Title	Solving the werewolf puzzle by communication between agents [課題研究報告書]
Author(s)	張, 夢柔
Citation	
Issue Date	2019-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/15900
Rights	
Description	Supervisor:東条　敏, 先端科学技術研究科, 修士（情報科学）

Master's Research Project Report

Solving the werewolf puzzle by communication between agents

1710125 ZHANG MENGROU

Supervisor	Tojo Satoshi
Main Examiner	Tojo Satoshi
Examiners	Shirai Kiyoaki
	Iida Hiroyuki
	Ogata Kazuhiro

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

February 2019

Abstract

In this research, we are talking about the methods that develop an AI agent for solving the werewolf game just like human beings. Werewolf game is party game at first. That means the information you get from other players is not only the conversation. In werewolf game of the real world, the emotions, the movements(body languages), the sound when you close your eyes and even the tones of voices, these are essential information to win the game. Moreover, the werewolf game is an incomplete information game. The deficiency of the information makes it is difficult to be resolved. And the werewolf game is a social game that players play the game by communication. It makes the werewolf game more difficult than others just like Majiang. However, this is also the charm point of the game that attracts many researchers to study it.

Here we introduce several pieces of research of werewolf agent. The first one, we talk about the AI Wolf Project. This is a group based on several researchers who are interested in the werewolf game. And they developed a platform that could let Agents play werewolf game together. They hold the competition of werewolf agents every year. This platform provides two types of agents to develop. One is the protocol type, and another one is natural language branch type. To consider that communication by natural language is too difficult to achieve, developers who want to concentrate on inference could choose the protocol type. The rules of the protocol provide a limited language option. The natural language type agent uses natural language(Japanese) to communicate with each other. So to reduce the difficulty of this type of agent, the rules of the game is more straightforward than protocol type.Next, we introduce a piece of research that use an extend BDI model to describe the werewolf game. They extended the BDI model by adding probabilistic intentions into the model to solve the problem that the beliefs of the agent are uncertain. Next, there is a piece of research that developed a werewolf game model based on play log from werewolf bbs. So then we introduce a study developed a werewolf game agent based on deep reinforcement learning. As a result, the win rate of the agent is better than the agents exist.

The most common method to make a werewolf agent is using machine learning, which is famous in solving perfect information games. Also, according to the results, machine learning and deep learning are useful to improve the win rate of werewolf agent. However, the data of human beings is not enough to make the agent powerful as human beings. Moreover, the log of the game is too difficult to analyze because it has a large amount of redundant information. Moreover, this research introduces several logic methods and show how to use these methods to solve the werewolf game. The first method is the modal logic. Modal logic is

developed in the 1960s. It is an extending of the first-order logic by adding two symbols, must and may. So the modal logic has several different axioms. The elementary axiom of modal logic is axiom K, and by extending the axiom K, we received the other axioms. Possible world theory is based on modal logic. In this theory they definite that W is a set of possible worlds, R is a set of connections and v is the value of the propositions which is in possible worlds. This theory could describe while the agent cannot be sure about the values of propositions by connecting the possible worlds.In the next part, we introduce Dynamic epistemic logic and Public Announcement Logic. Dynamic epistemic logic (DEL) is a logical framework for dealing with changes in knowledge and information. Public Announcement Logic (PAL) is a study of modal logic of knowledge, belief, and public communication. Combine the Public Announcement Logic with possible world theory, and we can show how the beliefs of agents change. This is considered that suitable to solve the werewolf game. But even you can describe the werewolf game by possible world completely. The uncertainty of this game and the less information let that making a decision is too complicated. While the werewolf agent model based on BDI logic also facing this problem. There still need an algorithm to resist the uncertainty of the werewolf game. The next part is the Mental Spaces. Mental spaces theory is an excellent way to analyze some difficult natural language. We can use mental spaces to describe the thinking of the agents. And it also could be dynamic to represent the changing of the thought of agent. But it is difficult to reason the real mind of the agent.

The next part is about a study that develops an Observation Model of Werewolf Game. In this study, they propose a model of belief and intention change throughout a dialogue. They use Situation Calculus to model to analyze the evolution of the world and an observation model to analyze the evolution of intentions and beliefs. This model is an observation model and it could describe dialogues combined with actions. The goal of this study is to model the interaction between beliefs, intentions, and utterances. By using this model they can predict decisions resulting from the dialogue is used as a performance measure.The Observation Model provides a new view of the werewolf game. By using this model, the agent could predict the effect of his act. In that example, the seer predicts the result of voting after his coming out. While the agent could predict the result of every act he doses, he could choose the most benefit act to reach the intention of the games. This is allowed the agent to persuade other agents and win the game.

In the past studies, researchers absorbed in finding the werewolves to prove the win rate. However, here is a new idea that not only exposes the werewolf but also persuades other agents to achieve success. The Observation Model concentrates on the observe so it can not describe the werewolf game complete. In future work, we could extend the model entirely and build an agent that good to persuade others.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays, making AI solves perfect information games is possible[1]. In 1997, Deep Blue defeated the world chess champion Garry Kasparov in chess. Recently, Alpha Go defeated several top players of Go. Go is considered as one of the most complicated perfect information game and what means AI has already become Invincible in this field. In perfect information games, the information received by both sides of the game is completely equivalent. In this type of game, AI only needs to search for the winning rate in each case after each calculation based on the current disk. But how about incomplete information games? In incomplete information games, the information received by both sides of the game is not completely equivalent. The player needs to guess the information that in other players' hand and inferential measurement the win rate for every choice.

Werewolf game is one of the incomplete information game. It is also a popular card game all over the world. The feature of the werewolf game is that the game is played in the player's communication. In this game, players communicate with each other and share information. In this process, some player may lie to others and try to persuade others to believe the lies. The AI which could solve the werewolf game perfectly is not developed until now.

To solve the werewolf game, AI needs to guess others' intention and try to let others believe itself. It is similar to social intercourse in our daily life. In the future, we may live a life with robots. This study may help the robot to understand the intention of human beings and make the communication between robots and human beings more smoothly.

More and more researchers pay attention to the werewolf game. Moreover, do reseaarches are interrelated to the werewolf game. There are some excellent results, but let agent solve the werewolf game as human beings is still not possible. The primary method to build the AI agent is machine learning. Moreover, we need more trying in this filed.

Chapter 2

Researches About Werewolf Game

2.1 Werewolf Games

The official name of the game is The Werewolves of Millers Hollow. Also, in Japan, the game called Are you a werewolf. The Werewolves of Millers Hollow is a card game based on a Russian game named Mafia. The story occurs in the deep of American countryside. There is a little town named Millers Hollow.Moreover, recently there are werewolves appear in this town. Every night, a townsperson is murdered. So the other inhabitants in Millers Hollow decide to find the evil murdered. In this case, the werewolf is looked as same as the human. They only killed people after everyone in the town falls asleep. So it is difficult to distinguish if someone is a human or a werewolf.

2.1.1 Rules of Werewolf Game

The rules of the werewolf game have several editions. And here I choose the edition written by Philippe des Pallières [2] and Hervé Marly to introduce. There two sides in this game, the townsfolk side, and the werewolves side. For the townfolk side, the goal of the game is to kill whole of the werewolves. And for the werewolves side, the goal of this game is to kill whole of the townsfolk.

Roles

The werewolf side: Werewolves kill 1 of townsperson at each night. The daytime, they try to hide those identities and a selfish act from a city person. The number of players decides the number of werewolves. There are 1 4 werewolves in one play. The townsfolk side: There is one townsfolk killed by werewolves at each night. This player is outing the playing. All of the survivors gather in the morning in front of the town square, and they try to find which one is a werewolf. The werewolves will hide their real face and join the discussion of the killer. Players will vote after the argument and discussion to kill a suspect. There are eight different kinds of Townsfolk: (Figure 2.1)

Figure 2.1: Roles Cards in Werewolf Game

Ordinary townsfolk: These people don't have any ability except for their intuition at all. Each ordinary Townsperson has to analyze player's behavior, not to make a mistake by mistake for a werewolf, to be killed by lynching excessively, to be sentenced to the gallows and to try to be burned to guess who a werewolf is.

The fortune teller: A fortuneteller can see one player's true celebrity at each night. A fortuneteller chooses which player this is. Without being found by a werewolf outside, a fortuneteller has to help that other city persons distinguish a werewolf right, follow and set their sights on him/herself.

