Strategies for an Autonomous Agent Playing the "Werewolf game" as a Stealth Werewolf

Shoji Nagayama[1] **Jotaro Abe**[1] **Kosuke Oya**[1] **Kotaro Sakamoto**[1]

Hideyuki Shibuki[2] **Tatsunori Mori**[1] **Noriko Kando**[2]

[1]Yokohama National University, Japan
[2]National Institute of Informatics, Japan
{nagayama, jotaro, kosuke-o, sakamoto, mori}@forest.eis.ynu.ac.jp
{shib, kando}@nii.ac.jp

Abstract

The "Werewolf game" is a popular multiplayer game wherein "villagers" try to figure out who is a "werewolf" through conversations. Werewolves usually pretend to be villagers. In this paper, we studied conversations in game logs in order to investigate how werewolves' cooperation contributed to increasing the winning percentage of the werewolves' team. As the number of "whispers" that are utterances via werewolves' private chat may be regarded as a measure of the werewolves' cooperation, we investigated the relation between the number of whispers and the winning percentage. As the result, we observed that the winning percentage of werewolves' team increased by 63 points at most when the number of whispers of at least two werewolves was more than 106.

1 Introduction

The "Werewolf game" is a popular multiplayer game wherein "villagers" try to figure out who is a "werewolf" through conversations. Werewolf game is actively researched, and competitions are also held in as shared task (Kano et al., 2019). As conversations in the game are open for all players, no player can talk with other players in secret. Therefore, working together with only his/her allies through conversations is difficult. Consequently, each player's thought and action become complicated. On the contrary, werewolves have a special talk channel, Whisper, through which they can secretly talk with other werewolves, allowing werewolves to work together. This is a strong advantage for werewolves, and it is an important factor so that werewolves win.

There are two basic strategies for werewolves. The first strategy is called as "swindle werewolf", wherein a werewolf makes himself/herself seem to be a leader of villagers, such as a seer or a medium. The second strategy is called as "stealth werewolf" wherein a werewolf hides himself/herself as one of the villagers. The swindle werewolf can have the initiative for misleading villagers, while it is easy to be a target of divination or execution. The stealth werewolf cannot have the initiative, but it is hard to raise a doubt of werewolf since he/she does not work directly on the subject of execution. We attempt to make the stealth werewolf an agent, and would like to clarify important factors for the stealth werewolf. If an agent can talk and mislead villagers without attracting attention from other players, it is a strong stealth werewolf. Although there are previous studies that have investigated conversations in the Werewolf game (Hirata et al., 2016), they are not done so from the standpoint of the stealth werewolf. Therefore, in this paper, we investigate the influences of the numbers of utterances, appearances in utterances of other players, and whispers, on the victory or defeat of werewolves.

2 Related work

There are the following previous researches about the Werewolf game. Toriumi et al., (2017) described the advantage of using the Werewolf game as "including the asymmetric diversity of player information, persuasion as a means of earning confidence, and speculation to detect fabrication." Gillespie et al.,(2016) used transcripts of the Werewolf game as the evaluation data of their semantic classifier. Takahashi et al.,(2017) measured trust between players through the arranged Werewolf game. Wang et al.,(2018) built a robot that had abilities such as casting a glance to play the real world Werewolf game. Xiong et al.,(2017) reported the optimal number of players to convey the attraction of the Werewolf game. The above researches did not aim to make agents in the Were-

Table 1: Number of data and players

players	files	the number of role						
		villager	seer	guard	medium	werewolf	possess	NPC
14	89	6	1	1	1	3	1	1
15	33	7	1	1	1	3	1	1
16	309	8	1	1	1	3	1	1
sum	431							

wolf game. Nide et al.,(2017) attempted to make a agent using extended BDI model, and conducted a thought experiment. However, no empirical experiment was conduct. Nakamura et al.,(2016) reported that estimating player roles based multiple perspectives increased winning rate. Hirata et al.,(2016) made an agent using action probabilities based on game logs of werewolf BBS for behaving like human beings. Their algorithms are not specialized in werewolf agent. We aim to construct a strategy for the stealth werewolf.

