Are Talkative AI Agents More Likely to Win the Werewolf Game?

Dolça Tellols
Tokyo Institute of Technology
tellols.d.aa@m.titech.ac.jp

Abstract

The Werewolf game is a communication game where, usually, two teams compete against each other. As players discuss and share ideas during the game to define their strategy, being talkative or not is one of the characteristics that define them. This paper presents a data analysis over logs from the shared task of The 1st International Workshop of AI Werewolf and Dialogue System to discuss if being talkative or not can be related to winning or losing when AI agents play the Werewolf game. Overall results show that the difference in the average of utterances sent by winning and losing players is not significant. However, they also suggest further analysis and discussion.

1 Introduction

In recent years, there have been approaches to implement Artificial Intelligence (AI) agents capable of playing and competing with other agents or human players in a variety of games. Proposals for games that do not require social interaction between the players, like chess, shogi and go, are achieving promising results (Silver et al., 2016, 2018). However, developing AI agents for communication games like the Werewolf game, where usually two teams (villagers and werewolves) discuss and compete against each other, remains a challenge.

The Artificial Intelligence based Werewolf project[1] is contributing to the previous aspect by researching and providing platforms to develop and test AI Werewolf agents. Researchers can implement agents to play in the protocol division (agents communicate in a middle language called the AI Werewolf protocol) or the natural language division (agents communicate using natural language utterances). The 1st International Workshop of AI Werewolf and Dialog System[2], taking place in the context of the International Natural Language Generation 2019 conference, proposed a shared task where participants implemented AI agents capable of playing the Werewolf game using natural language utterances (Kano et al., 2019).

In the frame of the shared task and based on the idea that being talkative or not may characterize the strategy that some players of the Werewolf game follow, this paper focuses on analyzing if the talkativeness level of the participant AI agents may be related to winning or losing the Werewolf game. To do so, and having as reference a talkative agent implemented to participate in the previously cited shared task, this work analyzes, in Section 4, the logs from the played games in which the agent participated and discusses the results in Section 5.

2 Related Work

In the last few years, research based on the Werewolf game is increasing.

On the one hand, a lot of researchers use the game to analyze human players' behaviors. For example, some created the Idiap Wolf Database and used it to show how it is possible to automatically detect suspicious actions and how the degree of speaker behavior influences on the outcomes of the game (Hung and Chittaranjan, 2010; Chittaranjan and Hung, 2010). Other researchers used machine learning to analyze video data of people playing the game and checked the importance that nonverbal information has to achieve victory (Katagami et al., 2014).

On the other hand, because of proposals like

[1] http://aiwolf.org/en/

[2] https://aiwolfdial.kanolab.net/home

the already cited AI based Werewolf project, the number of works defining implementation strategies for AI Werewolf agents is increasing. As an example, some researchers proposed psychological models to be used to implement AI Werewolf agents so that they achieve higher winning rates (Nakamura et al., 2016). And others developed a behavioral model for the implementation of agents based on logs from human players (Hirata et al., 2016).

This work contributes by providing a data analysis study that opens a discussion over how the talkativeness of an AI agent may be related or not to its winning rate. Results may serve as a reference for the future development of AI agents that can play the Werewolf game and communicate using natural language.

3 Dataset

As data, this paper uses a set of 60 game logs from the shared task of The 1st International Workshop of AI Werewolf and Dialogue System (AIWolfDial2019) from the 2019 International Natural Language Generation (INLG2019) conference (Kano et al., 2019).

In all games, the same five agents (A1, A2, A3, A4, and A5), play the werewolf game using natural language utterances written in Japanese. The talkative agent implemented for the shared task (A4) sends utterances as long as they do not become excessively repetitive (slight repetition may result in emphasis). All five agents participate in all games and, each time, they have randomly assigned one of the following roles: villager (has no special skill and there are two in each game), seer (can see if a player is human or werewolf at the end of each day), possessed (human from the werewolves side) or werewolf (can eliminate one player at the end of each day from day 1). Agents play each role 12 times (seer, possessed and werewolf cases) or 24 (villager case).

