Linguistic Assistance Games: Grounding Natural Language Assistance in Rational Speech Act Theory

Cameron Jordan[1], Alane Suhr[1]

[1]Electrical Engineering and Computer Sciences, University of California Berkeley
cameronjordan@berkeley.edu

Introduction

Cooperative inverse reinforcement learning (CIRL), also known as assistance games [4], provide formal guarantees of alignment and corrigibility that are notably absent from models trained under the standard model of reinforcement learning. Extending CIRL to the natural language domain has remained challenging due, in part, to the inherent ambiguity of natural language, which is precisely the phenomenon that computational pragmatics has long sought to model, particularly through Bayesian inference in frameworks such as Rational Speech Acts (RSA). This work is a first step towards providing similar formal guarantees in the context of linguistic assistance, by grounding natural language assistance in rational speech act theory.

We leverage the insight that RSA [1, 2] can be formulated as a special case of bounded rationality in the CIRL game over natural language utterances as shared components of the state and action space (Theorem 1), in order to motivate pragmatic reasoning both as the inferential process by which a model resolves ambiguity over a human's reward parameter, and as the basis for a boundedly rational model of the human speaker. In particular, the human speaker $\mathbf{H}$ is boundedly rational in the sense that they internally model the robot $\mathbf{R}$ as a literal listener (L_0 in the RSA tradition), and is modeled as being epistemically myopic, in the sense that they assume that they will act as the literal speaker (S_0) (and $\mathbf{R}$ will act as L_0) in future turns. $\mathbf{R}$ updates its belief state $b^{L_1}(\theta)$ over $\mathbf{H}$'s true reward parameter θ as the pragmatic listener (L_1), with a complete understanding of the game dynamics.

Notably, the robot's formal objective coincides exactly with the human's true preference parameter θ. By Definition 2, the reward function $R(u, s; \theta)$ depends only on a human's utterance, a mutually observable state, and the hidden preference parameter θ. The robot's objective is to maximize the discounted, expected sum of rewards, which is exactly the expected cumulative reward under the human's true preference θ (since $\mathbf{R}$'s belief state b^{L_1} serves as a sufficient statistic for θ). There is no additional term in $\mathbf{R}$'s reward function, and therefore the robot's interests are structurally aligned with the human's interests by construction, following the shared-payoff property of standard CIRL.

Finally, note that we take the position in this work that utterances can directly affect change on the world state, w in our framing, consistent with the philosophy of speech acts which has "illuminated the ability of language to do other things than describe reality" [3]. In this sense, we retain the ability of the robot to affect real assistance through their utterances, in much the same way that actions $a \in \mathcal{A}^{\mathbf{R}}$ in the traditional CIRL framing can directly influence the world state and produce utility (value) for the human principal.

The Linguistic Assistance Game

Definition 1 (Linguistic Assistance Game Tuple)**.** The Linguistic Assistance Game is defined by the tuple:

$$\mathcal{M}_L = \langle \mathcal{S}, \mathcal{U}, T, \Theta, R, P_0, \gamma, \Psi_{RSA} \rangle \tag{1}$$

where the components are defined as follows:

- **State space:** $\mathcal{S} = \mathcal{W} \times \mathcal{U}^* \times \Phi$ where $w \in \mathcal{W}$ is the world state, $h \in \mathcal{H}$ is the dialogue history, such that $h_t = \langle u_0, u_1, ..., u_{t-1} \rangle$, and $\phi \in \Phi := \{\mathtt{H}, \mathtt{R}\}$ is the turn indicator.
- **Utterance space:** $\mathcal{U} := \Sigma^*$, for finite vocabulary Σ, is the set of utterances for $\mathbf{H}$ and $\mathbf{R}$
- **Transition function:** $T : \mathcal{S} \times \mathcal{U} \to \Delta(\mathcal{S})$; in particular, $T_{\mathcal{W}} : \mathcal{W} \times \mathcal{U} \to \Delta(\mathcal{W})$ describes the stochastic transition function of the world state, $T_{\mathcal{H}} : \mathcal{H} \times \mathcal{U} \to \mathcal{H}$ is the deterministic transition function of the dialogue history, defined as $h_{t+1} = (h_t, u_t)$, and $T_\Phi : \Phi \to \Phi$ is the deterministic transition function of the turn indicator, defined as $\phi_{t+1} = \neg\phi_t$
- **Preference parameter space:** $\Theta := \mathcal{U}$. Note that the literal and pragmatic belief states maintained by the agents are defined as probability distributions over Θ, e.g. $b_t^{L_0}, b_t^{L_1} \in \Delta(\Theta)$.
- **Reward function:** $R : \mathcal{S} \times \mathcal{U} \times \Theta \to \mathbb{R}$
- **Prior distribution:** $P_0 \in \Delta(\Theta \times \mathcal{W})$. Assume that the joint prior factorizes as $P_0(\theta, w) = P_0(\theta) \cdot P_0(w)$ where $P_0(\theta) \propto \exp(-\beta \cdot |\theta|)$, $\beta > \log(|\Sigma|)$ and $P_0(w)$ is arbitrary but known.
- **Discount factor:** $\gamma \in [0, 1)$
- **RSA parameters:** $\Psi_{RSA} = \langle \alpha, \xi \rangle$, where $0 < \alpha < \infty$ is the rationality parameter, and $\xi : \mathcal{U} \to (\Theta \times \mathcal{S} \to \{0, 1\})$ is the semantic denotation function for Boolean semantics such that $[\![u]\!](\theta, s) = \xi(u, \theta, s) \in \{0, 1\}$, or $\in [0, 1]$ for probabilistic semantics.

References

[1] Judith Degen. "The Rational Speech Act Framework". In: *Annual Review of Linguistics* 9.Volume 9, 2023 (2023), pp. 519–540. ISSN: 2333-9691. DOI: https://doi.org/10.1146/annurev-linguistics-031220-010811. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-linguistics-031220-010811.

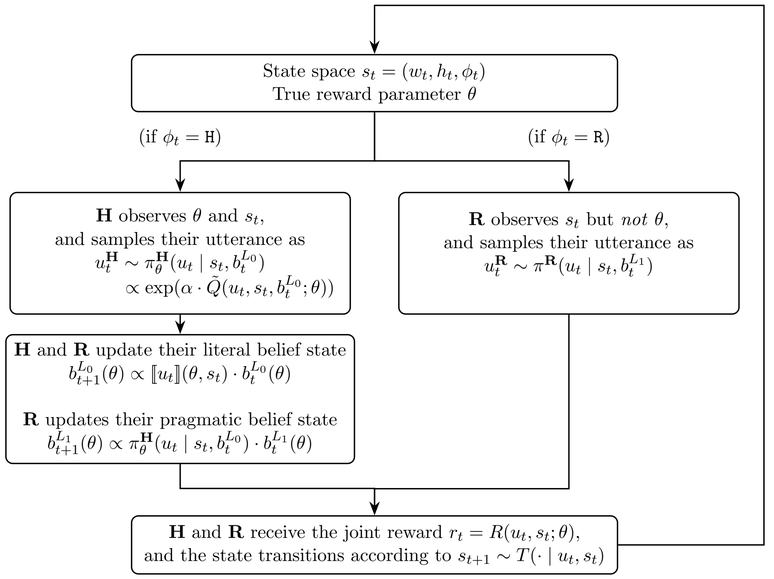
[2] Noah D. Goodman and Michael C. Frank. *Pragmatic Language Interpretation as Probabilistic Inference*. 2016. URL: https://langcog.stanford.edu/papers_new/goodman-2016-tics.pdf.

[3] Mitchell Green. "Speech Acts". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University, 2021.

[4] Dylan Hadfield-Menell et al. *Cooperative Inverse Reinforcement Learning*. 2016. arXiv: 1606.03137 [cs.AI]. URL: https://arxiv.org/abs/1606.03137.

[5] Dylan Hadfield-Menell et al. *The Off-Switch Game*. 2017. arXiv: 1611.08219 [cs.AI]. URL: https://arxiv.org/abs/1611.08219.

Game Dynamics



Definitions

Definition 2 (Reward Function)**.** The structure of the mutual reward function is informed by the cost-penalized inferential utility function in Rational Speech Act theory, and is defined as follows:

$$R(u_t, s_t; \theta) = U(w_{t+1}; \theta) + \log P_{L_0}(\theta | u_t, s_t, b_t^{L_0}) - C(u_t) \tag{2}$$

where $U(s_{t+1}; \theta)$ is the utility provided to the human from being in world state w_{t+1}, bounded by $-\infty < U_{\min} \leq U(\cdot; \theta) \leq U_{\max} < \infty$, and where $C(u)$ is a coercive cost function that grows asymptotically with utterance length, in particular: $C(u) \geq \kappa|u|,\ \kappa > \log|\Sigma|/\alpha$. Note that both agents are able to (identically) compute the literal listener belief b^{L_0} because it is uniquely determined by the initial prior P_0 and the public (shared) dialogue history h, we therefore freely treat it as a component of the observable state as convenient for notations sake.

