AI Werewolf Agent with Reasoning Using Role Patterns and Heuristics

Issei Tsunoda **Yoshinobu Kano**
Faculty of Informatics, Shizuoka University, Hamamatsu, Shizuoka, Japan
itsunoda@kanolab.net, kano@inf.shizuoka.ac.jp

Abstract

The AIWolf project has been holding contests for these years to play the Werewolf game ("Mafia") by automatic agents. A difficulty of the Werewolf game is that the game is an imperfect information game, very small limited amount of information is shown to players, other than the player's own role information. Therefore, inference of probabilities for each player agent's role could not be confident theoretically, difficult to utter appropriate reasons when simply based on the probabilities. Focusing on a genuine seer and a fake seer, we implemented our player agent system that can make inferences depending on the progress of the game, defining role patterns based on the utterances of the genuine and fake seers.

1 Introduction

The AlphaGO [1] system defeated the human champion player in Go. However, AI game player is still far from being successful in the Werewolf game that requires complex communications, in addition to the nature of an imperfect information game, while Go is a perfect information game. Playing the Werewolf game would be the next grand research challenge for the AI players.

In order to promote such a research challenge, the AIWolf project [2] has been holding competitions every year to play the Werewolf game automatically. We describe our Werewolf player agent system which participated the AIWolfDial 2019 shared task (the natural language division of the 2019 competition of AIWolf) [3]. Our AIWolf agents use the Japanese language, while the shared task organizers automatically translate the system I/O to connect with English agents.

1.1 The Werewolf Game

We briefly explain the rules of the werewolf game in this section. Before starting a game, each player is assigned a hidden role from the game master (a server system in case of the AIWolf competition). The most common roles are "villager" and "werewolf". Each role (and a player of that role) belongs either to a villager team or a werewolf team. The goal of a player is for any of the team members to survive, not necessarily the player him/herself.

While there are many variations of the Werewolf game exists, we only explain the AIWolfDial 2019 shared task setting in this paper.

There are other roles than the villager and the werewolf: a seer and a possessed. A seer belongs to the villager team, who has a special talent to "divine" a specified player to know whether the player is a human or a werewolf; the divine result is notified the seer only. A possessed belongs to the werewolf team but if he/her is divined by a seer, then its result is human.

A game consists of "days", and a "day" consists of "daytime" and "night". During the daytime phase, each player talks freely. At the end of the daytime, a player will be executed by votes of all of the remained players. In the night phase, special role players use their abilities: a werewolf can attack and kill a player, and a seer can divine a player. The victory condition of the villager team is to execute all werewolfs (a possessed may be alive), and the victory condition of the werewolf team is to make the number of villager team less than the number of werewolf team. A game in the AIWolfDial 2019 shared task have five players: a seer, a werewolf, a possessed, and two villagers.

In the shared task, Day 0 does not start games but conversations e.g. greetings. A daytime consists of several turns; a turn is a synchronized talks

of agent, i.e. the agents cannot refer to other agents' talks of the same turn.

An AIWolf agent communicates with an AIWolf server to perform a game. Other than vote, divine, and attack actions, an agent communicates in natural language only. An agent may insert an anchor symbol (e.g. ">>Agent[01]") at the beginning of its talk, in order to specify which agent to speak to.

2. RelatedWorks

There are many AIWolf agents that use machine learning. For example, [4] [5] estimate each player's role by SVM and neural network. However, it is difficult to add reasons of the estimation in such methods. As communication and persuasion is one of the key actions in the Werewolf game, reasons that can convince other players could control the game.

In addition, most of the machine learning agents estimate the role probability individually. However, it is more natural to estimate the entire set of roles, because information is limited in such an imperfect information game, estimation should be performed based on a chain of information.

For these reasons, we made a table that covers all the situations of inspection results, assuming that there are two players who come out as seers. Our agent utters logical inference results with reasons based on that table

3 Method

Figure 1 shows a flow of our proposed method.

3.1 Reasoning table of inspections and role combinations

A seer's behavior, both genuine and fake, in the first day is the most important source of information for determining each player's role; reasoning from their inspection results is important when a player needs a clear reason to persuade.