3 Werewolf BBS

Werewolf BBS is a bulletin board system for the online Werewolf game. A game session is called as "a village", which comprises 10 to 16 characters, including a non-player character (NPC)[1]. The game time synchronized with the real -world time, and it takes approximately a week to play a game. Each player can have up to 20 utterances per day. The non-verbal communication information is not allowed. As werewolves have a special talk channel, Whisper, they can discuss their strategy, for example, as to who takes charge of the swindle werewolf or the stealth werewolf.

In this study, we collected game logs from "Werewolf BBS: G villages" for analysis using Python library, Beautiful Soup[2]. We collected villages that included three werewolves, indicating that the number of villagers is 13 to 16, and of which players did not drop out, except execution or attack[3]. A village was collected as a file. Table 1 lists the number of collected files and the number of game roles in each village. There were no villages with 13 players that met the collection condition described above, and the total number of the files was 431 (243 MB). Although every file includes a prologue involving idle talk before the game roles are assigned to players, the prologue was excluded for analysis. The average number of utterances per file after excluding a prologue was 70.7.

4 Influence of utterances and appearances

Strong stealth werewolf talks into misleading villagers without attracting attention from other players. As judging whether an utterance can lead to misleading is difficult, we used the following two measures for attracting appearances.

The first measure is the number of utterances indicating how many times a player talks, because we considered that players with a lot of utterances were conspicuous. The second measure is the number of appearances indicating how many times a player comes up in utterances of other players, because we considered that it indicates how he/she attracts attention from other players. The more the number of utterances and appearances are, the more attention will be drawn.

Using decision trees, we analyzed how the numbers of utterances and appearances per player affected the winning percentage of werewolves. For making a decision tree of utterances, we used 16 character roles as attributes, the total numbers of utterances in a game as attribute values, and victory or defeat of werewolves as classes. The decision tree of appearance was made in the same manner. If a role such as werewolf or villager was assigned to two or more players, it was distinguished by the rank in the descending order of the number of utterances or appearances. If the number of players in a game was fewer than 16, we add dummy villagers to make up for the shortage. The utterance number of dummy villagers and the appearance number of those are assumed to be zero. We used the Python libraries scikit-learn[4] and dtreeviz [5] for making and showing decision trees, respectively. Figures 1 and 2 show the decision trees of utterances and appearances, respectively. Bifurcation occurs depending on a certain threshold for the number of utterances and ap-

[1]The number of actual players is 9 to 15.

[2]https://github.com/waylan/beautifulsoup

[3]Players who does not talk at least once a day are forcibly dropped out. Besides, players can stop playing the game of their own accord.

[4]https://github.com/scikit-learn/scikit-learn

[5]https: // github.com/parrt/dtreeviz

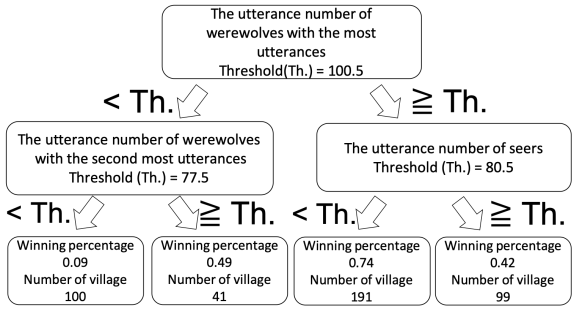

Figure 1: Decision tree based on the number of utterances

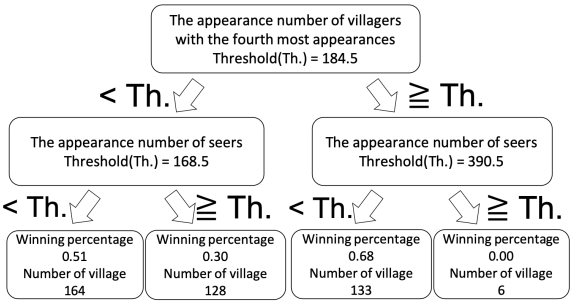

Figure 2: Decision tree based on the number of appearances

pearances. This corresponds to the case where the value of the right branch is greater than or equal to the threshold, and the case where the value of the left branch is less than the threshold. In the figures, "Winning percentage" represents the winning percentage of werewolves, and "Number of villages" represents the number of applicable villages. Each of leaves displays the winning percentage of werewolves and the number of applicable villages at the condition. From Figure 1, if the utterance number of the werewolf with the most utterances is 100.5 or more and the utterance number of seers is less than 80.5, the winning percentage of the werewolf teams is as high as 74 points across 191 villages.

