AIWolf - OKAMI - ANAC 2021

Ran Wolf
ranwolf162@gmail.com
Bar-Illan University

Ziv Ben-David
benda1237@gmail.com
Bar-Illan University

June 2021

1 Introduction

Games are a great platform for Artificial Intelligence research. The social game Werewolf (aka Mafia) is a particularly difficult game for which no AI has yet been created that can play at a competitive level against humans and win consistently. The difficulties of this game include the social factor, the communication based game-play, and the necessity of modeling the internal belief state of all other agents. All of this in an asymmetric game with imperfect information.

The goal of this competition is to develop an AI agent for the Werewolf game, that can perform reliably well against other AI agents. Our agent and team name for the competition was OKAMI - "wolf" in Japanese.

This work is submitted as part of our studies in the Advanced AI course in Bar-Illan University[1]. It was submitted to the ANAC competition's Werewolf League division[2]. You can view the code for our agent in a github repository[3].

1.1 About the game

Werewolf, also known as Mafia, is a popular social deduction game, created by Dimitry Davidoff in 1986 [4]. In Israel this game is known as "Harotzeah" or "Haayara".

The game models a conflict between two groups: an informed minority (the werewolves), and an uninformed majority (the villagers). Each player is assigned a role affiliated with one of these teams.

Every player gets a character. In this competition, there are 15 characters: 1 Possessed, 3 Werewolves, 1 Seer, 1 Medium, 1 Bodyguard and 8 Villagers.

The game has two alternating phases: first, a night phase, during which those with night killing powers may covertly kill other players, and second, a day phase, in which surviving players debate the identities of players and vote to eliminate a suspect. The game continues until a faction achieves its win condition; for the village, this means eliminating the evil minority, while for the minority this usually means reaching numerical parity with the village.

2 Previous Work

The organizers of the competition published the best 15 agents in the competition that took place last year. We decided to use as a baseline an agent named HALU out of these 15 agents.

This agent can grade all agent's movements and judge and vote based on the score. The agent has a strategy to keep monitoring all agents, all behaviors and all utterances. For each player, the agent, depending on it's role in the game, calculates a score and determines who is the player to vote for (and eliminate).

While looking at the agent, we identified some gaps in its implementation. We found that the villager agent doesn't participate in the conversation at all. In addition, the seer agent, the possessed agent and all werewolves agents are coming out as seers right at the beginning of the game, but with this technique, the wolves team can immediately identify who is the true seer, and conversely, the villagers team can also identify the werewolves. Also, the power of the medium and of the bodyguard were not sufficiently utilized by the HALU implementation.

3 Our Improvements

In order to improve on HALU, we will be implementing a set of Heuristics for each role in OKAMI. since HALU has already taken care of much of the actual learning process and message handling code, our improvements will focus on just that: improving the ability of HALU to make valid judgements according to a set of extra rules. These heuristics were thought up by us, and a few of them make assumptions based on other agents from last year's agent strategies. In this section we will detail different Heuristics we will implement in the agent for each role he has.

3.1 Villager

The villager role is the base for all of the other roles. This means that Heuristics we plan for the villager will affect all of the agent's other roles.

- **Talking** - Currently there is no incentive for the Villager to talk at all until the voting stage. Since we want to improve our standing in regards to our fellow humans, if we are not a Werewolf role, we should like to be honest about our role - as a villager. Being truthful at the start of the game ensures that we are not a target for voting while humans still outnumber the werewolves. We assume that the werewolves will first target players with roles, and that the seer will announce himself in the first round else risk a reverse power-play by the werewolves. This allows us to pass information that we are OK to the rest of the villagers, enabling cooperation.

3.2 Seer

The seer can divine with 100% accuracy the role of one agent each round. This makes him a source of pure information, but also a target for werewolves should he reveal himself.

- **Hide your identity** - Since in our heuristics we assume that the seer is automatically the target of the werewolf faction, as seer we should simply gather information to reveal at a later date. We can also use our divination powers to try to lead the other agents towards voting out werewolves - not by the power of divination, but by the power of **suggestion**! Instead of saying that we divine a werewolf or human, we can say that we *think* that someone is a villager or werewolf. This should tip the information scales in favor of the humans. We are considering (but have not tested) the option that we say we are a villager in order to increase the credibility of our voting suggestions.

- **Denounce a werewolf** - On the other hand, if we do divine a werewolf, we will want to announce that we are seer and denounce the werewolf in order to "take him down with us". As villagers, if we do not successfully vote out werewolves each round, we will inevitably lose. That said, in each round we either kill a werewolf and a human dies at night, or two humans are killed. We want to create as many nights of the first variety. So taking down a werewolf with us is actually beneficial, as the humans get another round since the werewolves are less in number after a round where we successfully vote out a werewolf.

