*Proceedings of the LREC 2020 Workshop Games and Natural Language Processing*, pages 85–93
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

# Automatic Annotation of Werewolf Game Corpus with Players Revealing Oneselves as Seer/Medium and Divination/Medium Results

**Youchao Lin, Miho Kasamatsu, Tengyang Chen, Takuya Fujita, Huaijin Deng, Takehito Utsuro**
Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan

**Abstract**

While playing the communication game "Are You a Werewolf", a player always guesses other players' roles through discussions, based on his own role and other players' crucial utterances. The underlying goal of this paper is to construct an agent that can analyze the participating players' utterances and play the werewolf game as if it is a human. For a step of this underlying goal, this paper studies how to accumulate werewolf game log data annotated with identification of players revealing oneselves as seer/medium, the acts of the divination and the medium and declaring the results of the divination and the medium. In this paper, we divide the whole task into four sub tasks and apply CNN/SVM classifiers to each sub task and evaluate their performance.

**Keywords:** werewolf game, annotation, CNN, players revealing oneselves as seer/medium

## 1. Introduction

Werewolf is a party game created in the USSR in 1986. It models a conflict between an informed minority, the werewolf, and an uninformed majority, the villager. The werewolf game has been popular in many countries including Japan. In Japan, particularly, not only the game itself, but several other activities such as "Werewolf TLPT" (Werewolf: the live playing theater) [1], a improvisation where the actors and actresses play the werewolf game, and a TV variety show[2]where comedians, actors, and actresses play the werewolf game.

In the research community of artificial intelligence, it has been well known that the werewolf game is one of games with imperfect information where certain information are hidden from some players. This situation is quite contrary to games with perfect information such as chess, shogi, and go, where it is known that computer programs won a human champion[3]. In the Japanese research community of artificial intelligence, the werewolf game has been considered to be used as one of standard problems to evaluate the performance of general artificial intelligence since 2014 (Shinoda et al., 2014). Also, research activities aiming at developing a computer agent program which participates in the werewolf game has started and the first competition of the AIWolf (artificial intelligence based werewolf)[4] was held in August 2015 (Toriumi et al., 2014).

However, in those previous studies aiming at developing a computer agent program which participates in the werewolf game, research issues that are closely related to natural language processing and knowledge processing research have not been studied extensively. Those higher level research issues should include, e.g., understanding natural language conversations among the participating players, inferring each player's roles considering the contents of their conversations, and deciding the player to be attacked or executed based on high level inference.

Considering the underlying goal of constructing an agent that can analyze the participating players' utterances and play the werewolf game as if it is a human, as the first step, this paper studies how to accumulate werewolf game log data annotated with identification of players revealing oneselves as seer/medium, the acts of the divination and the medium and declaring the results of the divination and the medium. In this paper, we divide the whole task into four sub tasks and apply CNN/SVM classifiers to each sub task and evaluate their performance.

## 2. Werewolf Game

In the werewolf game, each player is given a role and all the players are divided into one of the werewolf side and the villager side. Then, players of the both sides aim at winning the game. The werewolf side attacks one player of the villager side per day, while the villager side tries to execute one werewolf per day through arguments and votes. The players on the villager side do not know each player's identity of being a werewolf or a human, while those on the werewolf side know those identifies. The werewolf side tries to make the players on the villager side vote themselves to be executed through misleading arguments by providing false information. Table 1 shows a typical case the list of roles of the both sides of the werewolf game with 15 players. Among those roles, the role of the possessed is on the werewolf side and the possessed wins when the werewolf side wins, although the seer divines the possessed to be a human, the medium declares the possessed to be a human as the result of the act of the medium, and the possessed is counted as a human when one survives.

