

Reinforcement Learning of Multi-Issue Negotiation Dialogue Policies

Alexandros Papangelis

Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
apapa@cs.cmu.edu

Kallirroi Georgila

Institute for Creative Technologies
University of Southern California
Playa Vista, CA 90094, USA
kgeorgila@ict.usc.edu

Abstract

We use reinforcement learning (RL) to learn a multi-issue negotiation dialogue policy. For training and evaluation, we build a hand-crafted agenda-based policy, which serves as the negotiation partner of the RL policy. Both the agenda-based and the RL policies are designed to work for a large variety of negotiation settings, and perform well against negotiation partners whose behavior has not been observed before. We evaluate the two models by having them negotiate against each other under various settings. The learned model consistently outperforms the agenda-based model. We also ask human raters to rate negotiation transcripts between the RL policy and the agenda-based policy, regarding the rationality of the two negotiators. The RL policy is perceived as more rational than the agenda-based policy.

1 Introduction

Negotiation is a process in which two or more parties participate in order to reach a joint decision. Negotiators have goals and preferences, and follow a negotiation policy or strategy to accomplish their goals. There has been a lot of work on building automated agents for negotiation in the communities of autonomous agents and game theory. Lin and Kraus (2010) present a quite comprehensive survey on automated agents designed to negotiate with humans. Below we focus only on research that is directly related to our work.

English and Heeman (2005) and Heeman (2009) applied reinforcement learning (RL) to a furniture layout negotiation task. Georgila and Traum (2011) learned argumentation policies against users of different cultural norms in a one-issue negotiation scenario. Then Georgila (2013)

learned argumentation policies in a two-issue negotiation scenario. These policies were trained for some initial conditions, and they could perform well only when they were tested under similar conditions. More recently, Efstathiou and Lemon (2014) learned negotiation behaviors for a non-cooperative trading game (the Settlers of Catan). Again, in Efstathiou and Lemon (2014)'s work, the initial settings were always the same. Georgila et al. (2014) used multi-agent RL to learn negotiation policies in a resource allocation scenario. They compared single-agent RL vs. multi-agent RL and they did not deal with argumentation, nor did they allow for a variety of initial conditions. Finally, Hiraoka et al. (2014) applied RL to the problem of learning cooperative persuasive policies using framing. Due to the complexity of negotiation tasks, none of the above works dealt with speech recognition or understanding errors.

In this paper, we focus on two-party negotiation, and use RL to learn a multi-issue negotiation policy for an agent aimed for negotiating with humans. We train our RL policy against a simulated user (SU), which plays the role of the other negotiator. Our SU is a hand-crafted negotiation dialogue policy inspired by the *agenda paradigm*, previously used for dialogue management (Rudnicky and Xu, 1999) and user modeling (Schatzmann and Young, 2009) in information providing tasks.

Both the agenda-based and the RL policies are designed to work for a variety of goals, preferences, and negotiation moves, even under conditions that are very different from the conditions that the agents have experienced before. We vary the goals of the agents, how easy it is for the agents to be persuaded, whether they have enough arguments to accomplish their goals (i.e., shift their partners' preferences), and the importance of each issue for each agent. We evaluate our two models by having them negotiate against each other under various settings. We also ask human raters to rate

negotiation transcripts between the RL policy and the agenda-based SU, regarding the rationality of the two negotiators.

In our negotiation task, both the agenda-based SU and the RL policy have human-like constraints of imperfect information about each other; they do not know each other's goals or preferences, number of available arguments, degree of persuadability, or degree of rationality. Furthermore, both agents are required to perform well for a variety of negotiation settings, and against opponents whose negotiation behavior has not been observed before and may vary from one interaction to another or even within the same interaction. Thus our negotiation task is very complex and it is not possible (or at least it is very difficult) to compute an analytical solution to the problem using game theory.

