Learning non-cooperative dialogue behaviours

Ioannis Efstathiou
Interaction Lab
Heriot-Watt University
ie24@hw.ac.uk

Oliver Lemon
Interaction Lab
Heriot-Watt University
o.lemon@hw.ac.uk

Abstract

Non-cooperative dialogue behaviour has been identified as important in a variety of application areas, including education, military operations, video games and healthcare. However, it has not been addressed using statistical approaches to dialogue management, which have always been trained for co-operative dialogue. We develop and evaluate a statistical dialogue agent which learns to perform non-cooperative dialogue moves in order to complete its own objectives in a stochastic trading game. 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 games more often – against a variety of adversaries, and under various conditions such as risking penalties for being caught in deception. For example, we show that a non-cooperative dialogue agent can learn to win an additional 15.47% of games against 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 an agent can learn to use non-cooperative dialogue to effectively meet its 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 (Walker et al., 2001) or finding a place to eat (Young et al., 2010). Gricean cooperative principles have been shown to emerge from multi-agent decision theory, in a language task modelled using Decentralised Partially Observable Markov Decision Processes (Vogel et al., 2013a), and in related work conversational implicature was argued to be a by-product of agents who maximise joint utility (Vogel et al., 2013b).

However, non-cooperative dialogues, where an agent may act to satisfy its own goals rather than those of other participants, are also of practical and theoretical interest (Georgila and Traum, 2011), and the game-theoretic underpinnings of non-Gricean behaviour are actively being investigated (Asher and Lascarides, 2008). For example, it may be advantageous for an automated agent not to be fully cooperative when trying to gather information from a human, and 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 (Georgila and Traum, 2011; Shim and Arkin, 2013). Another arena in which non-cooperative dialogue behaviour is desirable is in negotiation (Traum, 2008), where hiding information (and even outright lying) can be advantageous. Furthermore, deception is considered to be an essential part of successful military operations. According to Sun Tzu "All warfare is based on deception" and Machiavelli clearly states in *The Discourses* that "Although deceit is detestable in all other things, yet in the conduct of war it is laudable and honorable"(Arkin, 2010). Indeed, Dennett argues that deception capability is required for higher-order intentionality in AI (Dennett, 1997).

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, which is usually defined in terms of meeting a user's goals (Young et al., 2010; Rieser

Proceedings of the SIGDIAL 2014 Conference, pages 60–68,
Philadelphia, U.S.A., 18-20 June 2014.

and Lemon, 2011). 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.

An example of the type of non-cooperative dialogue behaviour which we are generating in this work 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... I'll give you a wheat if you give me rock"
B: "OK"

Here, A is deceived into providing the wheat that B actually needs, because A believes that B needs rock rather than wheat. Similar behaviour can be observed in trading games such as Settlers of Catan (Afantenos et al., 2012).

1.1 Non-cooperative dialogue and implicature

Our trading dialogues are linguistically cooperative (according to the Cooperative Principle (Grice, 1975)) since their linguistic meaning is clear from both sides and successful information exchange occurs. Non-linguistically though they are non-cooperative, since they they aim for personal goals. Hence they violate Attardo's Perlocutionary Cooperative Principle (PCP) (Attardo, 1997). In our non-cooperative environment, the manipulative utterances such as "I really need sheep" can imply that "I don't really need any of the other two resources", as both of the players are fully aware that three different resources exist in total and more than one is needed to win the game, so therefore they serve as scalar implicatures (Vogel et al., 2013b). Hence we will show that the LA learns how to include scalar implicatures in its dialogue to successfully deceive its adversary by being cooperative on the locutionary level and non-cooperative on the perlocutionary level.

1.2 Structure of the paper

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. This is similar to (Vogel et al., 2013a) who show how cooperativity emerges from multi-agent decision making, though in our case we show the emergence of non-cooperative dialogue behaviours.

We begin with the case of a simple but challenging 2-player trading game, which is stochastic and involves hidden information.