The hunter: Whether a Hunter is killed by a werewolf or by voting, he/she'd be able to retaliate. When the hunter died, he/she could shoot one of the players.

Cupido: "Cupido" is a town go-between. He/she got this diminutive because he/she can make two persons be in love with each other. Cupido will link two persons together and make them fall in love on the first night of the game. So they will fall in love until the game ending. If the Cupido want, he/she could be half of the lovers. If one of the lovers dies, another one will suicide at once. A lover cannot vote to lynch the other one.

Special Case: If one person of the lovers is a Werewolf and the other person belongs to the human side, the objective of the lovers is changed. The lovers' only wish is to live together. For achieve the wish, they should kill other players (Werewolves side and Townsfolk side) in this game.

The witch: The witch has two bottles of powerful potions: one is a healing potion, it could bring the dead person killed by werewolves back to life. Another one is a poison, which can be used at night to kill one player. Every potion can only be used once in the game. The Witch can use either potion on him/herself if he/she wishes. The witch can use poison at night and the healing potion the next day. With this role, you can spend a night without someone dying or one player dying or two players dying.

The Little girl: The little girl is inquisitive. She can open her eyes and peep at the werewolf at night. However, if the werewolf catches her, she will immediately die instead of the designated victim. The little girl cannot spy all night, only in the ”Werewolf Wake Up” stage. When a little girl enters the game, all players should not hide their faces by (hands, cards, etc.) while sleeping.

The Sheriff: Leave a person inside the player instead of this card being distributed like other characters. This player was voted as a city man, just like a real sheriff. The player who gets the most votes will become the sheriff. Once selected, the Sheriff cannot refuse this honor. From now on, this player's vote will be counted as two votes (it always applies to a single player, it will not break, see the bottom vote.) When the sheriff is eliminated, he/she can choose a player to be the new sheriff.

Thief: If you use a thief, add two more regular Townsfolk cards to the deck at the beginning of the game. After shuffling and processing all the cards, place, two special cards face down on the table. In the initial turn of the game, the thief will look at the two cards and exchange one of the two individual cards for his/her cards. However, if both cards are werewolves, then the thief should exchange his cards with the wolf card. If the thief takes away an extra hand, he/she will play the role of the card in the remaining games.

Setting up the game

A. The player chooses one as the moderator (by random, voting or other means of choice). For a new set of players, the host should be familiar with the rules of the game. The host is the most critical player in this game. Their job is to create an intense atmosphere and make the game truly enjoyable.

B. The host takes out the appropriate number of cards, shuffles the cards, and gives the character cards (face down) to everyone. Then the player should look at their cards in secret and place them face down in front of them.

C. The host let the town fall asleep and said, ”The night falls, the town is asleep, everyone closes their eyes.” Each player now closes his eyes. Then, depending on the characteristics of the card being played.

D. The host said, ”The Thief woke up.” The player who got the thief card opened his/her eyes, quietly looked at the two cards on the table, and possibly switched the Thief character card with one of the two cards (See The Thief above). Then the host said: ”The thief is asleep.” The thief closed his eyes.

E. The host wakes up the Cupido by saying ”Cupido to wake up.” The player who got the Cupido card opened his/her eyes and silently designated two players (probably his/her own). The host walked around and carefully patted the shoulders of the two lovers (the lover should close his eyes). The host then said, ”Cupido is asleep.” Cupido closed his/her eyes.

F. The host called the lover ”the lover wakes up, knows each other, and returns to a happy sleep.” The lovers look at each other. They didn’t show each other their cards, so they didn’t know the true character of their lover.

Standard Turn

1 - The host calls the fortune teller.

The host said: ”The fortune teller wakes up and chooses a player she wants to know about her true character.”

The host presents the character card of the selected player to the Fortune teller. After that, the host said, ”The fortune teller is asleep.”

2 - The host called the werewolf.

The host said: ”werewolf wake up and know each other and choose new victims.”

The werewolves opened their eyes and looked at each other silently, designating a new victim. At this stage, the little girl can monitor the werewolf by opening her eyes, peeping or anything she thinks she can escape. She does not need to do this, but she can do it if she dares.

If the little girl is arrested, she will die instead of the designated victim.

Then, the host said: ”Werewolf falls asleep again after satisfying their human needs” (or something similar). Then the werewolf closed his eyes.

The moderator will announce the results the next morning.

3 - The host called the witch.

The host said: ”The witch woke up. I will tell you the victim of the werewolf. Do you want to use your potion or poison?”

The host showed the witch to the werewolf’s victim. If the player does not want to, the witch is not obliged to use the potion. If the witch uses the potion, he/she will indicate the treatment by showing a thumbs up. If the poison is poisoned, he/she will be poisoned and which player will be poisoned.

The moderator will announce the results the next morning.

4 - The host wakes up the Town.

The host said: ”The sun has risen. Everyone wakes up and opens their eyes... Everyone is all right except for the one who was killed at night.”

The host now points to the victim of the werewolf. The player has his/her face up and is not in the game. This player can no longer communicate with other players.

If the slain player is a hunter, he/she can retaliate before dying and choose other players to die immediately. If the player being eliminated is one of the lovers, then the other lover will immediately die.

5 - The Townsfolks argue and debate.

The moderator is gentle here and organizes debates (also known as angry thugs!).

Hearing a strange sound at night, suspicious behavior, a player who always votes with another person. All of these are clues that can make players suspect that they are werewolves. During the discussion, the player’s goals are different:

- Every townsfolks tried to uncover a werewolf and lynch him.
- Werewolf tries to disguise itself as an ordinary townsfolks.
- The ortune teller and the little girl try to help urban residents but don’t reveal themselves, which can be dangerous or fatal.
- Lovers try to protect each other.

All players can pretend they are anything. You can bluff or tell the truth, but you better convince you what to say. And you not only have to protect yourself, but you must also know as much as possible about other players.

Believe us, you always give up in one way or others. Everyone has a ”tell”.

6 - The Town votes.

Players must destroy players they suspect are werewolves.

According to the moderator’s signal, each player points his/her finger at the same time they want to eliminate the player (don’t forget the Sheriff’s voting value of two).

The player whose finger pointed at them was judged to be a werewolf and was immediately lynched, hanged, drowned, fired with silver bullets and burned (just to be sure).

This player has now disappeared from the game. If there is a relationship between the defendants, no one will be eliminated.

Players who are eliminated will show their cards and will not be able to communicate with other players until the game is over.

7 - The Town Falls asleep.

The host said, ”The night falls. The surviving players fall asleep.” All players close their eyes. Players who are eliminated must remain silent, especially when they find out who the werewolf is. The game starts again, the first step, the Fortune Teller...

Advice from the author.

Werewolf: Vote against your partner is an effective way to transfer your suspicions. Of course, this is only effective if the public is concerned.

Fortune teller: If you find a werewolf, be very careful. It may be worthwhile to show your pain in order to identify the player but avoid prematurely doing so.

Little girl: A compelling character, but it is very nerve-racking to play. Do not hesitate to spy. This can be terrible, but it must be profitable quickly before being eliminated.

Hunter: When you are accused of being a werewolf, you always try to make yourself a hunter.

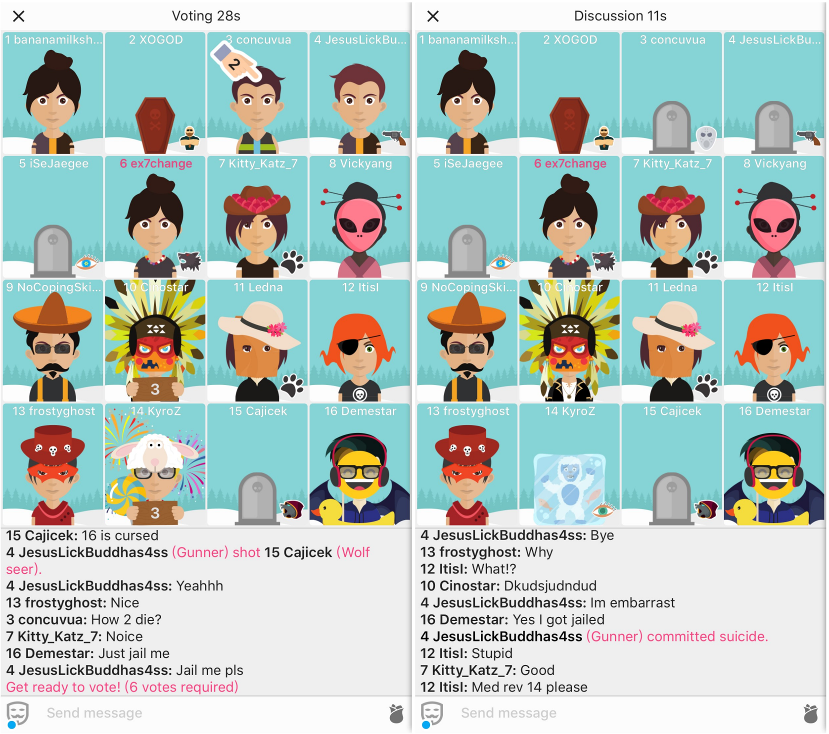
Cupido: If you choose yourself as a lover, don't choose a quarrel to be your partner.