The shared task allows agents to communicate freely using natural language during certain periods ("days"), without specifying a maximum number of utterances per day. Since day 0 only consists of greetings, Section 4 analyzes utterances performed from day 1. Because of the small number of players, each game only lasts for one day (20 games) or two (40 games). This is because, each day from day 1, all alive players vote to eliminate a player (villagers try to use this voting to eliminate the werewolf) before the werewolf eliminates another one.

Logs contain the following information: (i) status (keeps playing or not) and role of each player at the beginning of each day and the end of the game; (ii) utterances each player performs during the day; (iii) information of the seer divination at the end of each day; (iv) voting each player performs at the end of each day (excluding day 0) and the corresponding result (which agent stops playing); (v) information of the werewolf attack at the end of each day (excluding day 0); and (vi) result of the game indicating the status and role of each player and the winning side (villagers or werewolves).

4 Data Analysis

To discuss in Section 5 if the talkativeness of AI agents affects their odds of winning the Werewolf game, this work analyzed the data presented in Section 3 from different points of view: (i) game result; (ii) side; (iii) role; and (iv) agent. For each case, this paper presents the average (Avg.) and standard deviation (Std. Dev.) of the number of utterances sent during one day by a player belonging to one of the categories each point of view may consider. When appropriate, it also presents data depending on the game result (win or lose) and shows the winning rate.

Agent	Utterances	
Game Result	Avg.	Std. Dev.
Win	8.9	1.05
Lose	9.08	0.88
All	8.99	0.97

Table 1: Analysis results of the utterances sent per agent and day according to the game result.

Table 1 shows an overview result by comparing the average of utterances sent each day by winning players and by losing players. Since the average number of utterances sent by winning players (8.9) is similar to the one of losing players (9.08), it seems that talkativeness may not be a determining factor that leads to deciding the game result. Because the difference between the winning and losing agents' data samples follows a normal distribution, this study also performed a t-test, which confirms that the difference in the presented averages is not significant (p-value = $0.3214 > 0.05$).

Agent Side	Result	Utterances Avg.	Utterances Std. Dev.	Win. Rate
Villagers	Win	9.22	0.68	0.62
	Lose	8.91	0.6	
	All	9.1	0.67	
Werewolves	Win	8.38	1.3	0.38
	Lose	9.18	1.01	
	All	8.88	1.19	

Table 2: Analysis results of the utterances sent per agent and day according to the side of the agent.

Table 2 shows the result of the analysis performed according to the side (villagers or werewolves) each agent belongs to in a game. On the one hand, winning players from the villagers' side perform more utterances (9.22) than losing players (8.91) and their winning rate is 0.62. On the other hand, losing players from the werewolves' side perform more utterances (9.18) than winning players (8.38) and their winning rate is 0.38. There is almost no difference between the average number of utterances performed by villagers' side players (9.1) and the average number of utterances performed by the werewolves' side players (8.88).

Agent Role	Result	Utterances Avg.	Utterances Std. Dev.	Win. Rate
Villager	Win	9.26	0.79	0.62
	Lose	8.96	0.93	
	All	9.14	0.86	
Seer	Win	9.15	1.17	0.62
	Lose	8.8	1.15	
	All	9.02	1.18	
Possessed	Win	8.15	1.71	0.38
	Lose	9.32	1.19	
	All	8.88	1.52	
Werewolf	Win	8.61	1.45	0.38
	Lose	9.04	1.37	
	All	8.88	1.42	

Table 3: Analysis results of the utterances sent per agent and day according to the role of the agent.

Table 3 illustrates the analysis result of the utterances sent by an agent according to its role. Results are coherent with the ones from Table 2, as possessed and werewolf role players (werewolves' side) tend to send more utterances when they lose while villager and seer role players (villagers' side) send more utterances when they win. In this table, we can also see how possessed and werewolf players, which have a lower winning rate, send fewer utterances on average than villager and seer players. Note that in the case of possessed players, there is an increase of 1.17 points on the average of utterances sent when they lose the game compared to the times when they win.