In section 2 we describe and motivate the trading game used in this work, and in section 3 we describe the Learning Agent. In section 4 we explain the different adversaries for experimentation, in section 5 we provide results, and in section 6 we conclude and discuss areas for future work.

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". Motivated by the principle of Occam's razor we shaped this game as simply as possible, while including the key features of a resource trading game. The precise goal was also to implement mechanics that are not restrictive for the future of this research and therefore can be flexibly extended to capture different aspects of trading and negotiation. We call the 2 players the "adversary" and the "learning agent" (LA).

The two players can trade three kinds of resources with each other sequentially, in a 1-for-1 manner, in order to reach a specific number of resources that is their individual goal. The player who first attains their goal resources wins. Both players start the game with one resource of each type (wheat, rock, and sheep). 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 (i.e. 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 "Taikun" game characteristics

Taikun is a sequential, non-cooperative, non-zero-sum game, with imperfect information, where:

- The goal is to reach either 4 or 5 of two specific resources (4 wheat and 5 rocks for the learning agent and 4 wheat and 5 sheep for the adversary). The players share a goal resource (wheat).
- Each round consists of an update of resources turn, the learning agent's trading proposal turn (and adversary's acceptance or rejection), and finally the adversary's trading proposal turn (and LA's acceptance or rejection).
- The update turn, which is a hidden action, changes one of the resources of each player at random by +2 or -1.
- When a resource is "capped", that is if its number is 5 or more, then no update rule can be applied to it. Trading can still change its quantity though.

2.2 Actions (Trading Proposals)

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, or nothing (7 actions in total):

1. I will do nothing
2. I will give you a wheat if you give me a rock
3. I will give you a wheat if you give me a sheep
4. I will give you a rock if you give me a wheat
5. I will give you a rock if you give me a sheep
6. I will give you a sheep if you give me a wheat
7. I will give you a sheep if you give me a rock

Agents respond by either saying "No" or "OK" in order to reject or accept the other agent's proposal.

2.3 Non-cooperative dialogue moves

In our second experiment three manipulative actions are added to the learning agent's set of actions:

1. "I really need wheat"
2. "I really need rock"
3. "I really need sheep"

The adversary believes these statements, resulting in modifying their probabilities of making certain trades.

Note that in the current model we assume that only these 3 manipulative actions potentially have an effect on the adversary's reasoning about the game. An alternative would be to allow all the trading utterances to have some manipulative power. For example the LA's uttering "I will give you a wheat if you give me a rock" could lead the adversary to believe that the LA currently needs rock. For the present work, we prefer to separate out the manipulative actions explicitly, so as to first study their effects in the presence of non-manipulative dialogue actions. In future work, we will consider the case where all trading proposals can cause adversaries to change their game strategy.

3 The Learning Agent (LA)

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. We track up to 19 of each type of resource, and have a binary variable representing whose turn it is. Therefore there are 20 x 20 x 20 x 2 = 16,000 states.

The learning agent (LA) plays the game and learns while perceiving only its own set of resources. This initial state space can later be extended with elements of history (previous dialogue moves) and estimates of the other agent's state (e.g. beliefs about what the adversary needs).

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 (Sutton and Barto, 1998): 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 case, 500 for a draw and -100 when losing a game. The winning and draw cases have the same goal states and that would initially suggest the same reward but they can be achieved through different strategies. Experiments that we have conducted using either the above rewards or the same rewards for win and

draw have verified this. The learning agent's performance is slightly better when the reward for a win is 1000 and 500 for a draw.

The LA was trained using a custom SARSA(λ) learning method (Sutton and Barto, 1998) with an initial exploration rate of 0.2 that gradually decays to 0 at the end of the training games. After experimenting with the learning parameters we found that with λ equal to 0.4 and γ equal to 0.9 we obtain the best results for our problem and therefore these values have been used in all of the experiments that follow.