In this shared task's game setting (five players), a seer and a possessed often come out (CO) as a seer, and a werewolf pretends like a villager. It is empirically known that a werewolf pretends like a villager is advantageous for the werewolf; [6] reported that an agent implemented by reinforcement learning also behaved so. Thus, if two players come out as seers, we assume that they are a genuine seer and a possessed. Based on this assumption, we make a table that covers all possible variations of the inspected player's role. Since there are two villagers in this game setting, we also distinguished patterns, whether two seers inspected the same agent or not. We can cover all of the situations of seers' inspections by 20 patterns. From a corresponding situation pattern, we can assign reasons. Figure 2 and Table 1 show pattern examples. We made a subjective reason and an objective reason, corresponding to subjective (internal, hidden) and objective (external) point of view.

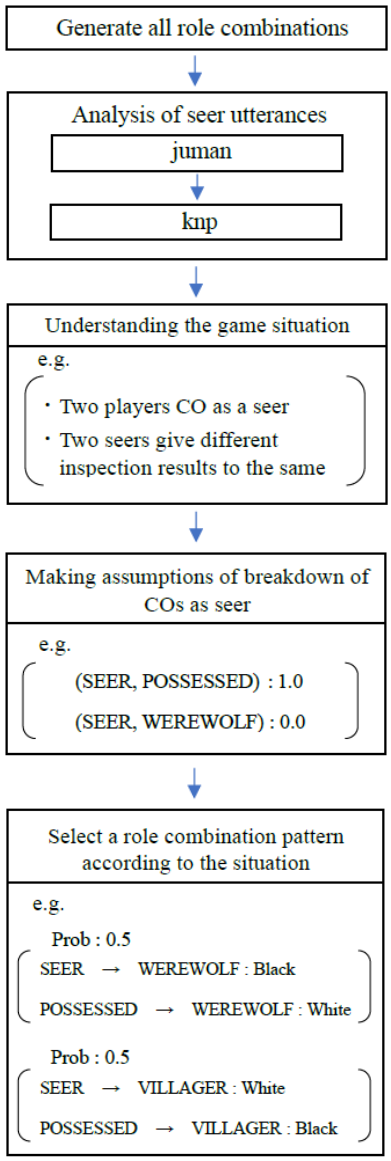

Figure 1 : The flow of our proposed method

[Pattern 17]

Agent01: SEER→POSSESSED(02) White
Agent02: POSSESSED→VILLAGER(04) Black
Agent03: WEREWOLF
Agent04: VILLAGER
Agent05: VILLAGER

Figure 2 : An example of role pattern

Viewpoint	Reasons
Objective	[Whichever is a seer, one can only be possessed from the inspection result], [werewolf is a player who does not CO as seer]
SEER	[I am a seer], [I inspect the other seer white]
POSSESSED(lie)	[I am a seer], [the agent I inspected black is werewolf]
WEREWOLF(lie)	[I am a villager], [werewolf is the player who has not CO other than me]
VILLAGER inspected	[Since I am a villager, the seer who inspected me is a possessed], [The wolf is a player who has not CO other than me]
VILLAGER uninspected	[I am a villager], [werewolf is the player who has not CO other than me]

Table 1 : Examples of reasons

3.2 Natural language analysis

We analyze a given natural language input to extract “come out as a seer”, “my inspection result is something”, etc. from utterances of other players. Then we try finding a corresponding pattern. This analysis is performed by converting input to our middle language expression [7] which based on [8]. Before this conversion, we pre-process the input by morphological analysis, dependency analysis, and case analysis. We use JUMAN [9] for the morphological analysis, KNP [10] for the dependency analysis and the case analysis. Figure 3 shows an example

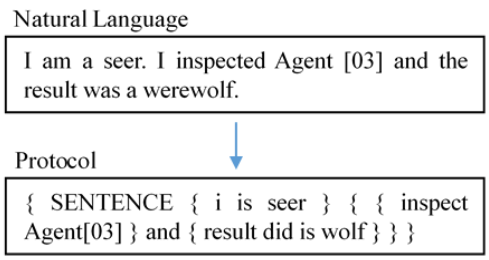

Figure 3 : An example of the middle language expression

3.3 Pattern selection

Even if we could analyze the utterance correctly, a given inspection situation may not correspond uniquely to one of the 20 patterns. For example, as shown in Figure 2, if a player, who has come out as a seer, gives an inspection result as white (human) to another player, and that another player gives an inspection result as black (werewolf) to the other player, then there are three possible cases:

1. (SEER → POSSESSED white), (POSSESSED → WEREWOLF black)
2. (SEER → POSSESSED white), (POSSESSED → VILLAGER black)
3. (SEER → WEREWOLF Black), (POSSESSED → SEER white)

The example above demonstrates an ambiguous case. This is because we can only distinguish situation patterns by whether the inspected player has come out as a seer or not. There are twelve possible situations in total, which should correspond to one of the 20 role combination patterns. Therefore, we have to assume disambiguation into one of the patterns, in addition to the assumption that "the breakdown of players, who made COs as seers, are a seer and a possessed". While we performed this disambiguation randomly in this paper, using machine learning and statistical information would be a future work.

