Reinforcement Learning in Werewolf Game AI: A Unified Framework for Multi-Agent Social Deduction

Arjun Agarwal (A0314489U), Hu Silan (A0304367E), Tanya Warrier (A0314513R), Wang Yuwen(A0211193R)

Abstract

This proposal presents a unified framework that leverages reinforcement learning (RL) for developing intelligent agents in "One Night Werewolf", a popular variant of Werewolf. In this version, the daytime phase comprises three rounds of sequential speeches, while the night phase features a single voting round with a reversed win-condition (votes by good-camp members yield a win for the werewolf camp and vice versa). We model the game as a Partially Observable Markov Decision Process (POMDP) and integrate AI planning techniques with deep RL (using PPO) to handle finite speech state design, belief updates, and hierarchical decision-making in a multi-agent setting.

1 Introduction

"One Night Werewolf" challenges players with uncertainty, deception, and strategic communication. Key modifications include:

- **Daytime Speeches:** Three rounds where players speak in sequence (ordered by player index), each round updating the speech history.
- **Night Voting:** A single round of voting with an inverted rule — if a vote comes from a good-camp agent, the werewolf camp wins, and vice versa.

Our framework aims to capture these unique game rules while employing advanced AI planning and RL methods to achieve robust multi-agent decision-making.

2 Methodology

2.1 Game Rules and Finite Speech State Design

We formalize the game as a POMDP:

$$\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O}, \gamma \rangle,$$

with global state defined as:

$$s = \langle \mathbf{R}, \mathbf{H}, \phi, \tau \rangle,$$

where $\mathbf{R}$ represents true player roles, $\mathbf{H}$ records historical actions (including night actions and daytime speeches), ϕ indicates the phase (Night, Day, Voting), and τ is the round number.

The finite *speech state* is structured as:

$$\mathcal{S}_{\text{speech}} = \langle \mathcal{S}_{\text{type}}, \mathcal{S}_{\text{target}}, \mathcal{S}_{\text{content}} \rangle.$$

During the daytime phase, three sequential rounds (in player order) update $\mathbf{H}_{\text{speech}}$ for subsequent belief updates.

2.2 Template-Based Speech Strategy

Given the finite nature of the speech state space, it is not necessary to employ a large language model (LLM) for generating speech. Instead, we can design a fixed set of templates that correspond to different speech types. This approach offers several advantages:

- **Simplified Decision Process:** By limiting the available speech options to a predefined set of templates, the policy learning problem is reduced to selecting among these discrete actions.
- **Consistency and Coherence:** Predefined templates ensure that the generated speech is both contextually relevant and strategically appropriate for the Werewolf game setting.
- **Reduced Computational Overhead:** Eliminating the need for an LLM decreases the complexity and computational resources required during both training and inference.

This template-based approach leverages domain knowledge, making it particularly well-suited for games with a restricted and well-defined communication space.

2.3 AI Planning and RL Integration

Our framework employs a deep RL algorithm (PPO) within an Actor-Critic architecture. Key components include:

- **Belief Update Module:** Incorporates sequential speech information to update each agent's belief B_i over player roles.
- **Hierarchical Action Selection:** Uses sub-policies for daytime speeches and night voting, explicitly accounting for the reversed win-condition.
- **Communication Strategy Module:** Generates structured speech actions within the finite speech state space.

The policy is parameterized as:

$$\pi_\theta(a \mid o, r, \tau),$$

and optimized to maximize:

$$\max_\theta \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \, \mathcal{R}(s_t, a_t, s_{t+1})\right],$$

with the reward function engineered to reflect the inverted outcomes during the night voting phase.

2.4 Technical Details and AI Planning Applications

We incorporate AI planning techniques such as:

- **Differentiable Belief Updates:** Using meta-learning to refine planning based on sequential communication.
- **PPO Optimization:** Robust learning in a multi-agent environment through self-play, experience replay, and curriculum learning.

These methods ensure that agents can plan, adapt, and execute strategies that integrate finite speech state updates with the critical decision-making required during the night voting phase.

References

[1] Xu, Zelai, Yu, Chao, Fang, Fei, Wang, Yu, and Wu, Yi. "Language agents with reinforcement learning for strategic play in the Werewolf game." In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, articleno. 2285, 2024.