Our contributions are as follows. First, this is the first time in the literature that the agenda-based paradigm is applied to negotiation. Second, to our knowledge this is the first time that RL is used to learn so complex multi-issue negotiation and argumentation policies (how to employ arguments to persuade the other party) designed to work for a large variety of negotiation settings, including settings that did not appear during training.

2 Agenda-Based Negotiation Model

The original agenda-based SU factors the user state S into an agenda A and a goal G (Schatzmann and Young, 2009), and was used in a restaurant recommendation dialogue system. We replaced the constraints and requests (which refer to slot-value pairs) with *negotiation goals* and *negotiation profiles*, and designed new rules for populating the agenda.

The agenda can be thought of as a stack containing the SU's pending actions, also called speech acts (SAs), that are required for accomplishing the SU's goal. For example, the agenda could be initialized with offers for each issue (with the values preferred by the SU) and with requests for the opponent's preferences for each issue. Based on hand-crafted rules, new SAs are generated and pushed onto the agenda as a response to the opponent's actions. For example, if the opponent requests the SU's preference for an issue, a SA for providing this preference will be pushed onto the agenda and no longer relevant SAs will be removed from the agenda. When the SU is ready to respond, one or more SAs will be popped off the agenda based on a probability distribution. In our experiments, the maximum number of SAs

that can be popped at the same time is 4 based on a probability distribution (popping 1 SA is more likely than popping 2 SAs, etc.).

The set of available SAs is: Offer(issue, value), TradeOff(issue₁, value₁, issue₂, value₂), ProvideArgument(issue, value, argument-strength), ProvidePreference(issue, value), RequestPreference(issue), Accept(issue₁, value₁, issue₂, value₂), Reject(issue, value), ReleaseTurn, and Null. TradeOff is a special action, where the agent commits to accept value₁ for issue₁, on the condition that the opponent accepts value₂ for issue₂. Accept refers to a TradeOff when all four arguments are present, or to an Offer when only two arguments are present. An agent is not allowed to partially accept a TradeOff. The agenda-based SU's internal state consists of the following features: "self standing offers", "self standing trade-offs", "agreed issues", "rejected offers", "self negotiation profile", "self goals", "opponent's standing offers", "opponent's standing trade-offs", "estimated opponent's goal", "estimated opponent's persuadability", "negotiation focus".

The *negotiation profile* models useful characteristics of the SU, such as persuadability, available arguments, and preferred/acceptable values (possible outcomes) for each issue. *Negotiation goals* represent the agent's best value (of highest preference) for each issue. *Negotiation focus* represents the current value on the table for each issue. Persuadability is defined as low, medium, or high, and reflects the number of arguments that the agent needs to receive to be convinced to change its mind. Arguments for an issue can be either strong or weak. We define strong arguments to count for 1 "persuasion point" and weak arguments to count for 0.25. Any combination of strong and weak arguments, whose cumulative points surpass the agent's persuadability (10 points for low, 5 points for medium, and 2 points for high persuadability), are enough to convince the agent and shift its negotiation goal for one issue. Also, the agent has a set number of arguments for each issue, not for each issue-value pair (this will be addressed in future work). Apart from persuadability, we model how important each issue is for the agent (a real number from 0 to 1). Rules, concerning whether a TradeOff or Offer should be accepted or not, take into account issue importance and number of available arguments for that issue (to see if there is any chance to convince the opponent).

There is a number of parameters used to con-

figure the SU: number of issues under negotiation and possible values for each issue (in our setup 4 and 3 respectively); probability of number of SAs popped (this is based on a probability distribution as explained above); and minimum and maximum available arguments per issue (this applies separately to strong and weak arguments and in our setup is 0 and 4 respectively). The SU also keeps track of an estimate of the opponent’s persuadability and the opponent’s goal. These estimates are more accurate for longer dialogues. Table 3 (in the Appendix) shows an example interaction between the SU and another agent, including how the agenda is updated.