4 Adversaries

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

A 'gullible' adversary is one who believes statements such as "I really need rock" and then acts so as to restrict the relevant resource(s) from the LA. Our experiments (see experiments 3.1-3.3) show that this gullible behaviour may originate from sound reasoning. The adversary confronts in this case a very important dilemma. It suddenly does not know if it should stay with its goal-oriented strategy (baseline) or instead it should boycott the LA's stated needed resources. A priori, both of these strategies might be equally successful, and we will show that their performances are indeed very close to each other.

4.1 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 will have no effect on it – it simply ignores statements like "I really need wheat".

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 of which it has more than it needs for another one that it still needs. This is a strong strategy that wins a significant number of games. In fact, it takes about 100,000 training games before the LA is able to start winning more games than this adversary, and a random LA policy loses 66% of games against this adversary (See Table 1, LA policy 'Random').

4.2 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, as explained above. For example, if the LA utters "I really need rock", weights of actions which transfer rock from 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 trading resources that it really needs.

4.3 The restrictive adversaries: experiments 3.1, 3.2, 3.3

Here we investigate performance against adversaries who cannot be manipulated, but their strategy is to always restrict the LA from gaining a specific type of resource. We need to explore how well a manipulated adversary (for example one who will no longer give rocks that only its opponent needs) performs. This will show us the potential advantage to be gained by manipulation and most important, it will generalise our problem by showing that the restriction (boycott) of a resource that only the opponent needs, or of a resource that both of the players need, are actually reasonably good strategies compared to the baseline case (Experiment 1). Hence, the manipulated adversary has indeed a reason for choosing to restrict resources (Experiment 2) rather than staying with its rule-based strategy. In other words it has a rational reason to become gullible and fall in the learning agent's trap.

4.4 Gullible-based adversary with risk of exposure: experiments 4.1, 4.2

Here we extend the problem to include possible negative consequences of manipulative LA actions. The adversary begins each game with a probability of detecting manipulation, that exponentially increases after every one of the LA's manipulative actions. In more detail, every time the LA performs a manipulation, there is an additional chance that the adversary notices this (starts at 1-in-10 and increases after every manipulative move, up to 100% in the case of the 10th manipulative attempt). The consequence of being detected is that the adversary will refuse to trade with the LA any further in that game (experiment 4.1), or that the adversary automatically wins the game (experiment 4.2). In these two cases there is always a risk associated with attempting to manipulate, and the LA has to learn how to balance the potential rewards with this risk.

5 Results

The LA was trained over 1.5 million games against each adversary for the cases of the rule-based (experiment 1), gullible (experiment 2) and restrictive adversaries (experiments 3.1, 3.2, 3.3). The resulting policies were tested in 20 thousand games.

For reasons of time, the LA was trained for only 35 thousand games for the case of the gullible adversary who stops trading when the LA becomes exposed (experiment 4.1), and 350 thousand games for the gullible adversary who wins the game when the LA becomes exposed (experiment 4.2). In the former case we used 2 thousand testing games and in the latter 20 thousand.

5.1 Baseline performance: Experiment 1

The LA scored a winning performance of 49.5% against 45.555% for the adversary, with 4.945% draws (Table 1), in the 20 thousand test games, see Figure 1. This represents the baseline performance that the LA is able to achieve against an adversary who cannot be manipulated at all. This shows that the game is 'solvable' as an MDP problem, and that a reinforcement learning agent can outperform a strict hand-coded adversary.

Here, the learning agent's strategy mainly focuses on offering the sheep resource that it does not need for the rocks that does need (for example $action7 > action2 > action6 > action3$ Table 2). It is also interesting to notice that the LA learnt not to use action 3 at all (gives 1 wheat that they both need for 1 sheep that only the adversary needs). Hence its frequency is 0. The actions 4 and 5 are never accepted by the adversary so their role in both of the experiments is similar to that of the action 1 (do nothing). The rejections of the adversary's trades dominate the acceptances with a ratio of 94 to 1 as our learning agent learns to become negative towards the adversarial trading proposals and therefore to prohibit its strategy.

