Detecting Deception in Non-Cooperative Dialogue: A Smarter Adversary Cannot be Fooled That Easily

Aimilios Vourliotakis, Ioannis Efstathiou, and Verena Rieser
Interaction Lab
Heriot-Watt University
Edinburgh, UK
www.macs.hw.ac.uk/InteractionLab

Abstract

Recent work has learned non-cooperative dialogue behaviour within a stochastic trading game, including dialogue moves such as bluffing and lying. Here, we introduce an adversary which can detect deception based on logical contradictions between dialogue moves. Being caught in deception, the adversary will penalise this behaviour by either refusing to trade or declaring victory. We compare our results to a learning agent trained with a gullible adversary and show that a more realistic adversary decreases the chances of winning by over 20%, if the penalty for cheating is to lose the game. In future work we will re-train the learning agent within this more challenging environment.

1 Introduction

Deception in artificial agents has been identified as important in variety of application areas, including education, military operations, video games and healthcare (Traum, 2008; Shim and Arkin, 2013). Recently, dialogue policies have been developed which can execute such non-collaborative behaviour using Reinforcement Learning (Georgila and Traum, 2011; Efstathiou and Lemon, 2014). In this research, we extend previous work by (Efstathiou and Lemon, 2014) by evaluating the learnt policy against an adversary which is able to detect deception based on logical inconsistencies between dialogue moves. In contrast, (Efstathiou and Lemon, 2014) have used a simple frequency-based approach, where the likelihood of detection linearly increases the more the agent lies or bluffs.

In the following, we first summarise the learning framework within a stochastic trading game (Section 2). We then describe three models of detecting deception (Section 3). In Section 4 we present some preliminary results, testing the trained learning agent from (Efstathiou and Lemon, 2014) against our extended adversary models.

2 Learning Non-Cooperative Behaviour in Taikun

Taikun is a 2-player, sequential, non-zero-sum game with imperfect information designed to investigate non-cooperative dialogue in a controlled settings environment (Efstathiou and Lemon, 2014). The goal of the game is for each participant to collect resources (Rock, Wheat and Sheep) via trading or by random game update. In the trading phase a player proposes a 1-for-1 trade of resources and the other player accepts or rejects the proposed trade. In the game update phase the game randomly modifies the resources each player has by adding two or subtracting one of them.

In (Efstathiou and Lemon, 2014) a Learning Agent (LA) is modelled as a Markov Decision Process (MDP) and is trained using SARSA(λ) against a rule-based adversary. In order to introduce deception, the LA was supplemented with additional Manipulation Actions (MAs) in the form of "I really need X", where X is a type of resource. The adversary will then adapt its strategy to not engage in or propose trades where the LA would receive this resource. The LA uses these MAs against the "gullible" adversary in order to mislead him into trading resources he actually needed (**Baseline Scenario**). An advanced scenario introduces a risk of deception detection, where the likelihood of discovery by the adversary is increased after each MA (**Frequency-based Approach**).

3 Detecting Deception

Here we detect deception based on a model of semantic inconsistencies (e.g. contradictions) between dialogue moves. The following examples show how deception could be detected:

Scenario	LA wins		ADV wins		Draws	
Baseline (no detection)	59.170		39.755		1.075	
Detection by:	Refusal to trade	Automatic win	Refusal to trade	Automatic win	Refusal to trade	Automatic win
Case1: Plain Lies	55.725	39.996	42.295	58.895	1.980	1.110
Case1+2: Naive Turn	54.035	35.950	43.920	62.945	2.045	1.105
Case1+3: Probabilistic Turn	54.275	36.985	43.810	62.025	1.915	0.990
Frequency-based	50.86	49.7	46.33	46.225	2.81	4.075

Table 1: Winning rates in % for different adversary models

(1) a. LA: I really need Wheat. (**MA**)
b. ADV: I give you Rock and I need Wheat.
c. LA: Ok! (**Contradiction**)
d. (*Game update*)
e. LA: I give you Wheat and I need Sheep. (**Contradiction**)

Note that in real world face-to-face spoken interaction, deception can also be detected from multimodal cues (Fitzpatrick et al., 2012). In our simulations we consider the following cases:

Case 1: Lies in the same trading-phase (Plain Lies). In Example 1 (a) the LA falsely declares that he needs wheat, while in the next dialogue turn it clearly contradicts itself by giving this resource away, see Example 1 (b).

Case 1+2: Lies in consecutive trading-phase (Naive Turn-based Approach). In addition to Case 1, we consider logical inconsistencies which occur between an MA and a subsequent LA action in the next trading phase, see Example (1e). In this case, we ignore the game update in (1d).

Case 1+3: Likelihood of consecutive lies (Probabilistic Turn-based Approach). This case now accounts for the game update, where the LA randomly receives/ loses resources and thus the probability the MA is still valid decreases by 1/3.

Once a MA is discovered, the lie can be penalised in two different ways, following (Efstathiou and Lemon, 2014):
Refusal to trade: After detecting a MA, the adversary will refuse to further trade with the LA.
Automatic win: After detecting a MA, the adversary will win automatically.

4 Results

We now test the trained learning agent ('Baseline') from (Efstathiou and Lemon, 2014) against our extended adversary models. The results in Table 1 show:

- As expected, the LA trained for a gullible adversary performs worse with adversaries which can detect deception.
- Within our three different cases, detecting plain lies within the same turn has the most effect. There is a negligible difference between detecting lies in consecutive turns between the naive approach (Case 2) and the approach which takes environmental uncertainty into account (Case 3).
- Surprisingly, the adversaries which can detect MAs based on logical contradictions perform worse than the frequency-based adversary. However, note that in this case (Efstathiou and Lemon, 2014) actually re-trained the LA and thus the LA had the chance to adapt to this more challenging scenario, so there is no direct comparison. This difference is highlighted by greying out this result in Table 1.
- Finally, when comparing the effect of penalties, we find that refusal to trade has less impact than automatic win, since there is still a high chance of winning through game updates only.

5 Discusion and Future Work

The above results show that an adversary trained against a gullible agent performs significantly worse against an agent with a more sophisticated technique of detecting deception based on logical contradictions between dialogue moves. This motivates the need for re-training the Learning Agent with these advanced adversaries using Reinforcement Learning (Rieser and Lemon, 2011). We will first target the case where the adversary can only detect lies within the same trading phase (Case 1), which we found to have the main impact on the agents' winning rates. We will present full results at the conference.

Acknowledgments

The research leading to this work has received funding from the European Community's ERC programme under grant agreement no. 269427 (STAC). The authors would like to thank Oliver Lemon for his comments on the draft.

References

Ioannis Efstathiou and Oliver Lemon. 2014. Learning non-cooperative dialogue behaviours. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 60–68, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari, editors. 2012. *EACL 2012: Proceedings of the Workshop on Computational Approaches to Deception Detection*, Avignon, France. Association for Computational Linguistics.

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

Verena Rieser and Oliver Lemon. 2011. *Reinforcement Learning for Adaptive Dialogue Systems*. Theory and Applications of Natural Language Processing. Springer.

Jaeeun Shim and Ronald C. Arkin. 2013. A taxonomy of robot deception and its benefits in HRI. In *SMC*, pages 2328–2335.

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