Sergeant: Do not hesitate to nominate yourself as a sheriff. However, if you are a werewolf, then this job will not be too open for a public campaign. Be proud of this position and wear the card proudly on the clothes.

Witch: This character becomes even more powerful at the end of the game - don't waste your potion!

2.1.2 Werewolf Game Online

You could play the werewolf game online now, by a computer or a smartphone. The system acts the moderator and the turn of the game set by programs. You can play werewolf game online with people all over the world. The online game is different from the party game. When you play with someone face to face, the movements, the emotions and even the tones of speaking will give you useful information about the game. But in the online game, you can not receive the messages like this. So you need to concentrate on the words and try to analyze the intention of other players. According to the online game system, agents play werewolf is becoming possible.



15 Cajicek: 16 is cursed
4 JesusLickBuddhas4ss (Gunner) shot **15 Cajicek** (Wolf seer).
4 JesusLickBuddhas4ss: Yeahhh
13 frostyghost: Nice
3 concuvua: How 2 die?
7 Kitty_Katz_7: Noice
16 Demestar: Just jail me
4 JesusLickBuddhas4ss: Jail me pls
Get ready to vote! (6 votes required)

4 JesusLickBuddhas4ss: Bye
13 frostyghost: Why
12 Itisl: What!?
10 Cinostar: Dkudsjudndud
4 JesusLickBuddhas4ss: Im embarrast
16 Demestar: Yes I got jailed
4 JesusLickBuddhas4ss (Gunner) committed suicide.
12 Itisl: Stupid
7 Kitty_Katz_7: Good
12 Itisl: Med rev 14 please

Figure 2.2: The Game Werewolf Online

Werewolf Online(Figure 2.2) is an example of online werewolf game. It provides above ten languages to players, but not include Japanese. By this application, you can enjoy the werewolf game very quickly. One game takes about ten minutes.

2.1.3 Werewolf BBS

Werewolf BBS is most like a BBS. At first you choose a village you like. The different village has different rules of the game. And when the number of players is enough, the game begins. Every player gets his/her own character security. And then you play the role and talk with each other. The special of werewolf BBS is that they use time just as the real world. It means a game takes over ten days. And you can say things have nothing to do with the game just as in a BBS. Here you can play the game and communication with others at the same time.

2.2 Preceding Works

2.2.1 Artificial Intelligence based Werewolf(AI Wolf project)

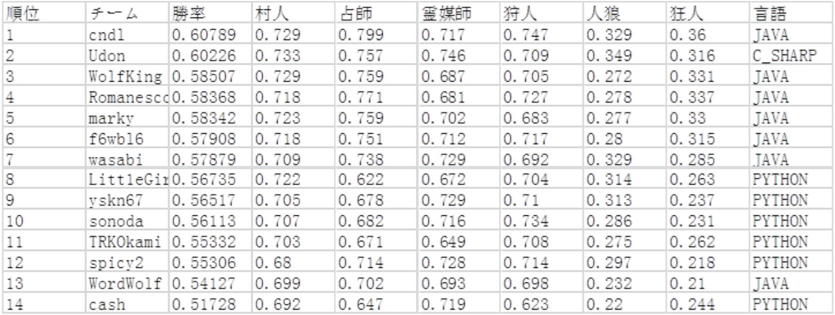
AI Wolf Project [3] is a group based on several researchers who are interested in werewolf game. Also, they are trying to attract more researchers to attend this work that to create AI which can play werewolf game as good as human beings. The final goal of this project is to build AI which can communicate with human beings naturally in the playing of werewolf game.

The researchers developed a platform that could let Agents play werewolf game together. Moreover, they designed rules of werewolf game between agents to let them play smoothly. Then they held events that allowed agents all over the world compete together to motivate the AI to become more powerful.

This platform provides two types of agents to develop. One is the protocol type, and another one is natural language branch type. To consider that communication by natural language is too difficult to achieve, developers who want to concentrate on inference could choose the protocol type. The rules of the protocol provide a limited language option. For example, in this game, agents can only say simple sentences like ”I am a seer,” ”I agree with agent A” and so on. The developers do not need to care how to deal with languages. The system will change these sentences to symbols. By this system, become a werewolf AI developer is becoming more easier. The natural language type agent uses natural language(Japanese) to communicate with each other. So to reduce the difficulty of this type of agent, the rules of the game is more straightforward than protocol type.

This is an open source project. So the standard way to join this project is to develop your own werewolf AI based on the winner of the last competition. The primary method is to improve the algorithm and use machine learning.

According to the result(Figure 2.3) of the last competition, the win rate of the werewolf is 0.329 while in the actual game the number is 0.522. There still a long road to solving the werewolf game by AI.



順位	チーム	勝率	村人	占師	霊媒師	狩人	人狼	狂人	言語
1	cndl	0.60789	0.729	0.799	0.717	0.747	0.329	0.36	JAVA
2	Udon	0.60226	0.733	0.757	0.746	0.709	0.349	0.316	C_SHARP
3	WolfKing	0.58507	0.729	0.759	0.687	0.705	0.272	0.331	JAVA
4	Romanescc	0.58368	0.718	0.771	0.681	0.727	0.278	0.337	JAVA
5	marky	0.58342	0.723	0.759	0.702	0.683	0.277	0.33	JAVA
6	f6wb16	0.57908	0.718	0.751	0.712	0.717	0.28	0.315	JAVA
7	wasabi	0.57879	0.709	0.738	0.729	0.692	0.329	0.285	JAVA
8	LittleGir	0.56735	0.722	0.622	0.672	0.704	0.314	0.263	PYTHON
9	yskn67	0.56517	0.705	0.678	0.729	0.71	0.313	0.237	PYTHON
10	sonoda	0.56113	0.707	0.682	0.716	0.734	0.286	0.231	PYTHON
11	TRKOkami	0.55332	0.703	0.671	0.649	0.708	0.275	0.262	PYTHON
12	spicy2	0.55306	0.68	0.714	0.728	0.714	0.297	0.218	PYTHON
13	WordWolf	0.54127	0.699	0.702	0.693	0.698	0.232	0.21	JAVA
14	cash	0.51728	0.692	0.647	0.719	0.623	0.22	0.244	PYTHON

Figure 2.3: The Result of last Competition

2.2.2 An Extended BDI Model of Werewolf Game

Nide Naoyuki and Shiro Takata [4] developed this model of werewolf game. They use an extend BDI model to describe the werewolf game. This model describe beliefs.desires and intentions of the agent. They extended the BDI model by adding probabilistic intentions into the model to solve the problem that the beliefs of the agent are uncertain. Moreover, by this model, some of the truth can be proved by the inference rule. However, there still lacks an algorithm to infer goals efficiently.

2.2.3 A Werewolf Game Modeling Based on Play Log

This study developed a werewolf game model based on play log from werewolf bbs by Yuya Hirata and his team[5]. In this study, they collect 467 play logs while 223 villager side wins and 224 werewolf side wins. Also, because of the analysis of a large amount of sentences is too difficult, They only pick up the log that the seer is coming out and other agent's decision about this. According to the data they create an agent. However, the testing result is a lot different from the win rate of the actual game. By using more data, the agent will be more potent in the future.

2.2.4 Werewolf Agent based on Deep Reinforcement Learning

This study developed a werewolf game agent based on deep reinforcement learning by Tianhe Wang and Tomoyuki Kaneko[6]. This study focuses on the problem that to decide whom to trust or to kill. They built the werewolf game agent by using Deep Q network. Also, they trained the agent by Q learning method. They used the reinforcement learning method instead of the heuristic method to solve the problem above. As a result, the win rate of the agent is better than the agents exist. In the future, they will try to use asynchronous methods to make the agent make proper decisions in diverse environments.