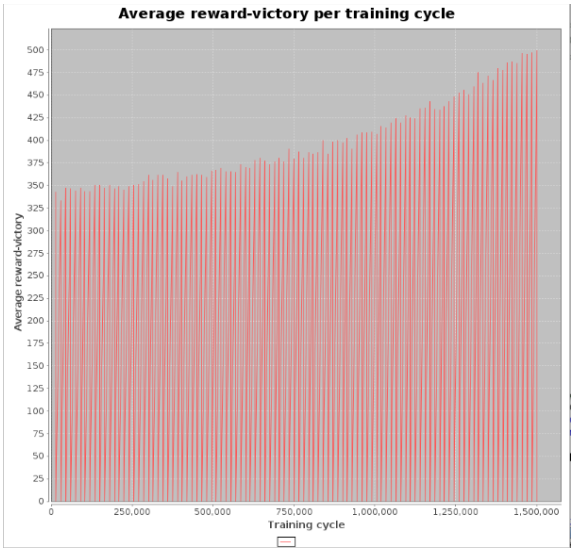

Figure 1: *Learning Agent's reward-victory graph over 1.5 million training games of Experiment 1.*

5.2 Non-cooperative actions: Experiment 2

In Experiment 2 the learning agent scored a winning performance of 59.17% against only 39.755% of its adversary, having 1.075% draws (Table 1), in the 20 thousand test games, see Figure 2.

Similarly to the previous experiment, the LA's strategy focuses again mainly on action 7, by offering the sheep resource that it does not need for rocks that it needs (Table 2). However in this case we also notice that the LA has learnt to use action 2 very often, exploiting cases where it will win by giving the wheat resource that they both need for a rock that only it needs. This is a result of its current manipulation capabilities. The high frequency manipulative actions 8 ("I really need wheat") and 9 ("I really need rock") assist in deceiving its adversary by hiding information, therefore significantly reinforcing its strategy as they both indirectly result in gaining sheep that only the adversary needs (experiment 3.2).

Rejections to adversarial trading offers over the

acceptances were again the majority in this experiment. However in this case they are significantly fewer than before, with a ratio of only 2.5 to 1, as our learning agent is now more eager to accept some trades because it has triggered them itself by appropriately manipulating its adversary.

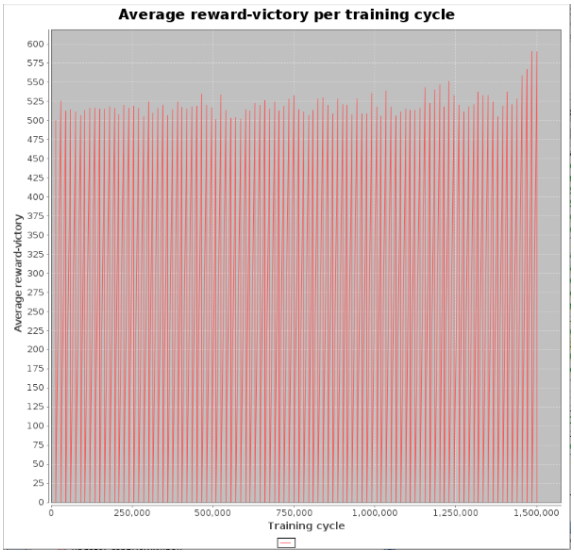

Figure 2: *Learning Agent's reward-victory graph in 1.5 million training games of Experiment 2.*

In Experiment 1 the LA's dominating strategy (mainly based on requiring the rocks resource from its adversary) provides it with a difference in winning performance of $+3.945\%$. In Experiment 2 the adversary, further being deceived by the learning agent's hiding information actions, loses 19.415% more often than the learning agent.