Looking at Figure 1, it can be confirmed that the first branch is made by the utterance number of werewolves; thus, the victory or defeat branches depending on the utterance number of werewolves. There is also the utterance number of werewolves at the second branch, and if the utterance number of werewolves is less than a certain number, the winning percentage of the werewolves reduces. The winning percentage of werewolves at this time is at least 9 percent. If the utterance number of werewolves with the most utterance is more than the threshold and the number of utterance of the seer is fewer than the threshold, the winning percentage was increased from 9 to 74 points.

From Figure 2, the appearance number of villager with the fourth most appearances is the first branch, and the appearance number of seers is the second branch. It seems to be the subject of conversation, whether the particular villager is suspected of being a werewolf or whether the seer is real. Especially, it can be confirmed that if the appearance number of seer is low, the winning percentage of werewolves increases. However, the number of appearances of werewolves does not appear as a factor of victory or defeat, and it is difficult to reflect this knowledge on the specific tactics of werewolves.

Based on the above, we consider a method of manipulating the number of utterances of a specific player, centering on the utterance number of werewolves involved in victory or defeat. Especially when the werewolf manipulates the number of utterances of a specific player, it is necessary to cooperate well with other werewolves so that the operation does not suffer. The details of such a werewolf collaboration are explained in Section 5. The number of appearances is not as good as expected, and the effect of the number of appearances of werewolves on victory or defeat is not so great. In particular, when reducing the number of appearances, unlike the number of utterances, it cannot be controlled even if it is excluded from the game by execution or attack. For this reason, we will consider methods for estimating roles in which werewolves utter so as not to raise, the number of their appearance in utterance of other players.

5 Influence of whispers

One of the unique abilities of a werewolf that cannot be found in other roles is the "whisper" described in Section 3. Using "whisper" makes it possible for the werewolves to cooperate secretly, which can have a big influence on victory. For example, as described in Section 4, manipulating the utterance number of a specific player as a method called "asking" that affects victory or defeat is possible. "Asking" increases the utterance number of a specific player intentionally by seeking a response by speaking to a specific player. However,

as the number of utterances that a player can make per day is limited, controlling the number of utterances of all players alone is difficult. If our role is that of a werewolf, we may ask another werewolf who has sufficient room of utterances to use "asking" through whispers. The utterance number of specific players can be controlled such that the werewolf teams is advantageous. We investigated how the number of werewolves' whispers affects victory or defeat by making a decision tree. The decision tree for whispers was made in the same manner as that mentioned in Section 4. For making a decision tree of whispers, we used three werewolf players as attributes, the total numbers of whispers in a game as attribute values, and the victory or defeat of werewolves as classes.

In Figure 3, increasing the number of whispers does not simply means that the werewolves are cooperated well. For example, when a werewolf asks questions or proposes a strategy, another werewolf will not always get on his proposal. To get his proposal accepted, persuading through dialogue is necessary, which is the essence of the Werewolf game. In a dialogue, a response from another werewolf maybe expected for the utterance of a werewolf. If there is not much difference in the number of whispers of each werewolf, we may infer that the dialogue has been established. Therefore, we assume that the number of whispers among the werewolves is considerable, and the strategy and situation are well discussed and coordinated, if there is no difference in the whisper number of each werewolf.

