ECAI 2014
T. Schaub et al. (Eds.)

doi:10.3233/978-1-61499-419-0-999

Learning non-cooperative behaviour for dialogue agents

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

Abstract. Non-cooperative dialogue behaviour for artificial agents (e.g. deception and information hiding) has been identified as important in a variety of application areas, including education and healthcare, but it has not yet been addressed using modern statistical approaches to dialogue agents. Deception has also been argued to be a requirement for high-order intentionality in AI. We develop and evaluate a statistical dialogue agent using Reinforcement Learning which learns to perform non-cooperative dialogue moves in order to complete its own objectives in a stochastic trading game with imperfect information. We show that, when given the ability to perform both cooperative and non-cooperative dialogue moves, such an agent can learn to bluff and to lie so as to win more games. For example, we show that a non-cooperative dialogue agent learns to win 10.5% more games than a strong rule-based adversary, when compared to an optimised agent which cannot perform non-cooperative moves. This work is the first to show how agents can learn to use dialogue in a non-cooperative way to meet their own goals.

1 Introduction

Research in automated conversational systems has almost exclusively focused on the case of cooperative dialogue, where a dialogue system's core goal is to assist humans in particular tasks, such as buying airline tickets [8] or finding a place to eat [9]. However, non-cooperative dialogues, where an agent may act to fulfil its own goals rather than the user's goals, are also of practical and theoretical interest [3], and the game-theoretic underpinnings of non-Gricean behaviour are actively being investigated [1]. For example, it may be advantageous for an automated agent not to be fully cooperative when trying to gather information from a human, when trying to persuade, argue, or debate, when trying to sell them something, when trying to detect illegal activity (for example on internet chat sites), or in the area of believable characters in video games, and educational simulations [3, 5]. Another arena in which non-cooperative dialogue behaviour is desirable is in negotiation [7], where hiding information (and even outright lying) can be advantageous. Dennett argues that deception capability is required for higher-order intentionality [2].

A complementary research direction in recent years has been the use of machine learning methods to automatically optimise *cooperative* dialogue management - i.e. the decision of what dialogue move to make next in a conversation, in order to maximise an agent's overall long-term expected utility [9, 4]. This research has shown how robust and efficient dialogue management strategies can be learned from data, but has only addressed the case of cooperative dialogue. These approaches use Reinforcement Learning with a reward function that gives positive feedback to the agent only when it meets the user's goals.

[1] Interaction Lab, Heriot-Watt University, Edinburgh, email: ie24,o.lemon@hw.ac.uk

An example of the type of non-cooperative dialogue behaviour which we are generating is given by agent B in the following dialogue:
A: "I will give you a sheep if you give me a wheat"
B: "No"
B: "I really need rock" [B actually needs wheat]
A: "OK"
A: "I'll give you a wheat if you give me rock"

Here, A is deceived into providing the wheat that B actually needs, because A believes that B needs rock rather than wheat.

In this paper we investigate whether a learning agent endowed with non-cooperative dialogue moves and a 'personal' reward function can learn how to perform non-cooperative dialogue. Note that the reward will not be given for performing non-cooperative moves themselves, but only for winning trading games. We therefore explore whether the agent can learn the advantages of being non-cooperative in dialogue, in a variety of settings.

2 The Trading Game

To investigate non-cooperative dialogues in a controlled setting we created a 2-player, sequential, non-zero-sum game with imperfect information called "Taikun". This game can be extended to capture different aspects of trading and negotiation. We call the 2 players the "adversary" and the "learning agent" (LA).

The two players trade three kinds of resources to each other sequentially, in a 1-for-1 manner, in order to reach a specific number of resources that is their goal. The player who first attains their goal resources wins. Both players start the game with one resource of each type (wheat, sheep, and rock). At the beginning of each round the game updates the number of resources of both players by either removing one of them or adding two of them, thereby making the opponent's state (the cards that they hold) unobservable. In the long run, someone will eventually win even if no player ever trades. However, effective trading can provide a faster victory.

2.1 Trading Proposals and Manipulation

Trade occurs through trading proposals that may lead to acceptance from the other player. In an agent's turn only one '1-for-1' trading proposal may occur for each resource, or nothing (7 actions in total). Agents respond by either saying "No" or "OK" in order to reject or accept the other agent's proposal.

In our second experiment three manipulative actions are added to the learning agent's set of actions, of the form "I really need X" where X is a resource type. The adversary might believe such statements, resulting in modifying their probabilities of making certain trades.

3 The Learning Agent (LA) and Adversaries

The game state can be represented by the learning agent's set of resources, its adversary's set of resources, and a trading proposal (if any) currently under consideration. The learning agent (LA) plays the game and learns while perceiving only its own set of resources. The LA is aware of its winning condition (to obtain 4 wheat and 5 rocks) in as much as it experiences a large final reward when reaching this state. It learns how to achieve the goal state through trial-and-error exploration while playing repeated games.

The LA is modelled as a Markov Decision Process [6]: it observes states, selects actions according to a policy, transitions to a new state (due to the adversary's move and/or a update of resources), and receives rewards at the end of each game. This reward is then used to update the policy followed by the agent.