3 Negotiation Policy Learning

To deal with the very large state space, we experimented with different feature-based representations of the state and action spaces, and used Q-learning with function approximation (Szepesvári, 2010). We used 10 state-action features: “issue and value under negotiation”, “are there enough arguments to convince the opponent?”, “will my offer be accepted?”, “opponent’s offer quality”, “opponent’s trade-off quality”, “are there pending issues?”, “is there agreement for the current issue?”, “is the agreed-upon value for the current issue good?”, “importance of current issue”, “current action”.

We worked on a *summary state space*, rather than the *full state space*. The full state space keeps track of the interaction in detail, e.g., what offers have been made exactly, and the summary state space keeps track of more abstract representations, e.g., whether an offer was made, out of which we extract the 10 state-action features that the RL policy uses to make decisions. This is also similar to how our agenda-based SU works; rules, that decide on e.g., whether a trade-off should be proposed or accepted, take into account the opponent’s estimated persuadability and context of the interaction, in essence allowing the agent to operate on a summary state space.

The learning algorithm was trained for 5 epochs (batches) of 20000 episodes each, with a limit to 35 iterations per episode, and was tuned with the following parameter values: α set to 0.95, decayed by $\frac{1}{1+N(s,a)}$ after each episode, where $N(s,a)$ is the number of times the state-action pair (s,a) has been explored so far, and γ set to 0.15. We varied the exploration rate ϵ . Initially it was set to 1, gradually decreasing until in the last epoch it was close to 0. To ensure that the policies did not converge

by chance, we ran the training and test sessions 10 times each and we report averages. Thus all results presented below are averages of 10 runs.

In our reward function (regular reward), we penalized each turn if no agreement was reached or, in the opposite case, assigned a reward value inversely proportional to how far the agreed-upon values are from the agent’s preferences.

During training we discovered that this reward function fails to capture the fact that depending on the initial conditions (agents’ goals, number of arguments, etc.) it may not be possible to reach an agreement or to achieve one’s goals. Therefore, we also calculated the best achievable score (BAS) of the policy, which is the best possible score that the agent can achieve given its resources (number of strong and weak arguments), the opponent’s persuadability, and assuming the best possible circumstances (i.e., that the opponent is very cooperative and accepts everything).

To assess whether Q-learning has converged, we calculate a normalized score, reflecting how well the goals were achieved, similar to the regular reward function presented above. The difference is that we do not have a turn penalty and that the maximum penalty is set lower (in training the penalty for sub-optimal agreements was higher to ensure that the policy learns to avoid such cases).

Figure 1 shows the scores of the policy and the SU as a function of the training episodes, when we use the regular reward. We can also see the BAS for both the RL policy and the SU. The maximum possible value for each agent is 100 (the agent accomplishes its exact goals) and the minimum is 0 (there is no agreement for any issue at all). In the last training epoch the exploration rate ϵ is almost 0, and the RL policy consistently outperforms the SU. During training, in each episode, we randomly initialize the following settings for both agents: number of available strong and weak arguments, persuadability per issue, importance per issue, and preferences per issue.

4 Evaluation

For our evaluation, we have the RL policy interact with the agenda-based SU for 20000 episodes varying the initial settings for both agents in the same fashion as for training. Similarly to training, we have 10 runs and report averages (see Figure 1). The RL policy outperforms the agenda-based SU. The RL policy learned to exploit trade-offs that while not being optimal for the SU, they are good enough for the SU to accept (the SU is

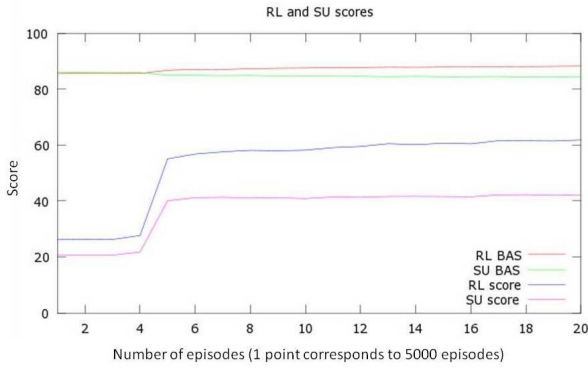


Figure 1: Average scores as a function of the number of episodes during training (10 runs). In the last 20000 episodes the exploration rate ϵ is almost 0 (similarly to testing).

designed to accept only trade-offs and offers that lead to reasonable agreements). Note that some decisions of the SU about what to accept are based on inaccurate estimates of its opponent’s persuadability and goals.