Table 2 and Table 3 also list the rules and common sense of the werewolf game. The players are usually requested to follow those rules listed in Table 2, while they are just assumed to follow those common sense listed in Table 3. Those common sense are considered to be a kind of conventional strategies that are recommended to adopt so as to

[1] `http://7th-castle.com/jinrou/index.php` (in Japanese)

[2] `http://www.fujitv.co.jp/jinroh/index.html` (in Japanese)

[3] `http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/`,
`http://www.shogi.or.jp/kisen/denou/` (in Japanese),
`https://www.deepmind.com/alpha-go.html`

[4] `http://cedec.cesa.or.jp/2015/session/AC/7649.html` (in Japanese)

| side | player type when counting the survivors | role | description | # of players |
|---|---|---|---|---|
| villager | human | villager | A human who does not have any special skill. | 8 |
| | | seer | A human who belongs to the villager side. Every night, the seer can choose one player and learn whether the player is ”werewolf” or ”human”. Learning the result, the seer can tell it to other players. | 1 |
| | | medium | A human who belongs to the villager side. The medium can learn whether the player who was voted to be executed on the previous day is ”werewolf” or ”human”. Learning the result, the medium can tell it to other players. | 1 |
| | | bodyguard | A human who belongs to the villager side. Every night the bodyguard can choose one player except the bodyguard oneself to defend so that the chosen player can avoid being attacked by the werewolves. However, the bodyguard can not learn whether the player chosen to be defended was actually attacked or not. | 1 |
| werewolf | | possessed | A human who belongs to the werewolf side. The possessed wins when the the werewolf side wins. However, the possessed and werewolves do not know each others' roles. | 1 |
| | werewolf | werewolf | Every night the werewolves choose one player on the villager side to be killed. The werewolves know each others' role and can communicate through a channel that are available only to the werewolves. | 3 |
| total | — | | | 15 |

Table 1: Roles in the Werewolf Game (for 15 players)

| | |
|---|---|
| 1 | The number of the players for each of the roles of the seer, the medium, the bodyguard, and the possessed is one. |
| 2 | The werewolves know each others role. |
| 3 | The werewolves can not attack themselves. |
| 4 | When the number of the werewolves is larger than that of humans, the werewolf side wins. |
| 5 | When all the werewolves are executed, the villager side wins. |

Table 2: Rules of the Werewolf Game

| | |
|---|---|
| 1 | The content of the utterances by the villagers, the seer, the medium, and the bodyguard do not conflict with the truth. |
| 2 | The seer / the medium reveal themselves as a seer / a medium. |
| 3 | The content of the utterances by the werewolves and the possessed may conflict with the truth. |

Table 3: Common Sense of the Werewolf Game

| task | class | | training | evaluation |
|---|---|---|---|---|
| task 1 | revealing oneself as a seer | | 881 | 178 |
| | revealing oneself as a medium | | 259 | 93 |
| | revealing oneself as neither a seer nor a medium | | 1,336 | 778 |
| task 2 | $X$ | ($X$ = “utterance declaring the results of divination / medium”) | 3,206 | 700 |
| | not $X$ | | 3,206 | 12,324 |
| task 3-1 | Dieter | | 145 | 41 |
| | Peter | | 195 | 33 |
| | Clara | | 165 | 35 |
| | Erna | | 134 | 40 |
| | Otto | | 183 | 35 |
| | Liesa | | 193 | 44 |
| | Nicolas | | 210 | 49 |
| | Katharina | | 218 | 31 |
| | Jacob | | 161 | 26 |
| | Walter | | 120 | 30 |
| | Fridel | | 202 | 38 |
| | Thomas | | 133 | 27 |
| | Albin | | 163 | 44 |
| | Simon | | 172 | 41 |
| | Pamela | | 185 | 40 |
| | Simson | | 176 | 44 |
| | Joachim | | 210 | 44 |
| | Moritz | | 152 | 29 |
| | Regina | | 89 | 23 |
| task 3-2 | human | | 2,398 | 500 |
| | werewolf | | 808 | 200 |

Table 5: # of Training and Evaluation Examples for Tasks 1, 2, 3-1 and 3-2

raise the winning rates of both the villagers' and the werewolves' sides.

## 3. Werewolf Game Log

In this paper, as the werewolf game log data, we use that of werewolf BBS[5], which is a werewolf game site on the Internet, where the participating players communicate with each other with a character-based text input communication channel. This werewolf game site keeps the record of the text data of the previous werewolf game log and makes them publicly available.