- **Use information to our advantage** - We can use the improved information to make requests about humans we divine. The most common request we will implement is a request for the bodyguard to protect a human we have divined. We will recommend each round the last human that we have divined since one of the humans is the *possessed* that registers as a human. we do not want to give him protection since the werewolves can't (and don't want to) kill him. Therefore, if we divine him, we will want to give him protection for only one round. The bodyguard does not have to agree to our request, but we will make it anyway as an attempt to generate credibility.

3.3 Medium

The second source of information in the game for the humans is the medium. The medium's power is to know the role of the voted out agent at the end of the round. Combined with the understanding that the werewolves kill one human each round during the "night", this allows the seer to know the exact balance of human and werewolf agents in the game at all times. This also makes the medium a great target for the werewolves to kill.

- **Suggest the killed agents roles** - As with the seer, we want to attempt a "honest" approach to the game. Instead of announcing that we are the medium and painting a target on our forehead, we will instead suggest the identity of the agents that were killed. This holds true *especially* if a werewolf was killed, as this is important information that will benifit other players.

- **Reveal your role at the critical point** - Only announce that you are the medium at the stage in the game where the werewolves are about to win. This is the point at which humans outnumber the werewolves by one or two people (the possessed is a human who wins with the werewolves!!). At this point, the only chance for success is for information to be as open as possible. This can be seen as a sort of "power play" where the medium and werewolves are the only agents in the game that can accurately discern the type of the agents at the table (human/werewolf).

3.4 Bodyguard

The bodyguard can "protect" one other player (not himself) from being killed by werewolves in a round. He has a different function in the game than the seer or the medium. Instead of attempting to gather information, his job is to prolong the game by guessing the target of the werewolves in the game and protect them. This causes the extension of the game since the werewolves can't kill their intended target this round.

- **No deaths == protected a human** - If no player dies during the night phase of the game, it is a sign that the bodyguard protected a player. The protected player, as the target of a werewolf attack, was a human! This is the only case of the bodyguard being able to actually gleam information from his own actions. As per our "honest recommendation" heuristic, we will tell the other players as a suggestion to assume that the player we protected is a human. In fact, since he was targeted, we will assume that he is safe to protect again in the following round.

- **Hidden identity** - We will not say our roles at all. The reason for this is that we do not want the werewolves selecting us more than someone else. Our job as bodyguard requires us to live for as long as possible. Even saying that we are a villager puts us at too much risk. We can only protect others, not ourselves.

3.5 Werewolf

The goal of a werewolf is to kill the villagers during the night phase, and cause the villagers to kill each other during the day phase. This means that it is advantageous to play with the information that the humans have in order to mislead them.

- **Be the seer** - Since until now we have discussed the power of being honest, the werewolf role is where we manipulate the trust. We will (at random) announce ourselves seer. Following this, we will begin "divining" humans as werewolves, and attempt to manipulate the human agents against one another. In this case, our goal is to kill the real seer as fast as possible so that they do not challenge us.

- **Be the medium** - As with the seer, we will at random announce ourselves as medium. Following this, we can manipulate the humans into thinking they killed a villager when they kill a werewolf and vice-versa. This should confuse the villager agents into making mistakes, and possibly mess up their inner models.

3.6 Possessed

As the possessed, your main goal is simply to survive. The werewolves will not kill you, and with proper play, the villagers will not vote you out. By simply surviving, you win. It is important to note that the possessed, like the medium, is aware of the current power balance between villagers and werewolves. That said, he should side with the werewolves and try to vote out villager agents.

- **Ask for protection** - Since we are being honest, let's say that we are a villager. Moreover, let's be the only person that can ask for the protection of the bodyguard without worry. Our goal with this tactic is to cause the bodyguard to mistakenly protect us instead of another legitimate target. Since we say we are human, this might be enough to convince the bodyguard to help us.

References

[1] S. Kraus, "Advanced ai." https://u.cs.biu.ac.il/~sarit/advai21.html, 2021.

[2] "Werewolf game league." http://aiwolf.org/, 2021.

[3] R. Wolf and Z. Ben-David, "Okami aiwolf agent code." https://github.com/ZivBenda/AIwolf-agent, 2021.

[4] F. Haffner, "Questions to dimitry davidoff about the creation of mafia on the french website." https://escaleajeux.fr/?principal=/jeu/mafid?, Jeuxsoc.fr. Retrieved 2011-04-11.