Action number	**Exp. 1** frequency	**Exp. 2** frequency
1. Do nothing	81969	144727
2. Give wheat for rock	8077	46028
3. Give wheat for sheep	**0**	10358
4. Give rock for wheat	80578	62874
5. Give rock for sheep	78542	55627
6. Give sheep for wheat	6429	24687
7. Give sheep for rock	**23888**	31132
8. I really need wheat	-	**68974**
9. I really need rock	-	**87123**
10. I really need sheep	-	18

Table 2: *Frequencies of LA actions.*

Table 2 shows that the LA's strategy in Experiment 1 mainly focuses on requiring rocks from the adversary by offering sheep (for example action 7 $>$ action 2 or 6). In Experiment 2 the agent's strategy is similar. However, it is now enhanced by the frequent use of the manipulative actions 8 and 9 (both hide information). The LA gathers mainly sheep (8 and 9) through its manipulation and then wheat ($9 > 8$) that the adversary needs to win. It also offers them 'selectively' back (2 and 7) for rock that only it needs in order to win.

5.3 Restrictive adversaries: Experiment 3

In experiment 3 the LA uses no manipulative actions. It is the same LA as that of Experiment 1. It is trained and then tested against 3 different types of restrictive adversaries. The first one (Experiment 3.1) never gives wheat, the second one (Experiment 3.2) never gives rocks, and the third one never gives sheep (Experiment 3.3). They all act randomly regarding the other 2 resources which are not restricted. In the first case (adversary restricts wheat that they both need), the LA scored a winning performance of 50.015% against 47.9% of its adversary, having 2.085% draws in the 20 thousand test games. In the second case (adversary restricts rocks that the LA only needs), the LA scored a winning performance of 53.375% against 44.525% of its adversary, having 2.1% draws in the 20 thousand test games. In the third case (adversary restricts sheep that only itself needs), the LA scored a winning performance of 62.21% against 35.13% of its adversary, having 2.66% draws in the 20 thousand test games. These results show that restricting the resource that only the opponent needs (i.e. LA only needs rocks) and especially the resource that they both need (i.e. wheat) can be as effective as the strategy followed by the rule-based adversary (see Table 1). The difference in the performances for the former case (rock) is +8.85% and for the latter (wheat) only +2.115%. That means the adversary has indeed a reason to believe that boycotting its opponent's resources could be a winning opposing strategy, motivating its gullibility in experiment 2 (section 5.2).[1]

5.4 Non-cooperative actions and risk of exposure: Experiment 4.1 (adversary stops trading)

In this case when the LA is exposed by the adversary then the latter does not trade for the rest of the

[1]Further experiments showed that having the same number of goal resources (i.e. both need 4 of their own goal resources, rather than 5) still produces similar results.

Exp.	**Learning Agent policy**	**Adversary policy**	**LA wins**	**Adversary wins**	**Draws**
	Random	Baseline	32%	66%	2%
1	SARSA	Baseline	49.5%	45.555%	4.945%
2	SARSA + Manipulation	Baseline + Gullible	59.17%*	39.755%	1.075%
3.1	SARSA	Restrict wheat	50.015%*	47.9%	2.085%
3.2	SARSA	Restrict rock	53.375%*	44.525%	2.1%
3.3	SARSA	Restrict sheep	62.21%*	35.13%	2.66%
4.1	SARSA + Manipulation	Basel. + Gull. + Expos.(no trade)	53.2%*	45.15%	1.65%
4.2	SARSA + Manipulation	Basel. + Gull. + Expos.(win game)	36.125%	61.15%	2.725%

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

game. The LA scored a winning performance of 53.2% against 45.15% for this adversary, having 1.65% draws in the 2 thousand test games, see Figure 3. This shows that the LA managed to locate a successful strategy that balances the use of the manipulative actions and the normal trading actions with the risk of exposure (Table 3). In more detail, the strategy that the LA uses here makes frequent use of the manipulative actions 8 ("I really need wheat") and 9 ("I really need rock") again which mainly result in the collection of sheep that only its adversary needs to win. Restriction of a resource that the opponent only needs is a good strategy (as our experiment 3.2 suggests) and the LA managed to locate that and exploit it. The next highest frequency action (excluding actions 4 and 5 that mostly lead to rejection from the adversary as it also follows its rule-based strategy) is 7 ("I will give you a sheep if you give me a rock") that is exclusively based on the LA's goal and along with 6 they 'selectively' give back the sheep for goal resources. Rejections to adversary's proposals over the acceptances were in a ratio of approximately 4 to 1. The LA is quite eager (in contrast to the baseline case of experiment 1) to accept the adversary's proposals as it has already triggered them by itself through deception.