The rewards that were used in these experiments were 1,000 for the winning and draw cases (because the goal states of these cases are the same) and -100 when losing a game. The LA was trained using a custom SARSA(0) learning method [6] with an initial exploration rate of 0.2 that gradually decays to 0 at the end of the training games. It was trained over 5.5 million games against each adversary, and the resulting policies were then tested in 20 thousand games.

We investigated performance with several different **adversaries**. As a baseline, we need to know how well a LA which does not have non-cooperative moves at its disposal can perform against a rational adversary. Our main hypothesis is that a LA with additional non-cooperative moves can outperform this when adversaries are somewhat gullible.

3.1 Rule-based adversary: experiment 1

This strategy was designed to form a challenging rational adversary for measuring baseline performance. It cannot be manipulated at all, and non-cooperative dialogue moves have no effect on it.

The strict rule-based strategy of the adversary will never ask for a resource that it does not need (in this case rocks). Furthermore, if it has an available non-goal resource to give then it will offer it. It only asks for resources that it needs (goal resources: wheat and sheep). In the case where it does not have a non-goal resource (rocks) to offer then it offers a goal resource only if its quantity is more than it needs, and it asks for another goal resource if it is needed.

Following the same reasoning, when replying to the LA's trading proposals, the adversary will never agree to receive a non-goal resource (rock). It only gives a non-goal resource (rock) for another one that it needs (wheat or sheep). It also agrees to make a trade in the special case where it will give a goal resource that is currently more than it needs for another one that it does need. This is a strong strategy that wins a significant number of games. In fact, it takes about 400,000 training games before the LA is able to start winning more than this adversary, and a random LA policy loses 66% of games against this adversary (See Table 1).

3.2 Gullible adversary: experiment 2

The adversary in this case retains the above strict base-line policy but it is also susceptible to the non-cooperative moves of the LA.

For example, if the LA utters "I really need rock", weights of actions which transfer rock from the the adversary will decrease, and the adversary will then be less likely to give rock to the LA. Conversely, the adversary is then more likely to give the other two resources to the LA. In this way the LA has the potential to mislead the adversary into giving the resources that it really needs.

4 Results

In Experiment 1 (baseline, no manipulation) the LA scored a winning performance of 49.23% against 45.62% for the adversary, with 5.15% draws (Table 1), in the 20 thousand test games. This represents the baseline performance that the LA is able to achieve against an adversary who cannot be manipulated at all. Hence the game is 'solveable' as an MDP problem. Here, the learning agent's strategy focuses on offering the sheep resource that it does not need for the rocks that it needs.

In Experiment 2 (adding manipulation) the learning agent scored a winning performance of 55.58% against 41.44% for the adversary, having 2.97% draws (Table 1), in the 20 thousand test games. Here the high frequency manipulative actions ("I need wheat" and "I need sheep") assist in deceiving the adversary by hiding information and lying respectively, therefore significantly increasing performance.

LA Policy	LA	Adversary	Draws
Random	32%	66%	2%
Exp 1 (baseline)	49.23%	45.62%	5.15%
Exp 2 (+manipulation)	**55.585%***	41.440%	2.975%

Table 1. *Performance (% wins) in 20 thousand testing games (*= significant improvement over baseline, $p < 0.05$)*

In Experiment 2 the adversary, further being deceived by the learning agent's hiding information and lying actions, loses 14.14% more often than the learning agent.

5 Conclusion & Future Work

We showed that a statistical dialogue agent can learn to perform non-cooperative dialogue moves in order to enhance its performance in trading negotiations. This demonstrates that non-cooperative dialogue strategies can emerge from statistical approaches to dialogue management. There are many extensions to pursue, including opponent modelling using belief states in POMDP-style models [9].

The long-term goal of this work is to develop intelligent agents that will be able to assist (or even replace) users in interaction with other human or artificial agents in various non-cooperative settings [5], such as trading, education, and healthcare.

Acknowledgements

This work was partially funded by the STAC project, funded ERC grant 269427, see http://www.irit.fr/STAC/

REFERENCES

[1] N. Asher and A. Lascarides, 'Commitments, beliefs and intentions in dialogue', in *Proc. of SemDial*, pp. 35–42, (2008).
[2] Daniel Dennett, 'When Hal Kills, Who's to Blame? Computer Ethics', in *Hal's Legacy:2001's Computer as Dream and Reality*, (1997).
[3] Kallirroi Georgila and David Traum, 'Reinforcement learning of argumentation dialogue policies in negotiation', in *INTERSPEECH*, (2011).
[4] 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.
[5] J. Shim and R.C. Arkin, 'A Taxonomy of Robot Deception and its Benefits in HRI', in *Proc. IEEE Systems, Man, and Cybernetics*, (2013).
[6] R. Sutton and A. Barto, *Reinforcement Learning*, MIT Press, 1998.
[7] David Traum, 'Computational models of non-cooperative dialogue', in *Proc. of SIGdial Workshop on Discourse and Dialogue*, (2008).
[8] M. Walker, R. Passonneau, and J. Boland, 'Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems', in *Proc. of ACL*, (2001).
[9] 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).