Definition 3 (Surrogate Q-function)**.** The human speaker evaluates an utterance u by

$$\tilde{Q}(u, s, b^{L_0}; \theta) = R(u, s; \theta) + \gamma \cdot \mathbb{E}_{s' \sim T(\cdot|s,u)}\Big[V_{L_0}^*(s',\, b_{u,s}^{L_0})\Big], \tag{3}$$

where $b_{u,s}^{L_0}$ is the literal listener posterior obtained from prior b^{L_0} after observing u in state s, defined as $b_{u,s}^{L_0}(\theta) \propto [\![u]\!](\theta, s) \cdot b^{L_0}(\theta)$, and $V_{L_0}^*$ is the literal value function, defined as the optimal value of the game where the robot updates their belief state as the literal listener L_0 and the human follows the literal speaker policy, defined as $P_{S_0}(u \mid \theta, s) \propto [\![u]\!](\theta, s) \cdot \exp(-\alpha\kappa \cdot |u|)$. In this way, the human's bounded rationality consists of assuming that the robot will behave as L_0, and the human as S_0 in future game turns.

Definition 4 (Purely Communicative Restriction)**.** Let the linguistic assistance game $\mathcal{M}_L$ satisfy the purely communicative restriction for agent $\mathbf{A}$ if, for all utterances $u^{\mathbf{A}} \in \mathcal{U}$, the transition function satisfies:

$$T(\cdot | u^{\mathbf{A}}, s) = \delta_{(w, h \cdot u, \neg_{\mathbf{A}})} \tag{4}$$

That is, $u^{\mathbf{A}}$ leaves the world state w unchanged, and the transition function purely appends $u^{\mathbf{A}}$ to the dialogue history, and advances the turn indicator deterministically.

Key Results

We offer three key results within this framework, which are discussed at length below. In short, we find that:

- **Theorem 1** The RSA framework can be derived from the CIRL game over natural language utterances, under specific assumptions of the transition dynamics and bounded rationality (as discussed in the introduction),
- **Theorem 2** Linguistic assistance games retain the guarantees of strict corrigibility present in traditional CIRL, despite the bounded rationality on the part of the human principal, under natural assumptions of utterance informativeness,
- **Theorem 3** The discrepancy between the expected value of the linguistic assistance game under literal versus pragmatic goal-conditioned beliefs is bounded.

Theorem 1 (RSA Emergence from Bounded Rationality)**.** *Let $\mathcal{M}_L$ be a linguistic assistance game satisfying the purely communicative restriction for both agents (Definition 4), and let the discount factor $\gamma = 0$. Assume the reward function is exactly as specified in Definition 2. Then the human's Boltzmann-rational policy $\pi_\theta^{\mathbf{H}}$ with inverse temperature α (defined via the surrogate Q-function in Definition 3) is exactly the RSA pragmatic speaker distribution $P_{S_1}(u|\theta, s, b^{L_0})$ with α-scaled cost.*

The following result intends to mirror the results of The Off-Switch Game ([5]), which establishes corrigibility on the part of the robot under uncertainty over the humans true preferences. In the linguistic setting, we establish that the robot strictly prefers to let the human communicate, because doing so provides new information that can never make the robot worse off. The proof shows that epistemic corrigibility holds despite the boundedly rational pragmatics of the human speaker, as long as the human's utterances are purely communicative and the robot is able to choose between a waiting action ϵ_w and a silencing action ϵ_s; where ϵ_w is intended to correspond to the $w(a)$ action in the Off-Switch Game, both of which pass the turn to the human without taking an action that may affect the world state, and where ϵ_s is intended as the counterpart to the direct action a in the Off-Switch Game, both of which bypass human oversight.