## 4. Werewolf Game Corpus Annotation Tasks

Table 4 overviews the werewolf game corpus annotation tasks we study in this paper. In this paper, we apply supervised classifier learning techniques to those tasks, and

[5] `http://www.wolfg.x0.com/` (in Japanese)

| task | task description |
|---|---|
| task 1 | Identifying Players Revealing Oneselves as Seer/Medium |
| | input: each player and his/her utterances of the first 3 days<br>output: one of the classes of task 1 ( $\in$ { (i) revealing oneself as a seer, (ii) revealing oneself as a medium, (iii) revealing oneself as neither a seer nor a medium} ) |
| task 2 | Identifying Utterances declaring the Results of Divination / Medium |
| | input: each utterance on the 2nd day or after, of the players who are judged as “revealing oneself as a seer / a medium”<br>output: one of the classes of task 2 ( $\in \{X, \text{not } X\}$, $X =$ “utterance declaring the results of divination / medium”} ) |
| task 3-1 | Identifying the Names of the Players whose Roles are Identified by the Act of Divination / Medium |
| | input: each utterance on the 2nd day or after, of the players who are judged as “revealing oneself as a seer / a medium”<br>output: one of the classes of task 3-1 (names of the 19 players listed in Table 5) |
| task 3-2 | Identifying Results of Divination / Medium |
| | input: each utterance on the 2nd day or after, of the players who are judged as “revealing oneself as a seer / a medium”<br>output: one of the classes of task 3-2 ( $\in$ { human, werewolf } ) |

Table 4: Overview of the Werewolf Game Corpus Annotation Tasks

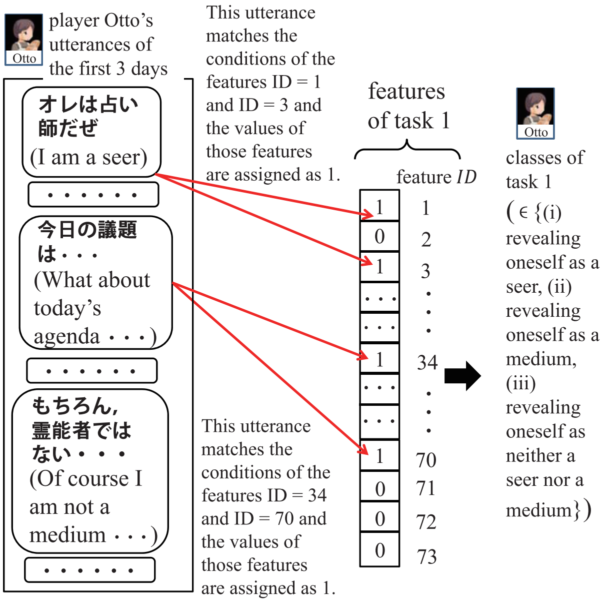

Otto
player Otto’s utterances of the first 3 days
オレは占い師だぜ
(I am a seer)
• • • • • •
今日の議題は・・・
(What about today’s agenda • • •)
• • • • • •
もちろん，霊能者ではない・・・
(Of course I am not a medium • • •)
• • • • • •
This utterance matches the conditions of the features ID = 1 and ID = 3 and the values of those features are assigned as 1.
This utterance matches the conditions of the features ID = 34 and ID = 70 and the values of those features are assigned as 1.
features of task 1
feature ID
1
0
1
• • •
• • •
1
• • •
• • •
1
0
0
0
1
2
3
34
70
71
72
73
Otto
classes of task 1
(∈ {(i) revealing oneself as a seer, (ii) revealing oneself as a medium, (iii) revealing oneself as neither a seer nor a medium})

Figure 1: Feature Representation of Task 1 (when manually crafted 73 rules are employed)

Table 5 lists the numbers of training and evaluation examples for each of the four tasks studied in this paper. The following sections introduce each of those four tasks. In the framework of applying classifier learning techniques, where we apply CNN and SVM, we employ manually crafted rules as well as character level text embeddings when designing feature representations of each of those four tasks. Rough idea of the feature representations of those four tasks when manually crafted rules are employed is illustrated in Figure 1 and Figure 2.