Agent	Result	Utterances Avg.	Utterances Std. Dev.	Win. Rate
A1	Win	7.63	1.08	0.52
	Lose	7.59	1.51	
	All	7.61	1.31	
A2	Win	9.63	0.77	0.52
	Lose	9.5	1.1	
	All	9.57	0.95	
A3	Win	9.54	0.46	0.38
	Lose	9.46	0.72	
	All	9.49	0.64	
A4	Win	9.57	0.63	0.58
	Lose	9.64	0.59	
	All	9.6	0.62	
A5	Win	8.64	1.65	0.62
	Lose	9.02	1.23	
	All	8.78	1.51	

Table 4: Analysis results of the utterances sent per agent in a day.

Finally, Table 4 presents the analysis results of the number of utterances sent per player in a day. As expected because of the talkativeness of A4, it presents the highest average of utterances sent from among all agents (9.6), which is 0.61 points above the average. Additionally, A4 also has the second-highest winning rate. It is interesting to observe though, how A5 has the highest winning rate but has the second-lowest average of sent utterances. Additionally, A4 and A5, the players with the highest winning rate, are the only ones with a higher average of utterances sent in losing games than in winning games.

5 Discussion

From the results obtained in Section 4, it seems that the number of utterances sent by the AI agents participating in the shared task was quite similar (around 9 utterances per player and day). The difference in the average of utterances sent by winning and losing players is also not significant. Consequently, we may be able to conclude that the number of utterances sent by the participant AI agents of the shared task may not be a significant

factor that determines the game result. However, since the winning rate of villagers' side players (0.62) is higher than the one of the werewolves' side players (0.38), it seems that other factors are leading certain players to victory.

One of the elements that may affect is the strategy implemented for each of the agents. As an example, the talkative agent implemented (A4) achieves a 0.46 winning rate when playing on the werewolves' side by following a strategy of faking the seer under certain circumstances.

Two other elements that may also affect are the content provided in an agent's utterances and the way it processes utterances performed by other agents. Note that, in the natural language division of the Werewolf game, performing appropriate utterances may be as important as listening and understanding the other agent utterances.

6 Conclusions and Future Work

This paper presented a data analysis on some logs from the shared task of AIWolfDial2019 workshop from INLG2019. The goal was to verify if the talkativeness of an AI agent could have an impact on the Werewolf game result.

Overall results showed that the number of utterances may not be a determinant factor influencing the result of the games played by the AI agents participating in the shared task.

Some questions that future work in this line may address could be: (i) are there some kinds of utterances that lead to winning or losing the game?; (ii) to what extent is the utterances' content important as far as a good game strategy is followed?; and (iii) how much do the other agents' strategy and conversational capabilities influence an agent's result?

This paper analyzed logs generated by only five AI werewolf agents, each with their own unique and independent strategies and conversational capabilities. Because the results of the analysis may depend on the implementation of the agents, it would also be interesting to analyze more data generated by a larger variety of agents.

References

Gokul Chittaranjan and Hayley Hung. 2010. Are you awerewolf? detecting deceptive roles and outcomes in a conversational role-playing game. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5334–5337. IEEE.

Yuya Hirata, Michimasa Inaba, Kenichi Takahashi, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Kousuke Shinoda. 2016. Werewolf game modeling using action probabilities based on play log analysis. In *International Conference on Computers and Games*, pages 103–114. Springer.

Hayley Hung and Gokul Chittaranjan. 2010. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 879–882. ACM.

Yoshinobu Kano, Claus Aranha, Hirotaka Osawa, Daisuke Katagami, Takashi Otsuki, and Fujio Toriumi. 2019. Overview of the aiwolfdial 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). Miraikan, Tokyo, Japan. 2019/10/29*.

Daisuke Katagami, Shono Takaku, Michimasa Inaba, Hirotaka Osawa, Kosuke Shinoda, Junji Nishino, and Fujio Toriumi. 2014. Investigation of the effects of nonverbal information on werewolf. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 982–987. IEEE.

Noritsugu Nakamura, Michimasa Inaba, Kenichi Takahashi, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Kousuke 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 (SSCI)*, pages 1–8. IEEE.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.