

The Complex Negotiation Dialogue Game

Romain Laroche

Microsoft Maluuba, Montréal, Canada
romain.laroche@microsoft.com

Abstract

This position paper formalises an abstract model for complex negotiation dialogue. This model is to be used for the benchmark of optimisation algorithms ranging from Reinforcement Learning to Stochastic Games, through Transfer Learning, One-Shot Learning or others.

1 Introduction

A negotiation is defined as a *bargaining process between two or more parties (each with its own aims, needs, and viewpoints) seeking to discover a common ground and reach an agreement to settle a matter of mutual concern or resolve a conflict*. From a dialogue point of view, one distinguishes negotiation dialogue from standard dialogue by the mutual sharing of information¹, by its required user adaptation², and by the non-stationarity induced by its non fully cooperative structure: the user and system objectives correlate but also differ to some extent, and they are consequently adversely co-adapting.

Research on negotiation dialogue experiences a growth of interest. At first, Reinforcement Learning (Sutton and Barto, 1998), the most popular framework for dialogue management in dialogue systems (Levin and Pieraccini, 1997; Laroche et al., 2009; Lemon and Pietquin, 2012), was applied to negotiation with mitigated results (English and Heeman, 2005; Georgila and Traum, 2011; Lewis et al., 2017), because the non-stationary policy of the opposing player prevents those algorithms from converging consistently. Then, Multi-Agent Reinforcement Learning (Bowling and Veloso, 2002) was applied but also with convergence

difficulties (Georgila et al., 2014). Finally, recently, Stochastic Games (Shapley, 1953) were applied successfully (Barlier et al., 2015), with convergence guarantees, but only for zero-sum games, which is inconsistent with dialogue since most tasks are cooperative.

Here, we extend (Laroche and Genevay, 2017)'s abstraction of the negotiation dialogue literature applications: (di Eugenio et al., 2000; English and Heeman, 2005) consider sets of furniture, (Afantenos et al., 2012; Efstathiou and Lemon, 2014; Georgila et al., 2014; Litman et al., 2016; Lewis et al., 2017) resource trading, and (Putois et al., 2010; Laroche et al., 2011; El Asri et al., 2014; Genevay and Laroche, 2016; Laroche and Féraud, 2017) appointment scheduling. Indeed, these negotiation dialogue problems are cast into a generic agreement problem over a shared set of options. The goal for the players is to reach an agreement and select an option. This negotiation dialogue game can be parametrised to make it zero-sum, purely cooperative, or general sum. However, (Laroche and Genevay, 2017) only consider elementary options: they are described through a single entity.

We formalise in this paper the game for options that are compounded in the sense that they are characterised by several features. For instance, *Tuesday morning* is defined by two features: the day and the moment of the day. Considering compounded options naturally leads to richer expressions, and therefore to a larger set of actions: *I'm available whenever on Tuesday*, or *I'd prefer in the afternoon*. Since the options are uttered in a compounded way, as opposed to their elementary definition in (Laroche and Genevay, 2017), the state representation also becomes more complex. This extension allows more realistic dialogues, and more challenging Reinforcement Learning, Multi-Agent Reinforcement Learning,

¹whereas standard dialogue mainly relies on discovering the user information or intent,

²whereas standard dialogue, such as form filling applications, is rather indifferent to the user's characteristics,

and Stochastic Games policy training.

2 The Negotiation Dialogue Game

This section recalls the negotiation dialogue game as described in (Laroche and Genevay, 2017). The goal for each participant is to reach an agreement. The game involves a set of m players $\mathcal{P} = \{\mathcal{P}^i\}_{i \in [1, m]}$. With $m > 2$, the dialogue game is said to be multi-party (Asher et al., 2016; ?). The players consider n options (in resource trading, an option is an exchange proposal, in appointment scheduling, it is a time-slot), and the cost to agree on an option τ is c_τ^i randomly sampled from distribution $\delta^i \in \Delta_{\mathbb{R}^+}$ to agree on it. Players also have a utility $\omega^i \in \mathbb{R}^+$ for reaching an agreement. For each player, a parameter of cooperation with the other players $\alpha^i \in \mathbb{R}$ is introduced. As a result, player \mathcal{P}^i 's immediate reward at the end of the dialogue is:

$$R^i(s_T^i) = \omega^i - c_\tau^i + \alpha^i \sum_{j \neq i} (\omega^j - c_\tau^j) \quad (1)$$

where s_T^i is the last state reached by player \mathcal{P}^i at the end of the dialogue, and τ is the agreed option. If players fail to agree, the final immediate rewards $R^i(s_T^i) = 0$ for all players \mathcal{P}^i . If at least one player \mathcal{P}^j misunderstands and agrees on a wrong option τ^j which was not the one proposed by the other players, this is even worse: each player \mathcal{P}^i gets the cost of selecting option τ^i without the reward of successfully reaching an agreement:

$$R^i(s_T^i) = -c_{\tau^i}^i - \alpha^i \sum_{j \neq i} c_{\tau^j}^j \quad (2)$$

The values of α^i give a description of the nature of the players, and therefore of the game as modelled in game theory (Shapley, 1953). If $\alpha^i < 0$, player \mathcal{P}^i is said to be antagonist: he has an interest in making the other players lose. In particular, if $m = 2$ and $\alpha^1 = \alpha^2 = -1$, it is a zero-sum game. If $\alpha^i = 0$, player \mathcal{P}^i is said to be self-centred: he does not care if the other player is winning or losing. Finally, if $\alpha^i > 0$, player \mathcal{P}^i is said to be cooperative, and in particular, if $\forall i \in [1, m]$, $\alpha^i = 1$, the game is said to be fully cooperative because $\forall (i, j) \in [1, m]^2$, $R^i(s_T^i) = R^j(s_T^j)$.