Table 1 reports results about the success percentages of the RL policy and the agenda-based SU. We show on average how many times (10 runs) the agents fully succeeded in their goals (score equal to 100), how many times they achieved roughly at least their second best values for all issues (score > 65), and how many times they achieved roughly at least their third best values for all issues (score > 30). A higher than 65 score can also be achieved when an agent achieves the best possible outcome in some of the issues and the third possible outcome in the rest of the issues. Likewise for scores greater than 30.

In a second experiment we asked human raters to rate negotiation transcripts between the agenda-based SU and the RL policy. The domain was organizing a party. The negotiators had to agree on 4 issues (food type, drink, music, day of week) and there were 3 possible values per issue. We replaced the speech acts with full sentences but for arguments we used sentences such as “here is a strong argument supporting jazz for music”. We randomly selected 20 negotiations between the RL policy and the agenda-based SU. In 10 of those the RL policy earned more points, and in the other 10 the agenda-based SU earned more points. This was to ensure that the transcripts were balanced and that we had not picked only transcripts where one of the agents was always better than the other. We did not tell raters that these were artificial dialogues. We deliberately included some questions with rather obvious answers (sanity checks)

to check how committed the raters were. We recruited raters from MTurk (www.mturk.com). We asked raters to read 2 transcripts and for each transcript rate the negotiators in terms of how rationally they behaved, on a Likert scale from 1 to 5. We excluded ratings that were done in less than 3 minutes and that had failed in more than half of our sanity checks. In total there were 6 sanity checks (3 per negotiation transcript). Thus we ended up with 89 raters. Results are shown in Table 2. The RL policy was perceived as more rational, and both agents were rated as reasonably rational. Interestingly, rationality was perceived differently by different human raters, e.g., revisiting an agreed-upon issue was considered as rational by some and irrational by others.

| | Full success (%) | At least second choice (%) | At least third choice (%) |
|--------------|------------------|----------------------------|---------------------------|
| Policy Score | 10.3 | 30.7 | 53.5 |
| SU Score | 0 | 11.2 | 55.1 |
| Policy BAS | 20.2 | 73.3 | 100 |
| SU BAS | 18.1 | 75.8 | 100 |

Table 1: Average success percentages (10 runs).

| | |
|-----------------------|-------|
| Learned Policy Score | 3.43 |
| Agenda-based SU Score | 3.02 |
| p-value | 0.027 |

Table 2: Human evaluation scores (the p-value is based on the Wilcoxon signed-rank test).

5 Conclusion

We built a hand-crafted agenda-based SU, which was then used together with RL to learn a multi-issue negotiation policy. Both the agenda-based SU and the RL policy were designed to work for a variety of goals, preferences, and negotiation moves. In both of our evaluation experiments, the learned model consistently outperformed the agenda-based SU, even though both models used similar features and heuristics, which shows the potential of using RL for complex negotiation domains. For future work, we plan to work on better estimates of the opponent’s persuadability and goals, and employ multi-agent RL techniques (Bowling and Veloso, 2002; Georgila et al., 2014). Finally, we will have our policies directly negotiate with humans.

Acknowledgments

This work was funded by the NSF Grant #1117313.