From Figure 3, the first branch shows the winning percentage of werewolves is higher when the number of whispers is larger throughout the game. In the second right branch, the value of threshold is the number of whispers posted by the second most whispering werewolf. That means the winning percentage of werewolves is higher when two werewolves establish the dialogues frequently. Specifically, the winning percentage of werewolves is high at 67 percent when both the first branch and the second branch are above the threshold. The winning percentage increases by 63 points compared to the case where the first branch and the second branch are both below the threshold. However, this analysis does not evaluate the difference in the number of whispers from the viewpoint of the degree of cooperation between the werewolves, owing to which another

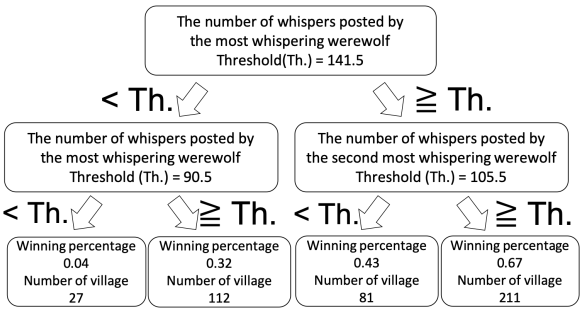

Figure 3: Decision tree based on the number of whispers

process is required. In addition, as the total of file size is only 40 MB, the number of villages used for analysis must be increased. However, when including villages where there are not more than three werewolves in the analysis, normalization is required because the number of roles and adopted roles set does not correspond to our current data set. In other words, investigating normalization conditions that do not depend on the number of werewolves and roles set is the immediate challenge.

6 Conclusion

In this paper, we analyzed 431 game logs of Werewolf BBS focusing on the "stealth werewolf", and confirmed that the winning percentage of werewolves increased by 65 points at most when the number of werewolf utterances was very frequent. We also confirmed that the winning percentage of werewolves increased by 63 points at most when the number of whispers was very frequent. In the future, we intend to proceed with research considering the content of utterances and whispers.

Acknowledgments

The research funded for / supported by the open collaborative research program at National Institute of Informatics (NII) Japan (FY2018).

References

Kellen Gillespie, Michael W. Floyd, Matthew Molineaux, Swaroop S. Vattam, and David W. Aha. 2016. Semantic Classification of Utterances in a Language-Driven Game. In *Communications in Computer and Information Science*, pages 116–129.

Yuya Hirata, Michimasa Inaba, Kenichi Takahashi, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Kosuke Shinoda. 2016. Werewolf Game Modeling using Action Probabilities based on Play Log Analysis. In *9th International Conference on Computers and Games*, pages 103–114.

Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Hirotaka Osawa, Daisuke Katagami, Takashi Otsuki, and Fujio Toriumi. 2019. Overview of the AIWolf-Dial 2019 Shared Task: Competition to Automatically Play the Conversation Game "Mafia". In *In proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial 2019), the 12th International Conference on Natural Language Generation (INLG 2019)*.

Noritsugu Nakamura, Michimasa Inaba, Kenichi Takahashi, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Kosuke Shinoda. 2016. Constructing a Human-like agent for the Werewolf Game using a psychological model based multiple perspectives. In *2016 IEEE Symposium Series on Computational Intelligence*. IEEE.

Naoyuki Nide and Shiro Takata. 2017. Tracing Werewolf Game by Using Extended BDI Model. In *Special Section on Frontiers in Agent-based Technology*, pages 2888–2896.

Hideyuki Takahashi, Midori Ban, Hirotaka Osawa, Junya Nakanishi, Hidenobu Sumioka, and Hiroshi Ishiguro. 2017. Huggable Communication Medium Maintains Level of Trust during Conversation Game. In *Frontiers in psychology*.

Fujio Toriumi, Hirotaka Osawa, Michimasa Inaba, Daisuke Katagami, Kosuke Shinoda, and Hitoshi Matsubara. 2017. AI Wolf Contest -Development of Game AI Using Collective Intelligence-. In *Communications in Computer and Information Science*, pages 101–115.

Bohao Wang, Hirotaka Osawa, Takuya Toyono, Fujio Toriumi, and Daisuke Katagami. 2018. Development of Real-World Agent System For Werewolf Game. In *17th International Conference on Autonomous Agents and Multiagent Systems*, pages 1838–1840.

Shuo Xiong, Wenlin Li, Xinting Mao, and Hiroyuki Iida. 2017. Mafia Game Setting Research using Game Refinement Measurement. In *14th International Conference Advances in Computer Entertainment Technology*, pages 830–846.