2.2.5 A Werewolf Match System for Humans and Lifelike Agent

This study developed a werewolf match system for human players and lifelike agents by Yu Kobayashi and his team [7] . In this study, they use MMD agent to create a platform where human beings and AI agent could play werewolf game together. MMD agent is a cartoon character. With the control of the player, MMD agent could move and speak. Both human beings and the lifelike agent could control the MMD agent to play werewolf game. The advanced of MMD agent is that it could communicate with other programs. In the future, they will develop a werewolf AI to have a conversation by controlling the MMD agent.

2.2.6 Nonverbal Information inWerewolf Game

In this study, the authors [8] describe that someday AI werewolf can sit beside human beings and play werewolf game together. To make a werewolf just like AI, not only the natural language but also the body language is very important. They analyzed movies of human beings playing werewolf game and combined the body languages, the roles, and the decisions they made. They deal with these data bu machine learning. According to the research, nonverbal information is essential in werewolf game. It will be the key to win. In future works, they will analyze the relationship between verbal and non-verbal information.

Chapter 3

Theoretical Background

3.1 Modal Logic and Possible World

3.1.1 Modal Logic

Modal logic[9] is developed in the 1960s. The first-order logic is not enough to describe all situations of natural language. So they extended the first-order logic by adding two new connectives and create modal logic. Modal logic has several different axioms(Figure 3.1) .

Connectives:$\neg,\rightarrow,\wedge,\vee,\diamond,\Box$

While A is a proposition,$\Box$A and $\diamond$A is also a proposition.

$\Box$A:A is necessary.

$\Box p := \neg\diamond\neg p$

$\diamond$A:A is possible.

$\diamond p := \neg\Box\neg p$

Axiom K(Figure 3.2)

In modal logic, axiom K, named after Saul Kripke, is a basic principle which almost all versions of propositional modal logic satisfy.

A_1: $(A\rightarrow(B\rightarrow A))$

A_2: $(A\rightarrow(B\rightarrow C))\rightarrow((A\rightarrow B)\rightarrow(A\rightarrow C))$

A_3: $(\neg B\rightarrow\neg A)\rightarrow((\neg B\rightarrow A)\rightarrow B)$

A_4: $\Box(A\rightarrow B)\rightarrow(\Box A\rightarrow\Box B)$

Axiom D=Axiom K + $\Box A\rightarrow\neg\Box\neg A$

Axiom T=Axiom K + $\Box A\rightarrow A$

Axiom S4=Axiom T + $\Box A\rightarrow\Box\Box A$

Axiom S5=Axiom T + $\diamond A\rightarrow\Box\diamond A$

Axiom B=Axiom T + $A\rightarrow\Box\diamond A$

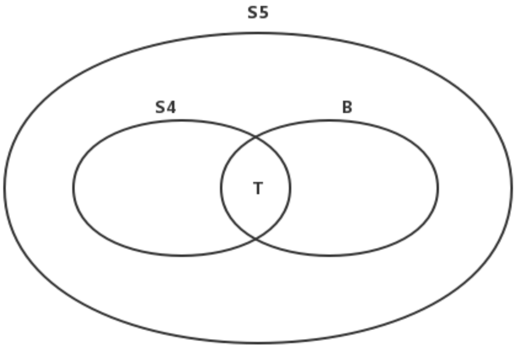
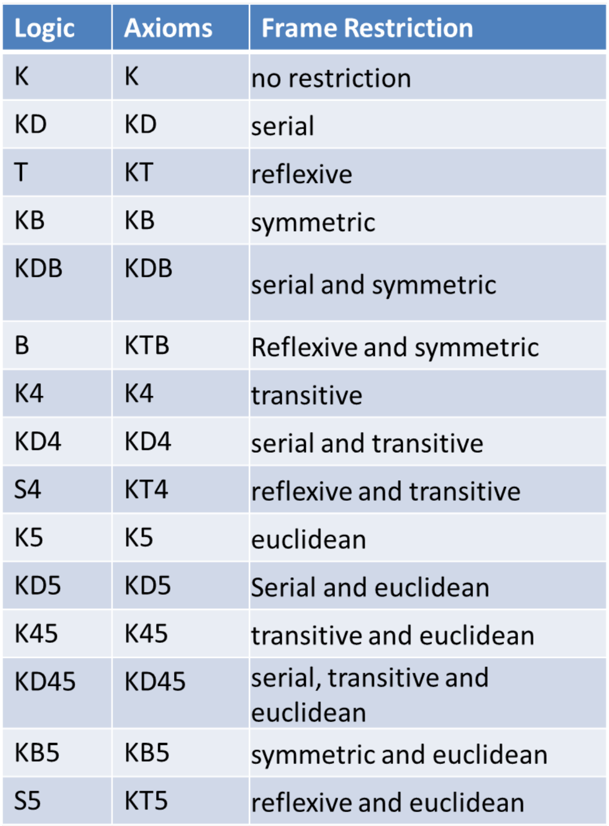


Figure 3.1: The relations of Axioms



Logic	Axioms	Frame Restriction
K	K	no restriction
KD	KD	serial
T	KT	reflexive
KB	KB	symmetric
KDB	KDB	serial and symmetric
B	KTB	Reflexive and symmetric
K4	K4	transitive
KD4	KD4	serial and transitive
S4	KT4	reflexive and transitive
K5	K5	euclidean
KD5	KD5	Serial and euclidean
K45	K45	transitive and euclidean
KD45	KD45	serial, transitive and euclidean
KB5	KB5	symmetric and euclidean
S5	KT5	reflexive and euclidean

Figure 3.2: The Restrictions of Axioms

3.1.2 Possible World

Definition: W= { w_1,w_2,w_3,... } is a set of possible worlds, R={w_1Rw_2,...} is a set of connections, $< W,R >$ is a Kripke frame. w$\models$φ means in possible world w, φ is true.

w$\models$□φ iff for all wRw′, w′$\models$φ.

w$\models$◇φ iff in all wRw′, there exist a w′, w′$\models$φ.

V is the value of the proposition.

V(w,φ)=T/F

For example (Figure 3.3) , here we have three possible worlds,w_1,w_2,w_3, the connections w_1Rw_2, w_1Rw_3, and values V(w_1,φ)=F, V(w_2,φ)=T, V(w_3,φ)=T.

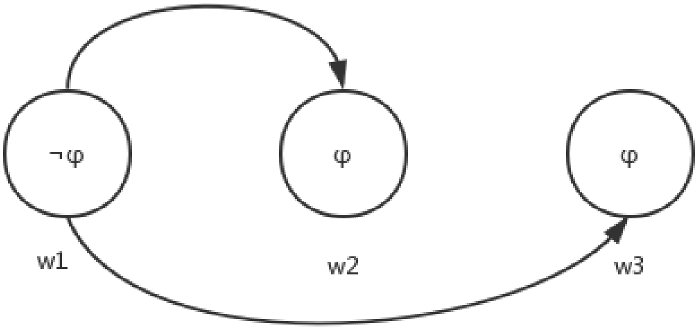


Figure 3.3: An Example of Possible World

For possible world w_1, there are $w_1$$\models$◇φ and $w_1$$\models$□φ.

For possible world w_2, there are $w_2$$\models$□φ and $w_2$$\models$φ.

For possible world w_3, there are $w_3$$\models$□φ and $w_3$$\models$φ.

T: □ φ → φ

For any w′, if w′$\models$□φ, then w′$\models$φ. It needs the start possible world w′ could access itself. This restriction is called reflexive(Figure 3.4) .

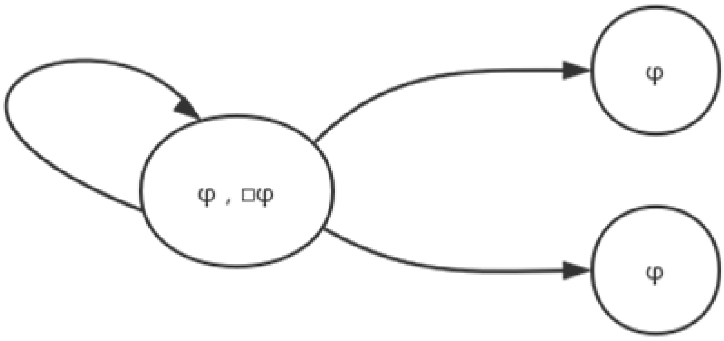


Figure 3.4: reflexive

4: $\Box\varphi\rightarrow\Box\Box\varphi$

In this axiom, if w_1Rw_2,and w_2Rw_3, then w_1Rw_3 is necessary. This restriction is called transitive (Figure 3.5) .