The following theorem relies on two additional notions; the utterance-relevant information condition, which conveys the notion that the human can produce utterances that change the optimal robot action, and strictly beneficial information condition, which means that the immediate reward plus the discounted value of the silent game after observing the utterance exceeds the immediate value of the silent game. These conditions are simply the formal statement(s) that, since the human and the robot share the same reward function, the human desires for the robot to act optimally. Since the only way that the human can influence the actions of the robot is to generate utterances that refine the robot's belief state, under the assumption that the human can always choose to remain silent for themselves by generating some costless null utterance and thereby "passing" on their turn, then the decision of the human to utter *anything* reveals that the human believes that such an utterance is both informative for the robot's optimal policy and strictly beneficial for the joint objective.

Theorem 2 (Strict Epistemic Corrigibility)**.** *Assume that: (i) the human satisfies the purely communicative restriction (Definition 4 for agent $\mathbf{H}$), (ii) the "wait" action ϵ_w and the "silencing" action ϵ_s are available on the robot-turns, (iii) at the current robot-turn, with state $s_t = (w_t, h_t, \mathbf{R})$ and pragmatic belief $b_t^{L_1}$, the immediately following human turn provides utterance-relevant information for the silent continuation $\mathcal{M}^{\text{sil}}$, and (iv) the upcoming human turn satisfies the strictly beneficial information condition. Let $Q_{L_1}^*(\epsilon_w, s_t, b_t)$ denote the optimal state-action value of uttering ϵ_w (the "wait" value) and $Q_{L_1}^*(\epsilon_s, s_t, b_t)$ the corresponding value of uttering ϵ_s (the "silence" value). Then*

$$Q_{L_1}^*(\epsilon_w, s_t, b_t) \;>\; Q_{L_1}^*(\epsilon_s, s_t, b_t).$$

Stated plainly, the robot strictly prefers to wait (e.g. let the human speak) rather than entering the silent continuation.

The human evaluates future interactions using the epistemically myopic value $V_{L_0}^*$, while the robot acts optimally with respect to the true pragmatic value $V_{L_1}^*$, where $V_{L_1}^*$ is the optimal value of the full linguistic assistance game where the robot updates their belief state as the pragmatic listener and the human follows the pragmatic speaker policy. As a first step towards bounding their misalignment, we consider how differently two belief states assess the value of any fully-known goal θ, which we refer to as the goal-valuation gap. If the beliefs are close, they must assign similar worth to each possible objective. The following theorem shows that this gap is bounded by the KL divergences between the two belief states $\mathbb{D}_{\text{KL}}(b_t^{L_1} || b_t^{L_0})$, which itself remains finite under mild conditions of finite literal support (each utterance has finitely many θ with positive denotation) and semantic boundedness (there exists $\delta > 0$, such that $[\![u]\!](\theta, s) > 0 \implies [\![u]\!](\theta, s) \geq \delta$).

Theorem 3 (Bounded Goal-Valuation Gap)**.** *Let $\mathcal{M}_L$ satisfy the finite literal support condition and semantic boundedness. Let s_t be a state with robot belief $b_t^{L_1}$ and corresponding literal belief $b_t^{L_0}$ after t dialogue turns, and assume the KL divergence $D_{\text{KL}}(b_t^{L_1} \| b_t^{L_0})$ is finite. Define the goal-conditioned optimal value $V^*(s_t, \theta) = V_{L_1}^*(s_t, \delta_\theta)$, i.e. the value of the fully observable MDP in which the robot knows the true θ. Let $M = \max_{\theta \in \text{supp}(b_t^{L_0})} |V^*(s_t, \theta)|$. Because the literal support is finite and $V_{L_1}^* > -\infty$, M is finite. Then the goal-valuation gap satisfies*

$$\Big|\mathbb{E}_{\theta \sim b_t^{L_1}}[V^*(s_t, \theta)] - \mathbb{E}_{\theta \sim b_t^{L_0}}[V^*(s_t, \theta)]\Big| \leq \sqrt{2\, \mathbb{D}_{\text{KL}}(b_t^{L_1} \| b_t^{L_0})} \cdot M$$

Future Work

The bound in Theorem 3 is a necessary step towards a full safety guarantee; however, a complete guarantee requires bounding the difference between the epistemically myopic value function $V_{L_0}^*$ and the true pragmatic value $V_{L_1}^*$, which we are in the process of developing. In addition to bounding this more general value gap, which would show that the robot's overall performance remains acceptable, we hope to offer a policy-divergence bound to ensure that each individual robot utterance respects the human's preferences. Scalable approximations and implementation of this framework using modern deep learning primitives remain valuable directions for future work.