References

- Michael Bowling and Manuela Veloso. 2002. Multi-agent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250.
- Ioannis Efstathiou and Oliver Lemon. 2014. Learning non-cooperative dialogue behaviours. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue*, Philadelphia, Pennsylvania, USA.
- Michael S. English and Peter A. Heeman. 2005. Learning mixed initiative dialogue strategies by using reinforcement learning on both conversants. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vancouver, Canada.
- Kallirroi Georgila and David Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In *Proc. of Interspeech*, Florence, Italy.
- Kallirroi Georgila, Claire Nelson, and David Traum. 2014. Single-agent vs. multi-agent techniques for concurrent reinforcement learning of negotiation dialogue policies. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland, USA.
- Kallirroi Georgila. 2013. Reinforcement learning of two-issue negotiation dialogue policies. In *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue*, Metz, France.
- Peter A. Heeman. 2009. Representing the reinforcement learning state in a negotiation dialogue. In *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Merano, Italy.
- Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Reinforcement learning of cooperative persuasive dialogue policies using framing. In *Proc. of COLING*, Dublin, Ireland.
- Raz Lin and Sarit Kraus. 2010. Can automated agents proficiently negotiate with humans? *Communications of the ACM*, 53(1):78–88.
- A. Rudnicky and W. Xu. 1999. An agenda-based dialogue management architecture for spoken language systems. In *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Keystone, Colorado, USA.
- Jost Schatzmann and Steve Young. 2009. The hidden agenda user simulation model. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):733–747.
- Csaba Szepesvári. 2010. *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers.

| |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Initial Conditions: SU: args(strong: 3, weak: 1), persuadability: high prefs: A: a ₁ -a ₂ -a ₃ , B: b ₃ -b ₁ -b ₂ , C: c ₁ -c ₃ -c ₂ , D: d ₁ -d ₃ -d ₂ importance: 0.2,0.2,0.7,0.9 |
| Agent: args(strong: 2, weak: 2), persuadability: high prefs: A: a ₂ -a ₁ -a ₃ , B: b ₂ -b ₁ -b ₃ , C: c ₁ -c ₂ -c ₃ , D: d ₃ -d ₂ -d ₁ importance: 0.7,0.3,0.3,0.9 |
| Agenda: ReqB,ReqC,ReqD,OffA,OffB,OffC,OffD,ReqA (the agenda is randomly initialized) SA: RequestPreference(issueB) SU: <i>What do you prefer for issue B?</i> Agenda: ReqC,ReqD,OffA,OffB,OffC,OffD,ReqA |
| SA: ProvidePreference(issueB) Agent: <i>I prefer b₂ for issue B.</i> Agenda: ArgB ,ReqC,ReqD,OffA,OffB,OffC,OffD,ReqA SA: ProvideArgument(issueB,b ₃ ,strong) SU: <i>Here's a strong argument, supporting b₃ for issue B.</i> Agenda: ReqC,ReqD,OffA,OffB,OffC,OffD,ReqA |
| SA: Trade-off(issueB,b ₃ ,issueA,a ₂), Offer(issueC,c ₁) Agent: <i>I'll accept b₃ for issue B, if you accept a₂ for issue A.</i> Agent: <i>How about c₁, for issue C?</i> Agenda: ArgB,AccC ,ReqC,ReqD,OffA,OffB,OffC,OffD,ReqA SA: ProvideArgument(issueB,b ₃ ,strong) SU: <i>Here's a strong argument, supporting b₃ for issue B.</i> Agenda: AccC,ReqC,ReqD,OffA,OffB,OffC,OffD,ReqA |
| SA: Offer(issueB,b ₃) Agent: <i>How about b₃, for issue B?</i> (after 2 strong arguments the agent is convinced) Agenda: AccB,AccC ,ReqC,ReqD,OffA,OffB,OffC,OffD,ReqA SA: Accept(issueB,b ₃) SU: <i>I accept b₃, for issue B.</i> SA: Accept(issueC,c ₁) SU: <i>I accept c₁, for issue C.</i> Agenda: ReqD,OffA,OffD,ReqA |

Table 3: Example interaction between our agenda-based SU and an agent. Actions that are pushed onto the SU agenda based on previous actions of the agent are shown in bold.