From now on, and until the end of the article, we suppose that there are only $m = 2$ players: a system \mathcal{P}_s and a user \mathcal{P}_u . They act each one in turn, starting randomly by one or the other. They

have four possible actions. ACCEPT(τ) means that the user accepts the option τ (independently from the fact that τ has actually been proposed by the other player; if it has not, this induces the use of Equation 2 to determine the reward). This act ends the dialogue. REFPROP(τ) means that the user refuses the proposed option and proposes instead option τ . ASKREPEAT means that the player asks the other player to repeat his proposition. And finally, ENDDIAL denotes the fact that the player does not want to negotiate anymore, and terminates the dialogue.

Understanding through speech recognition of system \mathcal{P}_s is assumed to be noisy with a sentence error rate $SE R_s^u$ after listening to a user \mathcal{P}_u : with probability $SE R_s^u$, an error is made, and the system understands a random option instead of the one that was actually pronounced. In order to reflect human-machine dialogue reality, a simulated user always understands what the system says: $SE R_s^s = 0$. We adopt the way (Khouzaimi et al., 2015) generates speech recognition confidence scores: $score_{reco} = \frac{1}{1+e^{-X}}$ where $X \sim \mathcal{N}(c, 0.2)$ given a user \mathcal{P}_u , two parameters (c_\perp^u, c_\top^u) with $c_\perp^u < c_\top^u$ are defined such that if the player understood the right option, $c = c_\top^u$ otherwise $c = c_\perp^u$. The further apart the normal distribution centres are, the easier it will be for the system to know if it understood the right option, given the score.

3 Allowing compounded options

This section extends the negotiation dialogue game recalled in Section 2 with compounded options. Each option τ is now characterised by a set of ℓ features: $\tau = \{f_\tau^k\}_{k \in [1, \ell]}$, with $f_\tau^k \in \mathcal{F}^k$. Not all feature combinations might form a valid option, but for the sake of simplicity, we consider that the set of the n options contain all of them and that the cost for inconsistent ones is infinite. This way, we can express that an option is invalid but that the user is not aware of it.

The cost of an option needs to be revisited consequently. The costs of two options that only differ by a feature are similar in general. Without loss of generality, we define the cost of one player \mathcal{P}^i for agreeing on a given option τ as follows:

$$c_\tau^i = \hat{c}_\tau^i + \sum_{k=1}^{\ell} \hat{c}_k^i, \quad (3)$$

where \hat{c}_k^i is the cost of agreeing on feature f^k and

\hat{c}_τ^i is the cost for selecting this option in particular. In an appointment scheduling negotiation task, the feature related costs \hat{c}_k^i can generally be considered as null: there is no correlation between being booked on Monday morning and being available on Tuesday morning. Most of the constraints are therefore expressed in the \hat{c}_τ^i term. On the opposite, in a furniture set application, the preferences are expressed on specific features of the furniture: colour, price, etc. In this case, the constraints mainly lie in \hat{c}_k^i terms.

The option-as-features definition naturally induces new ways of expressing one's preferences over the option set. The PROPFEATURES(f^{k_1}, f^{k_2}, \dots) dialogue act replaces the previously defined REFPROP(τ): it means that the speaker wants the $k_1^{\text{th}}, k_2^{\text{th}}, \dots$ features to be set to values f^{k_1}, f^{k_2}, \dots . ASKREPEAT still asks to repeat the whole last utterance, but its partial version is added: ASKPARTIALREPEAT(k_1, k_2, \dots) consists in asking to repeat values of features k_1, k_2, \dots . ACCEPT still accepts the last grounded option, but it can only be performed once all features have been grounded. Its partial version is also introduced: PARTIALACCEPT(k_1, k_2, \dots) determines an agreement on the last grounded value of features k_1, k_2, \dots .

The compounded options imply complex actions, which in turn imply a complex understanding model: the sentence level understanding rate and score need to be extended. The sentence error rate SER_s^u is therefore replaced with a feature error rate FER_s^u . The same speech recognition confidence score generation is used at the feature level, meaning that, at each PROPFEATURES, ASKPARTIALREPEAT, and PARTIALACCEPT acts, the player receives an array of feature values (or feature names), each associated with a confidence score.