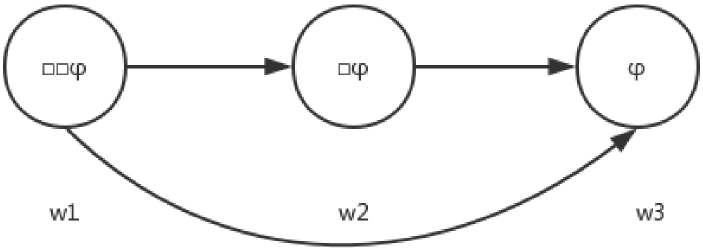


Figure 3.5: transitive

D:$\Box\varphi\rightarrow\diamond\varphi$

For every possible world w, there must exist a possible world w′ that wR w′. This restriction is called serial (Figure 3.6) .

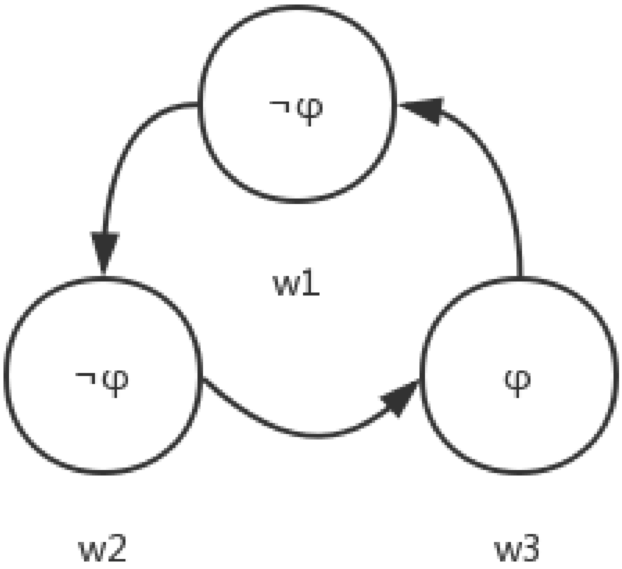


Figure 3.6: serial

B:φ→□◇φ

For every possible world w, if there is a connection wRw′,then there must be a connection that w′Rw. This restriction is called symmetric.

(Figure 3.7)

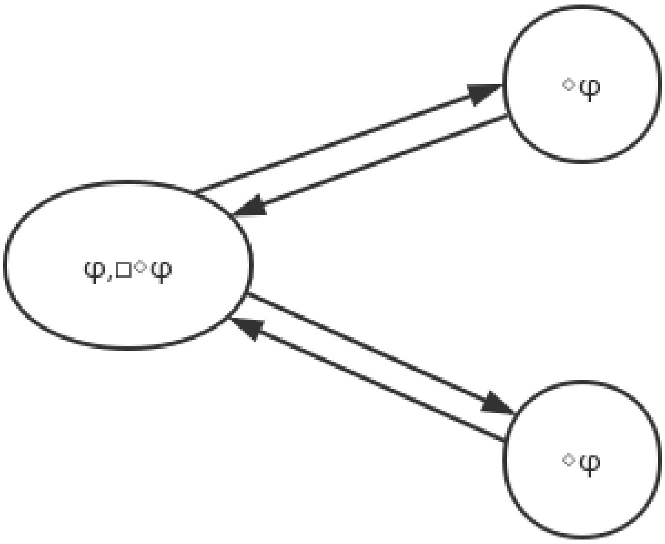


Figure 3.7: symmetric

5:◇φ→□◇φ

For every possible world w, if there are connections wRw′ and wRw″, then there must be a connection that w′Rw″. This restriction is called Euclidean (Figure 3.8) .

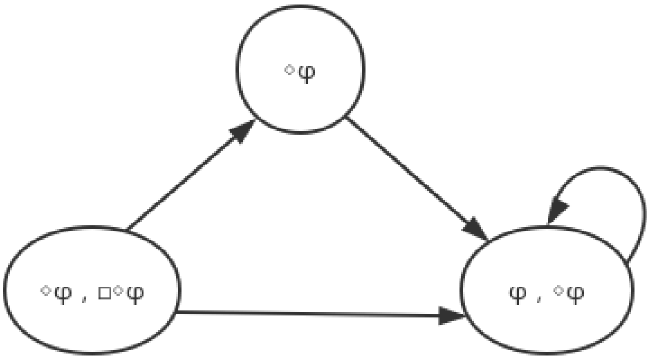


Figure 3.8: Euclidean

3.2 Dynamic Epistemic Logic

Dynamic epistemic logic (DEL)[10] is a logical framework for dealing with changes in knowledge and information. Often, dynamic cognitive logic focuses on situations involving multiple agents and examines how their knowledge changes when events occur. These events can change the factual attributes of the real world.

Public Announcement Logic (PAL) is a study of modal logic of knowledge, belief, and public communication. PAL is used to infer knowledge and beliefs, as well as changes in knowledge and beliefs based on the occurrence of entirely true, authentic announcements. Examples of the most common motivations for PAL include Muddy Children Puzzle and Sum and Product Puzzle.

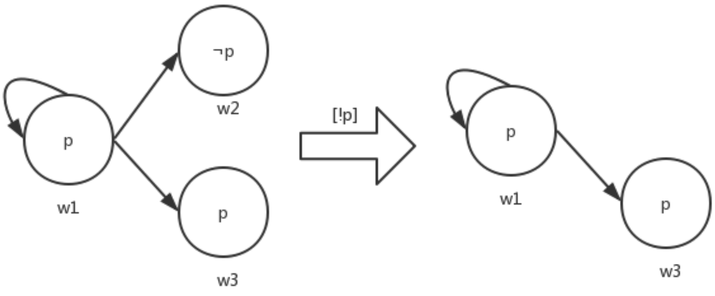
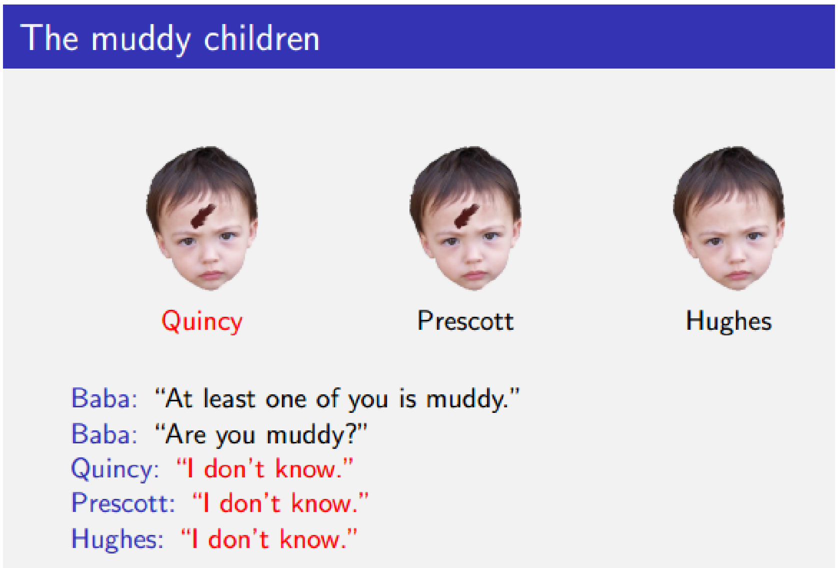


Figure 3.9: An Example of PAL

[!p] means there is a public announcement that p is true (Figure 3.9) . Before the announcement, the agent in possible world w_1 could see p and ¬ p, and when it received the announcement, the beliefe of the agent is changed, so the agent cut the connection of possible world w_2, because it believes the value of p is true.

A famous example of PAL is Muddy Children Puzzle (Figure 3.10) .



The muddy children

Quincy Prescott Hughes

Baba: "At least one of you is muddy."
Baba: "Are you muddy?"
Quincy: "I don't know."
Prescott: "I don't know."
Hughes: "I don't know."

