialogue. Lately it has been shown [5] that when given the ability to perform both cooperative and non-cooperative (manipulative) dialogue moves, a dialogue agent can learn to bluff and to lie during trading so as to win games more often, under various conditions such as risking penalties for being caught in deception – against a variety of adversaries [4]. Here we transfer those learned bilateral policies to more complex multilateral negotiations, and evaluate them.

2 Learning in Bilateral Negotiations

To learn trading policies in a controlled setting we initially [5] used a 2-player version of the non-cooperative 4-player board game "Catan". We call the 2 players the "adversary" and the "Reinforcement learning agent" (RLA). The goal of the RLA was to gather a particular number of resources via trading dialogue. Trade occurred through proposals that might lead to acceptance or rejection from the adversary. In an agent's proposal (turn) only one 'give 1-for-1' or 'give 1-for-2' trading proposal might occur, or nothing (41 actions in total), e.g. "I will give you a brick and I need two rocks". To overcome issues related to long training times and high memory demands, we have implemented a state encoding mechanism [5] that automatically compresses all of our numeric trading game states.

We first investigated the case of learning trading policies against adversaries which always accepted a trading proposal. The *goal-oriented RLA* did not use any manipulative actions and learned to reach its goal resources as soon as possible. In the case where the goal was to build a city it learned to win 96.8% of the time [5]. We then trained the *manipulative (dishonest) RLA* [5], which could ask for resources that it did not really need. It could also propose trades without checking whether the offered resource was available. The manipulated adversary [5] was implemented based on the intuition that a rational adversary will act so as to hinder other players in respect of their expressed preferences. The above trained policies of both of the agents are now evaluated in JSettlers [11].

3 Evaluating in Multilateral Negotiations

The experiments here are all conducted using JSettlers [11], a research environment developed in Java that captures the full multiplayer version of the game Catan, where there is trading and building. 10k games were played for each experiment. The players are:

The original STAC Robot (Bot) is based on the original expert rule-based agent of JSettlers [11] which is further modified to improve its winning performance. This agent (the Bot), which is the "benchmark" agent described in [7], uses complex heuristics to increase performance by following a dynamic building plan according to its current resource needs and the board's set-up.

Our trained RLA is in fact a Bot which has been modified to make offers based on our four learnt policies (for the development of city, road, development card, and settlement) in our version of the game "Catan" (Section 2). These policies were either the *goal-oriented* ones or the *manipulative (dishonest)* ones.

The Bayesian agent (Bayes) [8] is a Bot whose trading proposals are made based on the human corpus that was collected from Afantenos et al. [1]. The Bayesian agent was 65.7% accurate in reproducing the human moves.

The Manipulated Bot is a Bot which can be manipulated by our trained dishonest agent (i.e. the weights of the resources that they offer and ask for change according to the trained manipulative RL proposals). There are 3 types of manipulated Bots as we will see.

[1] Interaction Lab, Heriot-Watt University, Edinburgh, email: i.efstathiou, o.lemon@hw.ac.uk

3.1 Evaluation without Manipulation

Trained RLA (goal-oriented) vs. 3 Bots: Our trained RLA resulted in a performance of 32.66%[2], while those of the Bots were 22.9%, 22.66% and 21.78% respectively. This was interesting because it proved that our generic 2-player version of the game (Section 2) was enough to train a successful policy for the multi-player version of the game, by effectively treating all three opponents as one. Hence our RLA proposed only *public trades*. Furthermore the 32.66% performance of our RLA was *around 7% better than that of* [8], who trained it in the real multilateral negotiations environment (JSettlers).

Trained RLA (goal-oriented) vs. 3 Bayes: In this experiment our trained agent scored a performance of 36.32%, which is much higher than those of the three Bayes agents. Their performances were 21.43%, 21.02% and 21.23% respectively.

3.2 Evaluation with Manipulation

Here we evaluated our previously trained dishonest RL policies against the 3 types of Manipulated Bots and the Bayes agents.

Trained Dishonest RLA vs. 3 Manipulated Bots: In this experiment the 3 manipulated Bots win rates were 21.44%, 20.79% and 21.42% respectively. Our trained Dishonest RLA won by 36.35%.

Trained Dishonest RLA vs. 3 Manipulated Bots (Weights based on Building Plan): The Bot's probabilities are adjusted further according to the building plan (BP) in this case. That means that the Bots are initially biased towards specific resources, as the BP indicates the next piece to build (e.g. city). The results of this experiment were still satisfying: the 3 manipulated Bots won by 22.53%, 21.47% and 21.8% respectively. Our trained Dishonest RLA won by 34.2%.