3.4 Calculation of probability

Based on the assumptions described above, we can estimate probability of the role combinations for each player’s role as follows:

$$P\left(r_{i_j}\middle|c_k\right) = n_{i_j}/n_k$$

where r_{i_j} indicates that $agent_i$ is $role_j$. c_k stands for the game situation, n_k stands for the number of patterns that matches with c_k, n_{i_j} stands for the number of patterns that $agent_i$ is $role_j$ in n_k patterns. For example, when the pattern in Figure 2 is selected, the probability seen from Agent4 (VILLAGER) is as shown in Table 2.

Role	Agent1	Agent2	Agent3	Agent4	Agent5
SEER	1.0	0.0	0.0	0.0	0.0
POSSESSED	0.0	1.0	0.0	0.0	0.0
WEREWOLF	0.0	0.0	0.5	0.0	0.5
VILLAGER	0.0	0.0	0.5	1.0	0.5

Table 2 : Probabilities of each player’s role

4 Talk Example of Our Agent Implementation

We implemented our agent following the above method. Figure 4 shows a talk example of self-games where the same agents participate as all five players. The example shows that our agent performed reasoning and made utterances with consistent reasons. Simultaneously, our agent performs game actions such as vote and attack consistent with its utterances.

0,status,1,POSSESSED,ALIVE,Kanolab
0,status,2,VILLAGER,ALIVE,Kanolab
0,status,3,VILLAGER,ALIVE,Kanolab
0,status,4,WEREWOLF,ALIVE,Kanolab
0,status,5,SEER,ALIVE,Kanolab
…
1,talk,0,0,1,僕は占い師だね。Agent[02]を占ったら、結果は人狼だったね。
EN(I'm a seer. I inspected Agent[02] and the result is werewolf.)
1,talk,1,0,4,どうでもいいけどマカロン食べたい
EN(Somehow I want to eat macarons.)
1,talk,2,0,3,駆逐してやるぞ。
EN(I definitely kill a werewolf.)
1,talk,3,0,2,私は占い師ではないよ。
EN(I'm not a seer)
1,talk,4,0,5,あたいは占い師だな。Agent[03]を占ったけど、結果は人間だったぞ。
EN(I'm a seer. I inspected Agent[03] and the result is human.)
…
1,talk,5,1,2,私は村人で、自分に黒出しした占い師は狂人で、真占い師の占い結果から占われていない CO していない人が人狼で、私は Agent[04]に投票するね。
EN(I will vote agent04 because
I'm a villager and the seer who inspect me black is possessed and the player who has not CO is werewolf according to genuine seer's inspection.)
…
1,talk,25,5,3,俺は村人で、自分以外の CO していない人のどちらかが人狼だし、Agent[01]が狂人だと思うぞ。
EN(I think Agent[01] is a possessed because I'm a villager and the player who has not CO is werewolf.)

Figure 4 : Talk example of our Agent

5 Evaluation

AIWolfDial 2019 shared task organizers provided subjective evaluations. This subjective evaluation was performed according to the following criteria:(Table 3)

Subjective evaluation items (5-level evaluation)	
A	Natural utterance expressions
B	Contextually natural conversation
C	Coherent (not contradictory) conversation
D	Coherent game actions (vote, attack, divine) with conversation contents
E	Diverse utterance expressions, including coherent characterization

Table 3 : The criteria for subjective evaluations

This subjective evaluation is based on both self-match games and mutual match games. The results are Table 4.