Figure 3.10: Muddy Children Puzzle

Before the game start, eight possible worlds of all possibilities could be created. We use 1 of muddy, 0 of not muddly. Here we choose the visual angle of Quincy to analyze the puzzle. While Baba says at least one is muddy. This could be regarded as a public announcement. And then Quincy cut the connection of the possible world (000) (Figure 3.11) . Then Quincy looks at others. He knows that Prescott is muddy and Hughes is not muddy. So he cut the connections of the possible worlds that different from the real world. So after several times of communication there will be one possible world leave, and that is the real world where Quincy is.

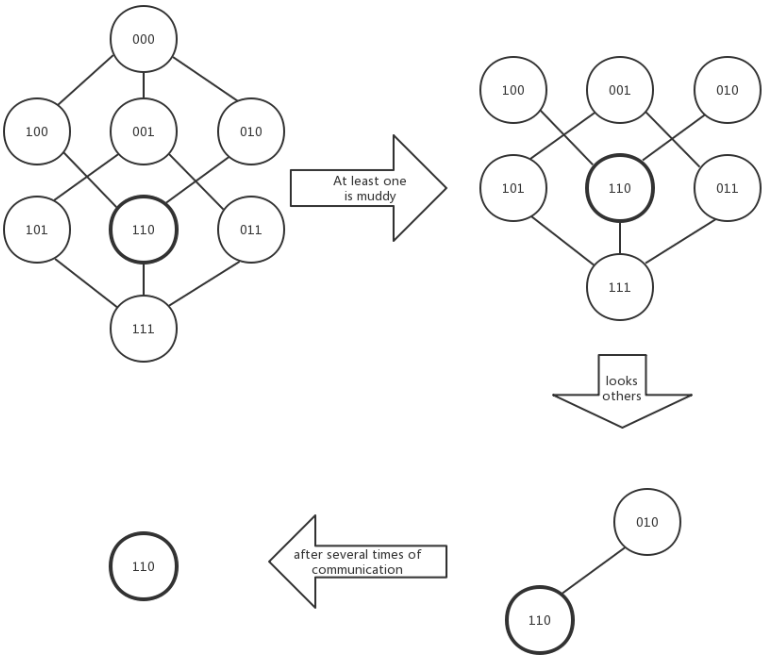
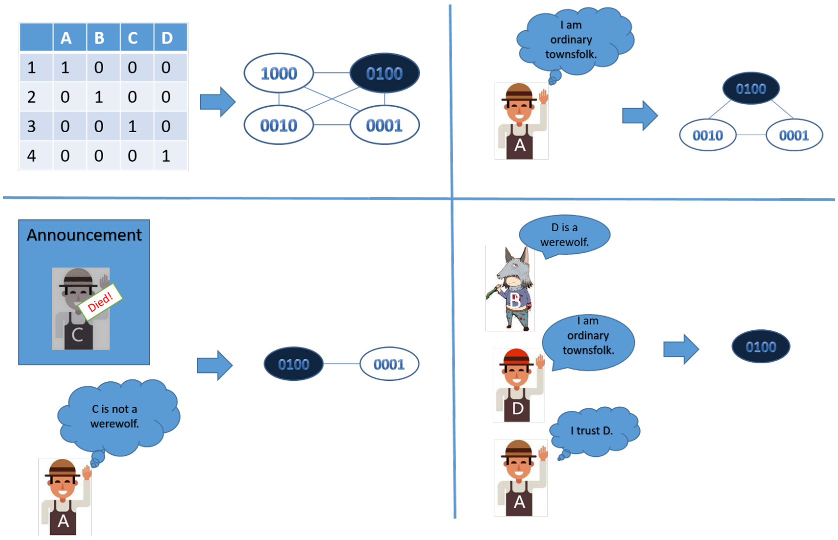


Figure 3.11: Possible World of Muddy Children Puzzle

3.2.1 Using possible world theory to Solve Werewolf Puzzle

This is a simple example(Figure 3.12) to show how to solve werewolf game by using possible world theory.

There are four players in this game, one is a werewolf, three are townsfolks. Here we choose the visual angle of agent A. Before the game start, we could create four possible worlds include every possibility of the game. We use 0 to mean townsfolk and 1 to mean werewolf. While the game starts, agent A gets its character card. The card of agent A is a townsfolk card, so he cut the connection of the possible world where A is a werewolf. Then system announcement says C is killed. That means C is not a werewolf. A cut the connection of a possible world where C is a werewolf. And then players have a meeting to talk about who is the werewolf. After the talking, agent A chooses to believe agent D and get the conclusion that agent B is a werewolf.



	A	B	C	D
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

Figure 3.12: The Game Werewolf Online

3.3 Mental Spaces

Gilles Fauconnier[11] develops mental spaces theory to connect the trigger and the target. In analyzing natural languages, there are a large amount of reference, descriptions, and coreference. It is challenging to let agent understand the real meaning of these sentences. For example, the meaning of ”If I were you, I would hate me.” is different from”If I were you, I would hate myself.”. Mental spaces theory could solve the problem like this.

Identification (ID) Principle

If two objects(in the most general sense), a and b, are linked by a pragmatic function F(b=F(a)), a description of a, d$_a$, may be used to identify its counterpart b.

For example: Tom is on the top shelf (Figure 3.13) .

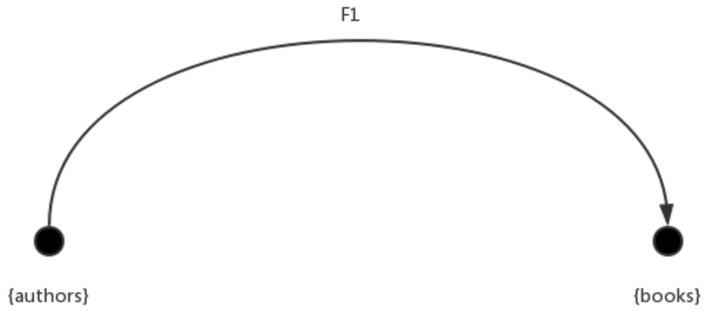


Figure 3.13: Tom is on the top shelf

In this case, a=”Tom,” b=F_1=”books written by Tom”. By the identification principle, this sentence could mean ”The books written by Tom are on the top of shelf”

According to the identification principle, call a the reference trigger (Figure 3.14) , call b the reference target, and F the connector.

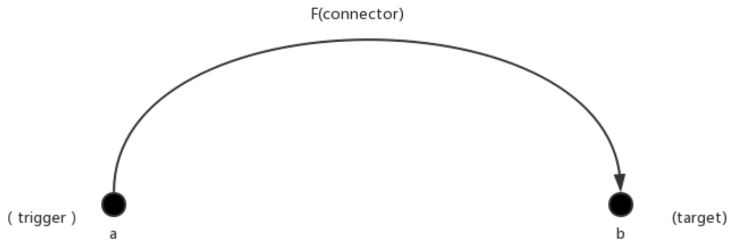


Figure 3.14: trigger and target

An example of images(Figure 3.15) : In the painting, the girl with blue eyes has green eyes.

The images, pictorial representations,photographs, etc, are clearly linked to their models by pragmatic connectors.

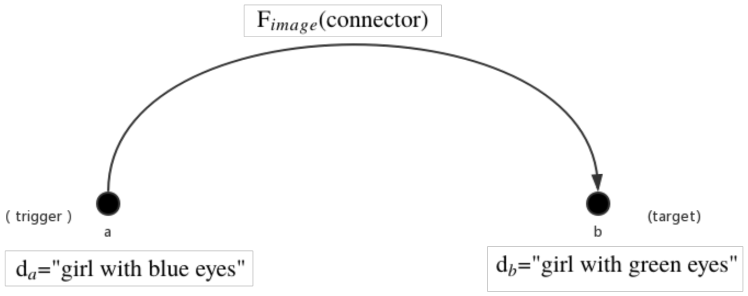


Figure 3.15: An example of images

In this case, the model a, the girl with blue eyes, triggers the image connector F, and the target b, the girl with green eyes, is the representation in thr painting.

Build Mental spaces

For example (Figure 3.16) : Tom believes that in Jack's picture, the flowers are yellow.

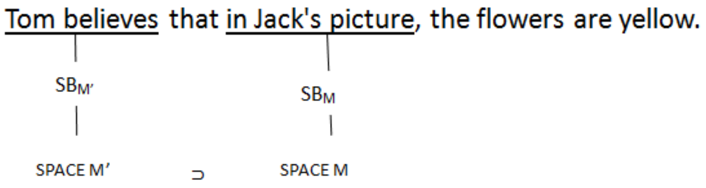

Tom believes that in Jack's picture, the flowers are yellow.
$SB_{M'}$
SB_M
SPACE M'
⊃
SPACE M

Figure 3.16: Build Mental spaces

In this example, Space M is the Jack's picture in Tom's belief, and space M′ is Tom's belief, it is a parent space.