5.5 Non-cooperative actions and risk of exposure: Experiment 4.2 (adversary wins the game)

In this case if the LA becomes exposed by the adversary then the latter wins the game. The LA scored a winning performance of 36.125% against 61.15% of its adversary, having 2.725% draws in 20 thousand test games, see Figure 4. It is the only case where the LA so far has not yet found a strategy that wins more often than its adversary, and therefore in future work a larger set of training games will be used. Note that this was only trained for 350 thousand games – we expect better performance with more training. In fact, here we would expect a good policy to perform at least as well as experiment 1, which would be the case of learning never to use manipulative actions, since they are so dangerous. Indeed, a good policy could be to lie (action 10) only once, at the start of a dialogue, and then to follow the policy of experiment 2. This would lead to a winning percentage of about 49% (the 59% of experiment 2 minus a 10% loss for the chance of being detected after 1 manipulation).

The LA has so far managed to locate a strategy that again balances the use of the manipulative actions and that of the normal ones with the risk of losing the game as a result of exposure (Table 3). According to Figure 4 we notice that the LA gradually learns how to do that. However its performance is not yet desirable, as it is still only slightly better than that of the Random case against the Baseline (Table 1). It is interesting though to see that the strategy that the LA uses here makes frequent use of the action 10 ("I really need sheep") that lies. On the other hand, the actions 8 and 9 are almost non-existent. That results in accepting wheat that they both need and rocks that it only needs, showing that the main focus of the manipulation is on the personal goal. The LA has learned so far in this case that by lying it can get closer to its personal goal. Rejections to adversary's proposals over the acceptances resulted in a ratio of approximately 1.7 to 1, meaning that the LA is again quite eager to accept the adversarial trading proposals that it has triggered already by itself through lying.

We report further results on this scenario in an updated version of this paper (Efstathiou and

Lemon, 2014).

Action number	**Exp. 4.1** frequency	**Exp. 4.2** frequency
1 Do nothing	8254	74145
2 Give wheat for rock	2314	3537
3 Give wheat for sheep	1915	4633
4 Give rock for wheat	5564	46120
5 Give rock for sheep	4603	57031
6 Give sheep for wheat	2639	2737
7 Give sheep for rock	3132	3105
8 I really need wheat	**7200**	4
9 I really need rock	**7577**	7
10 I really need sheep	548	**19435**

Table 3: *Frequencies of LA actions.*

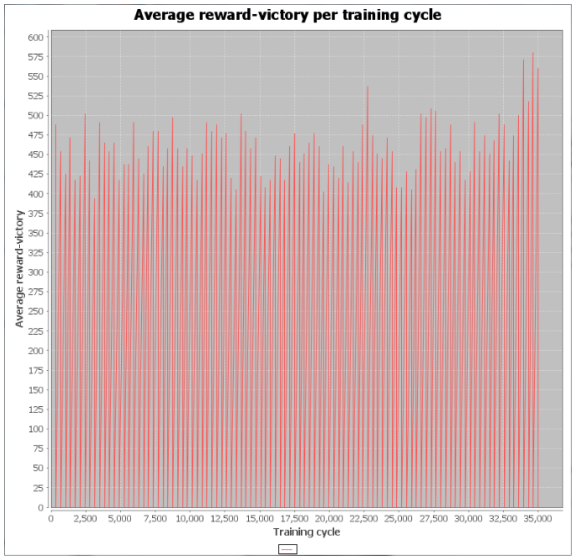

Figure 3: *Learning Agent's reward-victory graph in 35 thousand training games of Experiment 4.1.*

6 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, similarly to the emergence of cooperative behaviour from multi-agent decision theory (Vogel et al., 2013a).