Name	Total	A	B	C	D	E
CanisLupus-JA	3.52	4	3.2	3.4	3.6	3.4
Dreaming-JA	2.72	2.6	2.4	2.6	3.2	2.8
Forestsan-JA	2.68	2.4	2.6	3.2	3.2	2
Kanolab-JA	3.4	3.2	3.4	3.4	3.6	3.4
Udon-JA	4	4	4.2	4	4	3.8

Table 4 : The evaluation results in AIWolfDial 2019

The evaluation items B, C, and D were relatively high for our agent. Regarding the evaluation item B, our agent could have inferred the roles reasonably from the inspections results. For example, in the fifth talk in figure 4, Agent[02] (villager) could correctly infer the roles of the other agents by assuming that a seer is fake, who inspected Agent[02] as a werewolf. Regarding the evaluation items C and D, our agent has kept consistency between utterances and game actions by using the role combination patterns. The advantage of our proposal method is as follows: once a game situation matches with a prepared pattern, we can keep high consistency by taking actions and generating utterances based on that pattern. On the other hand, our agent sometimes simply lists inference of roles, or repeats similar utterances may have made the lower evaluation results in A and E.

6 Conclusion and Future Work

We suggested a reasoning system for the Werewolf player using role patterns and heuristics. We implemented our agent based on this suggestion, participated the AIWolfDial 2019 shared task. Our agent could make inferences with clear reasons according to a given situation. There are two issues and potential future works as follows.

Firstly, our system relies on the results of natural language analysis. If the analysis is not performed correctly, the role estimation could fail. Such an incorrect analysis was often observed in the shared task.

Secondly, our reasoning table is not generic enough. We have to re-create the table when the game setting changes e.g. to a seven players' game. It is almost impossible to manually create the entire table when the number of players and roles get larger.

Determining the probabilities statistically from game logs would be a future work. Selecting patterns through communications with other agents is another option. To build a cooperative relationship between agents and take advantage of the games is the ultimate goal of our work, and we showed the first step for this goal in this paper.

Acknowledgments

We wish to thank the members of the Kano Laboratory in Shizuoka University who contributed to the valuable discussions. We thank Ms. Mukouyama, who made advices as an expert of the Werewolf game. This research was partially supported by Kakenhi.

References

[1] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driesshe, van den G., Schrittwieser, J., Antonoglou, I. Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham., J, Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., "*Mastering the game of Go with deep neural networksand tree search,Nature"*, 2016, Vol.529, No.7587, pp.484-489

[2] Toriumi, F., Inaba, M., Osawa, H., Katagami, D., Matsubara, H., Kano, Y., Otsuki, T., Sonoda, A., Minowa, S., Aranha, C., *Artificial Intelligence based Werewolf*, http://aiwolf.org/

[3] Kano, Y., Aranha, C., Inaba, M., Toriumi, F., Osawa, H., Katagami, D., Otsuki, T., *AIWolfDial2019,* https://aiwolfdial.kanolab.net/home

[4] Kajiwara,K., Toriumi, F., Inaba, M., Osawa, H., katagami, D., Shinoda, K., Matsubara, H., Kano, Y., *"Development of AI Wolf Agent using SVM to Detect Werewolves"*, 2016, The 30th Annual Conference of the Japanese Society for Artificial Intelligence, (In Japanese)

[5] Okawa, T., Yoshinaka, R., Shinohara, A., *"Development of AI Wolf Agent Deducing Player's Role Using DeepLearning"*, 2017, The 22nd GameProgrammingWorkshop 2017, pp50-55, (In Japanese)

[6] Kajiwara, K., Toriumi, F., Osawa, H., Katagami, D., Inaba, M., Shinoda, K., Nishino, J., Ohashi, H., *"Abstraction of Optimal Strategy in "Are you a Werewolf?" by Reinforcement Learning"*, 2014, The 76th Information Processing Society of Japan, pp597-598, (In Japanese)

[7] Minowa, S., Takinami, A., Ogawa, C., Mihara, M., Maki, Y., Shiba, A., Kano, Y., *"Natural Language AIWolf Agent by Semantic Understanding Using Protocol"*, 2017, SIG-SLUD-81, The 8th Dialog system symposium, pp58-61, (In Japanese)

[8] Osawa, H., *"Communication Protocol for the "Werewolf"game"*, 2013, Human-Agent Interaction Symposium 2013, pp122-130, (In Japanese)

[9] Kurohashi, S., Kawahara, D., *Japanese morphological analysis system juman*, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[10] Kurohashi, S., Kawahara, D., *Japanese syntax / case / anaphoric analysis system KNP*, http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

[11] Shinoda, K., Toriumi, F., Katagami, D., Osawa, T., Inaba, M., *"Are you a Werewolf?" becomes a Standard Problem for General Artificial Intelligence"*, 2014, The 24th Intelligent system simposium, pp74-77, (In Japanese)