Trained Dishonest RLA vs. 3 Manipulated Bots (Weights based on Building Plan and Resource Quantity): This case is identical to the above but the trade probabilities are additionally adjusted according to the goal resource quantity. The results of this experiment for the trained Dishonest RL policies were as good as the above: the 3 manipulated Bots win rates were 21.72%, 21.5% and 22.47% respectively. Our trained Dishonest RLA won by 34.33%. This result, along with the two above, suggested that the RLA's dishonest manipulative policies were very effective against the Bots of the multi-player version of the game, showing that our transition from a bilateral negotiation environment to a multilateral one was successful.

Trained Dishonest RLA vs. 3 Bayes: We hypothesised in this case that the human players might have been affected by their opponents' manipulation (if any occurred in the data collection [1]), and we wanted to test that by using our Dishonest policy. The results proved our hypothesis: the 3 Bayes agents won by 21.97%, 20.58% and 21.64% respectively. Our trained Dishonest RLA won by 35.81%. This was an evidence that the Bayes agents were indeed affected by manipulation, and now by the Dishonest RLA's manipulative policy too, and its success resulted in almost 14% more winning games.

4 Conclusion

We showed that our trained bilateral RL policies from our generic version of "Catan" were able to outperform (by at least 10%) the agents of the JSettlers [11] environment and even managed to successfully manipulate them. That demonstrated how successful trained policies from bilateral negotiations can be, when evaluated in more complex multilateral ones, even compared to those which are trained in these multilateral negotiations. Hence training RL policies in complex multilateral negotiations may be unnecessary in some cases. Furthermore, by considering all of the opponents as one player, and by proposing public trades for all players, we bypass complexities that arise by personalizing the agent's trading proposals for each distinct opponent. Our findings show that an explicit model of each adversary is not required for successful RL policies to be learned in this case. Ultimately, it suggests that an implicit model of a complex trading scenario may be enough for effective RL, providing that efficient selection of the state representation and of the actions has been made.

Further work explores Deep Reinforcement Learning approaches to trading dialogue [2].

[2] The baseline performance in a four-player game is 25%

Acknowledgements

This work is funded by the ERC, grant no. 269427 (STAC project).

REFERENCES

[1] Stergos Afantenos, Nicholas Asher, Farah Benamara, Anais Cadilhac, Cedric Degremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, Philippe Muller, Soumya Paul, Verena Rieser, and Laure Vieu, 'Developing a corpus of strategic conversation in The Settlers of Catan', in *Proceedings of SemDial 2012*, (2012).

[2] Heriberto Cuayahuitl, Simon Keizer, and Oliver Lemon, 'Strategic Dialogue Management via Deep Reinforcement Learning', in *NIPS workshop on Deep Reinforcement Learning*, (2015).

[3] Ioannis Efstathiou and Oliver Lemon, 'Learning non-cooperative dialogue behaviours', in *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 60–68, Philadelphia, PA, U.S.A, (2014).

[4] Ioannis Efstathiou and Oliver Lemon, 'Learning to manage risks in non-cooperative dialogues.', in *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2014 - DialWatt)*, pp. 173–175, Edinburgh, Scotland, U.K., (2014).

[5] Ioannis Efstathiou and Oliver Lemon, 'Learning non-cooperative dialogue policies to beat opponent models: "the good, the bad and the ugly"', in *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2015 - GoDial)*, pp. 33–41, Gothenburg, Sweden, (2015).

[6] Kallirroi Georgila and David Traum, 'Reinforcement learning of argumentation dialogue policies in negotiation', in *Proc. INTERSPEECH*, (2011).

[7] Markus Guhe and Alex Lascarides, 'Game strategies in *the settlers of catan*', in *Proceedings of the IEEE Conference on Computational Intelligence in Games*, Dortmund, (2014).

[8] Simon Keizer, Heriberto Cuayahuitl, and Oliver Lemon, 'Learning trade negotiation policies in strategic conversation', in *The 19th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2015 - goDIAL)*, pp. 104–112, (2015).

[9] Verena Rieser and Oliver Lemon, *Reinforcement Learning for Adaptive Dialogue Systems: A Data-driven Methodology for Dialogue Management and Natural Language Generation*, Theory and Applications of Natural Language Processing, Springer, 2011.

[10] Verena Rieser, Oliver Lemon, and Xingkun Liu, 'Optimising information presentation for spoken dialogue systems', in *Proc. ACL*, (2010).

[11] R. Thomas and K. Hammond, 'Java settlers: a research environment for studying multi-agent negotiation', in *Proc. of IUI '02*, pp. 240–240, (2002).

[12] David Traum, 'Extended abstract: Computational models of non-cooperative dialogue', in *Proc. of SIGdial Workshop on Discourse and Dialogue*, (2008).

[13] Steve Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu, 'The Hidden Information State Model: a practical framework for POMDP-based spoken dialogue management', *Computer Speech and Language*, **24**(2), 150–174, (2010).