In future work we will investigate more complex non-cooperative situations. For example a real dialogue example of this kind is taken from

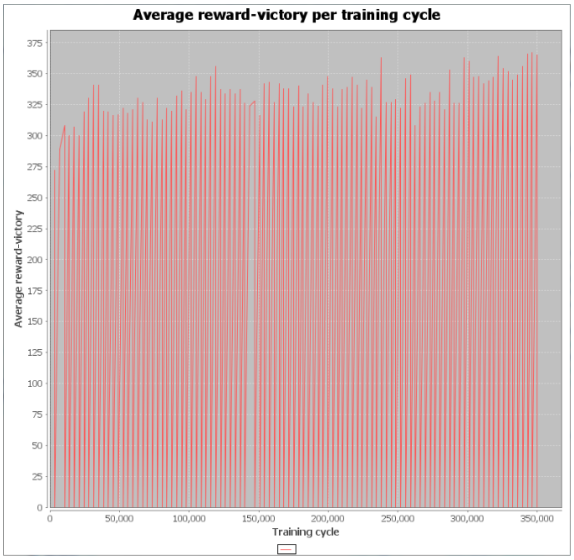

Figure 4: *Learning Agent's reward-victory graph in 350 thousand training games of Experiment 4.2.*

the "Settlers of Catan" game corpus (Afantenos et al., 2012):

- A: Do you have rock?
- B: I've got lots of wheat [in fact, B has a rock]
- A: I'll give you 2 clay for a rock
- B: How about 2 clay for a wheat?
- A: I'll give 1 clay for 3 wheat
- B: Ok, it's a deal.

In future more adversarial strategies will also be applied, and the learning problem will be made more complex (e.g. studying 'when' and 'how often' an agent should try to manipulate its adversary). Alternative methods will also be considered such as adversarial belief modelling with the application of interactive POMDPs (Partially Observable Markov Decision Processes) (Gmytrasiewicz and Doshi, 2005). 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 (Shim and Arkin, 2013), such as education, military operations, virtual worlds and healthcare.

References

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. 2012. Developing a

corpus of strategic conversation in The Settlers of Catan. In *Proceedings of SemDial 2012*.

R. Arkin. 2010. The ethics of robotics deception. In *1st International Conference of International Association for Computing and Philosophy*, pages 1–3.

N. Asher and A. Lascarides. 2008. Commitments, beliefs and intentions in dialogue. In *Proc. of SemDial*, pages 35–42.

S. Attardo. 1997. Locutionary and perlocutionary cooperation: The perlocutionary cooperative principle. *Journal of Pragmatics*, 27(6):753–779.

Daniel Dennett. 1997. When Hal Kills, Who's to Blame? Computer Ethics. In *Hal's Legacy:2001's Computer as Dream and Reality*.

Ioannis Efstathiou and Oliver Lemon. 2014. Learning to manage risk in non-cooperative dialogues. In *under review*.

Kallirroi Georgila and David Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In *Proc. INTERSPEECH*.

Piotr J. Gmytrasiewicz and Prashant Doshi. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79.

Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3.

Verena Rieser and Oliver Lemon. 2011. *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.

J. Shim and R.C. Arkin. 2013. A Taxonomy of Robot Deception and its Benefits in HRI. In *Proc. IEEE Systems, Man, and Cybernetics Conference*.

R. Sutton and A. Barto. 1998. *Reinforcement Learning*. MIT Press.

David Traum. 2008. Exrended abstract: Computational models of non-cooperative dialogue. In *Proc. of SIGdial Workshop on Discourse and Dialogue*.

Adam Vogel, Max Bodoia, Christopher Potts, and Dan Jurafsky. 2013a. Emergence of Gricean Maxims from Multi-Agent Decision Theory. In *Proceedings of NAACL 2013*.

Adam Vogel, Christopher Potts, and Dan Jurafsky. 2013b. Implicatures and Nested Beliefs in Approximate Decentralized-POMDPs. In *Proceedings of ACL 2013*.

M. Walker, R. Passonneau, and J. Boland. 2001. Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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