4 Potential use of the complex negotiation dialogue game

(Genevay and Laroche, 2016) already used the simple negotiation dialogue game to study Knowledge Transfer for Reinforcement Learning (Taylor and Stone, 2009; Lazaric, 2012) applied to dialogue systems (Gašić et al., 2013; Casanueva et al., 2015). It appears in this paper that the optimal policies are rather simple. Making the interaction process more intricate and more reality reflecting allows to put the computational

tractability of the methods to the test. Following the same purpose, one-shot learning (Schaal et al., 1997; Fei-Fei et al., 2006) may also be used for negotiation dialogues.

Cooperative co-adaptation in dialogue has been tackled only in one previous article: (Chandramohan et al., 2012). Similarly, but for the adversary case, the negotiation dialogue game offers a good empirical test bed for a generalisation to the general-sum games of (Barlier et al., 2015).

We believe that this line of research is complementary with the more applied one of (Lewis et al., 2017) that work on real human dialogues and are more focused on dealing with natural language within a negotiation task. Their mitigated results indicate that negotiation generalisation over simulated users to real users is difficult, even when the simulated user is trained on human data.

References

- Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cédric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, et al. 2012. Developing a corpus of strategic conversation in the settlers of catan. In *SeineDial 2012-The 16th Workshop On The Semantics and Pragmatics Of Dialogue*.
- Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *Proceedings of the 11th Edition of Language Resources and Evaluation Conference (LREC)*.
- Merwan Barlier, Julien Perolat, Romain Laroche, and Olivier Pietquin. 2015. Human-machine dialogue as a stochastic game. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (Sigdial)*.
- Michael Bowling and Manuela Veloso. 2002. Multiagent learning using a variable learning rate. *Artificial Intelligence* 136(2):215–250.
- Inigo Casanueva, Thomas Hain, Heidi Christensen, Richard Marxer, and Phil Green. 2015. Knowledge transfer between speakers for personalised dialogue management. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (Sigdial)*.
- Senthilkumar Chandramohan, Matthieu Geist, Fabrice Lefèvre, and Olivier Pietquin. 2012. Co-adaptation in Spoken Dialogue Systems. In *Proceedings of the 4th International Workshop on Spoken Dialogue Systems (IWSDS)*. Paris, France, page 1.

- Barbara di Eugenio, Pamela W. Jordan, Richmond S. Thomason, and Johanna D. Moore. 2000. The agreement process: an empirical investigation of humanhuman computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies* 53(6):1017 – 1076.
- 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)*.
- Layla El Asri, Remi Lemonnier, Romain Laroche, Olivier Pietquin, and Hatim Khouzaimi. 2014. Nastia: Negotiating appointment setting interface. In *Proceedings of the 9th Edition of Language Resources and Evaluation Conference (LREC)*.
- Michael S English and Peter A Heeman. 2005. Learning mixed initiative dialogue strategies by using reinforcement learning on both conversants. In *Proceedings of the conference on Human Language Technology (HLT)*.
- Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28(4):594–611.
- Milica Gašić, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young. 2013. Pomdp-based dialogue manager adaptation to extended domains. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (Sigdial)*.
- Aude Genevay and Romain Laroche. 2016. Transfer learning for user adaptation in spoken dialogue systems. In *Proceedings of the 15th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. International Foundation for Autonomous Agents and Multiagent Systems.
- Kallirroi Georgila, Claire Nelson, and David Traum. 2014. Single-agent vs. multi-agent techniques for concurrent reinforcement learning of negotiation dialogue policies. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Kallirroi Georgila and David R Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*. pages 2073–2076.
- Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre. 2015. Optimising turn-taking strategies with reinforcement learning. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (Sigdial)*.
- Romain Laroche and Raphaël Féraud. 2017. Algorithm selection of off-policy reinforcement learning algorithm. *arXiv preprint arXiv:1701.08810*.
- Romain Laroche and Aude Genevay. 2017. The negotiation dialogue game. In *Dialogues with Social Robots*, Springer, pages 403–410.
- Romain Laroche, Ghislain Putois, Philippe Bretier, Martin Aranguren, Julia Velkovska, Helen Hastie, Simon Keizer, Kai Yu, Filip Jurcicek, Oliver Lemon, and Steve Young. 2011. D6.4: Final evaluation of classic towninfo and appointment scheduling systems. *Report D6 4*.
- Romain Laroche, Ghislain Putois, Philippe Bretier, and Bernadette Bouchon-Meunier. 2009. Hybridisation of expertise and reinforcement learning in dialogue systems. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)*. pages 2479–2482.
- Alessandro Lazaric. 2012. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, Springer, pages 143–173.
- Oliver Lemon and Olivier Pietquin. 2012. *Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces*. Springer.
- Esther Levin and Roberto Pieraccini. 1997. A stochastic model of computer-human interaction for learning dialogue strategies. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*.
- Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.
- Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues .
- Ghislain Putois, Romain Laroche, and Philippe Bretier. 2010. Online reinforcement learning for spoken dialogue systems: The story of a commercial deployment success. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Citeseer, pages 185–192.
- Stefan Schaal et al. 1997. Learning from demonstration. *Advances in neural information processing systems* pages 1040–1046.
- Lloyd S Shapley. 1953. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America* 39(10):1095.
- Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.
- Matthew E Taylor and Peter Stone. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10:1633–1685.