Roles and properties

Example: In 1929, the president was a baby (Figure 3.17) citeref9.

The sentence has more than one meanings.

- (i)In 1939, a baby was president.

  ∃x[w⊩president(x)∧w'⊩baby(x)]

- (ii)The current president was a baby in 1939.

  w' ⊩ ∃x[president(x)∧baby(x)]

- (iii)In 1939, the president were chosen among babies.
  $w' \Vdash \forall x[\text{president}(x) \rightarrow \text{baby}(x)]$

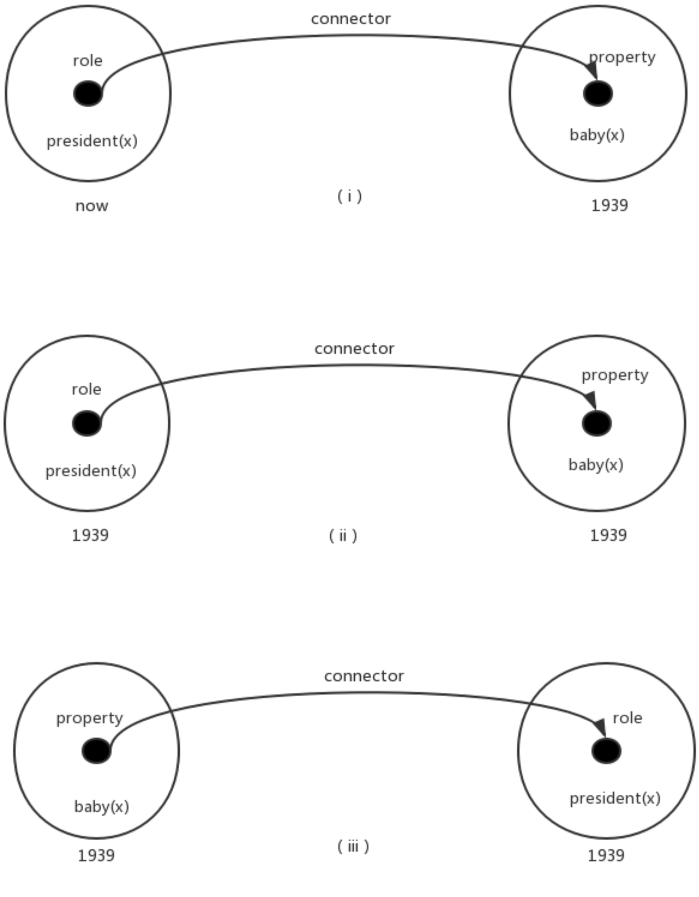


Figure 3.17: Roles and properties

Using Mental Spaces to decribe the werewolf game.

For example: By using Mental Spaces, we could describe the beliefs in the werewolf game (Figure 3.18) .

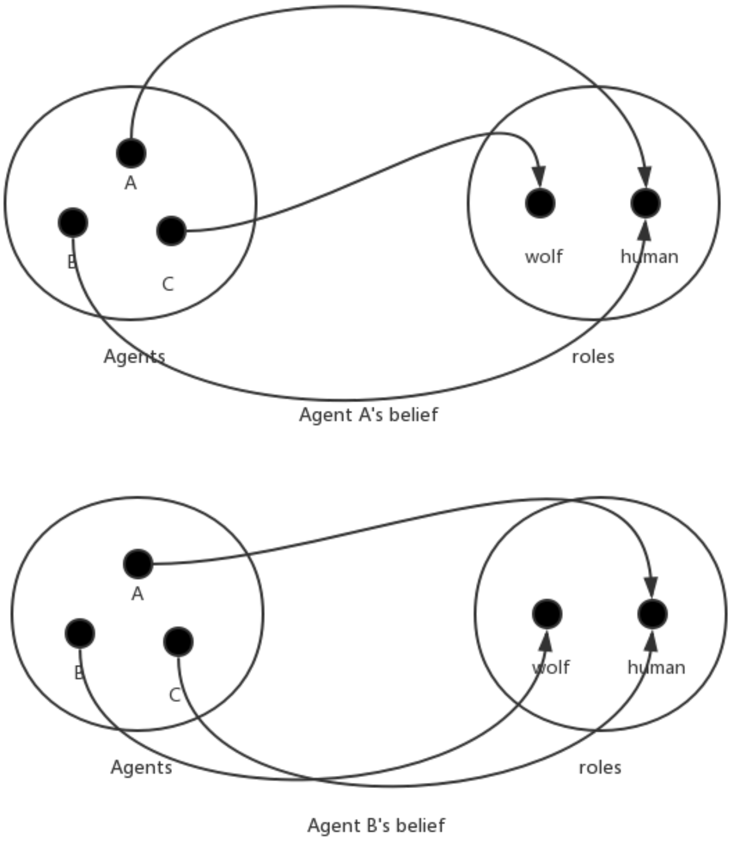


Figure 3.18: Using Mental Spaces to decribe the werewolf game

In agent A’s belief, the role of agent A, B, C is human, human,werewolf. And in agent B’s belief, the role of agent A, B, C is human, werewolf, human.

Chapter 4

An Observation Model of Werewolf Game

This study[12] is developed by Codruta Girlea, Eyal Amir and Roxana Girju. In this study, they propose a model of belief and intention change over the course of a dialogue. They use Situation Calculus to model to analyze the evolution of the world and an observation model to analyze the evolution of intentions and beliefs. This model is an observation model and it could describe dialogues combined with actions. The goal of this study is to model the interaction between beliefs, intentions, and utterances. By using this model they can predict decisions resulting from the dialogue is used as a performance measure.

4.1 Definitions

Ag is a set of agents. R is a set of roles. S is a set of situations. Some roles $R_u \subseteq$R are unique. The set of unique roles for a certain game is known by all agents. The signature specifies sorts α (agent) and σ (situation). M is a set of modalities, F is a set of fluents, C is a set of constant symbles. A is a set of actions.

M={K,B,I,P}(knowledge, belief, intention, persistent intention)

F=$\{\mathrm{r}{:}\alpha\sigma\}_{r \in R} \cup \{\mathrm{dead}{:}\alpha\sigma,\mathrm{alive}{:}\alpha\sigma,\mathrm{voted}{:}\alpha\alpha\sigma\} \cup \{\mathrm{claim}_r : \alpha\alpha\sigma\}_{r \in R}\sigma\{\mathrm{claim}_{\neg r} : \alpha\alpha\sigma\}_{r \in R}$

C=$\{\mathrm{c}{:}\alpha\}_{c \in Ag} \cup \{\mathrm{s}_0{:}\sigma\}$

A=$\{\mathrm{vote(x)}\} \cup \{\mathrm{round}_n\mathrm{(x)}\}_{3 \leqq n \leqq |C|}$

Modalities:

Modalities are ternary relations understood as each agent's accessibility relation on situations. More specifically, for a modality M∈{K,B,P,I},M(x,s′,s) is used for situation s′ being accessible from situation s to agent x according to M.

For modality M, agent x, situation s, and formula ϕ_z with free situation variable z, we use the shorthand notation:

$M(x, \phi_z , s) \equiv \forall s'{:}\sigma.\ M(x,s',s) \rightarrow \phi_z[z/s']$

For improving readability, free situation variables will be mention by symbol_ .

Observation Axioms

Accusing another player of being the werewolf:

$\forall x,y{:}\alpha.(claim_{wolf}(x,y,s) \wedge x \neq y) \rightarrow (wolf(x,s) \wedge$

$\neg B(x,wolf(y,\_),s) \wedge B(x,B(y,wolf(x,\_))) \vee \neg wolf(x,s) \wedge B(x,wolf(y,\_),s))$

Defending another player or oneself:

$\forall x,y{:}\alpha.claim_{\neg wolf}(x,y,s) \rightarrow (wolf(x,s)$

$\wedge B(x,\exists z{:}\alpha.z \neq x \wedge z \neq y \wedge B(z.wolf(y,\_),\_),s)$

$\wedge B(x,\exists z{:}\alpha.z \neq x \wedge z \neq y \wedge \neg B(z.wolf(y,\_),\_),s)$

$\wedge (x \neq y \vee \neg B(x,wolf(y,\_),s))) \vee (\neg wolf(x,s)$

$\wedge B(x,\exists z{:}\alpha.z \neq y \neg B(z,wolf(y,\_),s),s)$

$\wedge \neg B(x,wolf(y,\_),s)))$

Claiming the role of seer:

$\forall x{:}\alpha.claim_{seer}(x,x,s) \rightarrow (wolf(x,s) \vee seer(x,s))$

$\forall x,y{:}\alpha.(claim_{seer}(x,y,s) \wedge x \neq y) \rightarrow (\neg wolf(x,s)$

$\wedge B(x,seer(y,\_),s) \wedge \neg B(x,wolf(y,\_),s)$

$\wedge B(x,\exists z{:}\alpha.I(z,dead(y,\_),\_),s))$

Claiming that someone is not a seer:

$\forall x,y{:}a.claim_{\neg seer}(x,y,s) \rightarrow seer(x,s) \wedge B(x,\exists z{:}\alpha.B(z,seer(y,\_),s),s)$

Not only utterances can show the beliefs and intentions of the agent, action such as voting also can show intentions:

$voted(x,y,do(z,a,s)) \rightarrow I(x,dead(y,\_),s)$

Belief and Intention Axioms At first of the game, for each agent the goal of the game is to stay alive.

$P(x,\neg dead(x,\_),s)$

Game rounds end with agents voting to have someone executed, which also reflects in the agents' intentions:

$P(x,\vee_{y \in C, y \neq x} dead(y,\_),s)$

Persistent intentions are intentions and are known by all.

In a reciprocal manner, any agent will intend another agent be dead if she concluded the latter intends her death:

$K(x,I(y,dead(x,\_),s),s) \rightarrow I(x,dead(y,\_),s)$

If an agent knows who the werewolf is, unless she is herself a werewolf, she will intend the werewolf be dead:

$K(x,wolf(y,\_),s) \wedge \neg wolf(y,s) \rightarrow I(x,dead(y,\_),s)$

The werewolf will have a similar approach to the seer.

Conversely, knowing an agent's intention will also give some insight into his belief regarding who the werewolf is:

$\forall x,y{:}\alpha.I(x,dead(y,\_),s)\rightarrow(\neg wolf(x,s)\wedge B(x,wolf(y,\_),s))\vee(wolf(x,s)\wedge B(x,\neg wolf(y,\_),s))$

Using the Observation Model to Solve Werewolf Game

4.2 Using the Observation Model to Solve Werewolf Game

Here is an example that how to use the system to solve werewolf game. In this example, agent x is a seer, agent y is a villager and z is a werewolf.

At the beginning of the game, the agent only knows his own role. And the intention for each agent is to be alive until the game ends.

Game start:S_0

For agent x : $K(x,seer(x,\_),)S_0$; $P(x,\neg dead(x,\_),S_0)$

Agent x knows that he is a seer and his persistent intention is to stay alive until the game ends.

For agent y : $K(y,villager(y,\_),)S_0$; $P(y,\neg dead(y,\_),S_0)$

Agent y knows that he is a villager and his persistent intention is to stay alive until the game ends.

For agent y : $K(z,wolf(z,\_),)S_0$; $P(z,\neg dead(z,\_),S_0)$

Agent y knows that he is a villager and his persistent intention is to stay alive until the game ends.

After some turns agent x knows that agent z is a werewolf (by divination).

$claim_{seer}(x,x,s)$

$claim_{wolf}(x,z,s)$

Agent x is coming out. And agent x claims that agent z is a werewolf.

Then the knowledge and intentions of agents have changed.

For agent x:

$K(x,wolf(z,\_),s)\wedge\neg wolf(x,s)\rightarrow I(x,dead(z,\_),s)$

While x knows that agent z is a werewolf and agent x is not a werewolf. The intention of agent x will change to be that agent x wants agent z dies.

For agent y:

$k(y, seer(x,\_),s)\vee\neg k(y,seer(x,\_),s)$

While x is coming out, agent y may trust him or not .

$K(y,wolf(z,\_),s)\vee K(y,wolf(x,\_),s)$

While y trust agent x, he will think that agent z is a werewolf. While y do not trust agent x, he will think that agent x is a werewolf.

$I(y,dead(z,\_),s) \lor I(y,dead(x,\_),s)$

The intentions of agent y will have two possibilities.

For agent z:

$k(z,seer(x,\_),s) \land wolf(z,s) \land K(z,I(x,dead(z,\_),s),s) \rightarrow I(z,dead(x,\_),s)$ Agent z knows that agent x is a seer and agent z is a werewolf and agent z knows that agent x wants he dies,so the intention of agent z will become that agent z wants agent x to die.

Voting time

$(poss(x,vote(z),s) \land poss(y,vote(z),s) \land poss(z,vote(x),s)) \lor (poss(x,vote(z),s) \land poss(y,vote(x),s) \land poss($

$\rightarrow dead(z,\_) \lor dead(x,\_)$

The result of the voting will be killing agent z or killing agent x. By using this system, agent A could predict the knowledge changes and the intention changes for all agents after his coming out and he could predict the result of the voting.

Chapter 5

Conclusion

1. Werewolf game is party game at first. That means the information you get from other players is not only the conversation. In werewolf game of the real world, the emotions, the movements(body languages), the sound when you close your eyes and even the tones of voices, these are essential information to win the game. In werewolf game online, the only valid data is the conversations. I tried to play the werewolf online several times and I found that it is tough because of the lack of information. This is also the big problem in developing of werewolf AI. Some researchers used the data of werewolf game played by human beings to training the AI. The problem is that analyzing a large number of data of natural languages is almost impossible. And the record of human being game is not enough to train AI.

2. Modal logic is considered that suitable to solve the werewolf game. But even you can describe the werewolf game by possible world completely. The uncertainty of this game and the less information let that making a decision is too complicated. While the werewolf agent model based on BDI logic also facing this problem. There still need an algorithm to resist the uncertainty of the werewolf game.

3. Mental spaces theory is an excellent way to analyze some difficult natural language. We can use mental spaces to describe the thinking of the agents. And it also could be dynamic to represent the changing of the thought of agent. But it is difficult to reason the real mind of the agent.

4. The Observation Model provides a new view of the werewolf game. By using this model, the agent could predict the effect of his act. In my example, the seer predicts the result of voting after his coming out. While the agent could predict the result of every act he doses, he could choose the most benefit act to reach the intention of the games. This is allowed the agent to persuade other agents and win the game.

5. In the past studies , researchers absorbed in finding the werewolves to prove

the win rate. But here is a new idea that not only exposes the werewolf but also persuades other agents to achieve success. The Observation Model concentrates on the observe so it can not describe the werewolf game complete. In future work, we could extend the model entirely and build an agent that good to persuade others.

Bibliography

[1] Fujio Toriumi, Daisuke Katamata, and Hiroshi Osawa. *Artificial intelligence to cheat, find out, persuade people's werewolf AI*, Mori Kita Publishing, (2016)

[2] Philippe des Pallières and Hervé Marly. *The Werewolves of Millers Hollow*, Asmodee, (2001)

[3] http://aiwolf.org/

[4] Nide Naoyuki and Shiro Takata. *Tracing Werewolf Game by Using Extended BDI Model*, (2016)

[5] Yuya Hirata, Michimasa Inaba and Kenichi Takahashi. *Werewolf Game Modeling Using Action Probabilities Based on Play Log Analysis*, (2016)

[6] Tianhe Wang and Tomoyuki Kaneko. *Application of Deep Reinforcement Learning in Werewolf Game Agents* (2018)

[7] Yu Kobayash, Hirotaka Osawa and Michimasa Inaba . *Development of werewolf match system for human players mediated with lifelike agents*, (2014)

[8] Daisuke Katagami, Shono Takaku and Michimasa Inaba. *Investigation of the effects of nonverbal information on werewolf*, (2014)

[9] Tojo Satoshi. *Logic of language, knowledge and belief*, Ohm Publishing, (2006)

[10] Hans van Ditmarsch, Wiebe vab der Hoek and Barteld Kooi. *Dynamic Epostemic Logic*, Springer, (2008)

[11] Gilles Fauconnier. *MENTAL SPACES*, Cambridge University Press, (1998)

[12] Codruta Girlea, Eyal Amir and Roxana Girju. *Tracking Beliefs and Intentions in theWerewolf Game*, (2014)