### 4.1. Task 1: Identifying Players Revealing Oneselves as Seer/Medium

The most important information which werewolf game players keep tracking throughout the whole werewolf game, and especially, in the early stage of the game, is that of players who reveal oneselves as a seer / a medium. This is obviously because the players of the roles of the seer / medium provide other players with true information on human / werewolf roles of other players. Thus, among the four tasks, the most important task 1 is that of identifying players who reveal oneselves as a seer / a medium. The input to the task 1 is one of all the 15 players and his/her utterances on the first three days[6], and then as the output of the task 1, the player given to the task 1 is judged as one of the three classes: (i) revealing oneself as a seer, (ii) revealing oneself as a medium, and (iii) revealing oneself as neither a seer nor a medium. Table 6 shows four examples of the task 1, for each of which, the name of the player given as the input, the role of the player, and the date of the utterance in which he/she indicates exactly that he/she is or is not a seer/medium are shown[7].

As illustrated in Figure 1, in the framework of applying classifier learning techniques such as CNN and SVM, we employ manually crafted rules as well as character level text embeddings when designing feature representations of task 1. In the case of task 1, we used 73 manually crafted rules in total, each of which is combined with the quoting notation that is commonly used in the werewolf BBS. Out of the total 73 rules, 14 are for matching Japanese expressions for revealing oneself as a seer, 7 for matching Japanese expressions for revealing oneself as a medium, and 15 for matching Japanese expressions for revealing oneself as not a seer nor a medium, where these in total amount to 36 rules. Another 36 rules are designed to examine the temporal order of the utterances that are matched to

[6] The task 1 considers each player’s utterances for only first three days, but not for later days. This is because, in werewolf BBS, generally, the seer reveals oneself by the end of the second day and the medium does so by the end of the third day.

[7] In those utterances of werewolf BBS, as shown in the underlined part of each utterance of Table 6, players use notation of quoting the part where he/she exactly indicates that he/she is or is not a seer/medium.

| | | | | |
|---|---|---|---|---|
| name of the player who utters, the role of the player, the date of the utterance in which he/she indicates exactly that he/she is or is not a seer/medium | Joachim, seer, day 1 | Liesa, possessed, day 1 | Pamela, werewolf, day 1 | Fridel, villager, day 1 |
| utterance | おれおれおれだよ占い師··· (*Here I AM a seer* ···) | ··· あたしは占い師でも霊能者でもないじょー。··· (··· *I am neither a seer nor a medium* ···) | ··· 霊能者ＣＯ、霊見えます··· (··· *medium CO, I can tell the role of the executed* ···) | ··· 私は占い師や霊能者ではありません ··· (··· *I am neither a seer nor a medium* ···) |
| class of task 1 ( ∈ { (i) revealing oneself as a seer, (ii) revealing oneself as a medium, (iii) revealing oneself as neither a seer nor a medium} ) | revealing oneself as a seer | revealing oneself as neither a seer nor a medium | revealing oneself as a medium | revealing oneself as neither a seer nor a medium |

Table 6: Examples of Reference Dataset of Task 1 (Underlined part of the utterance is quoted by the player who utters, indicating exactly in that part that he/she is or is not a seer/medium) (CO: abbreviation of "coming out")

one of those 36 rules. More specifically, each of another 36 rules judges whether the utterance matching the rule is around the end of the third day compared to remaining other 35 rules. Finally, the last rule is designed for judging whether at least one of the aforementioned 36 rules matches any of the utterances on the first three days, whose value is assigned as 1 when none of the 36 rules matches any of the utterances. The results of matching those 73 rules are represented as feature value assignments as shown in Figure 1.

### 4.2. Task 2, Task 3-1, Task 3-2: Identifying Utterances declaring the Results of Divination / Medium, the Names of the Players whose Roles are Identified, and Results of Divination / Medium

Once a player is identified as revealing oneself as a seer / a medium as the result of the task 1, then, each of his/her utterances on the 2nd day or after is given to the following task 2, task 3-1, and 3-2 as the input. In those following tasks, task 2 first identifies the utterance $U$ which declares the results of divination / medium, task 3-1 then identifies the name of the player whose role is identified by the act of divination / medium in the utterance $U$ (identified in task 2 ), and task 3-2 finally identifies the result of the act of divination / medium in the utterance $U$ (identified in task 2 ). The input to those three tasks task 2, task 3-1, and task 3-2 is each utterance (on the 2nd day or after) of the player who is identified as revealing oneself as a seer / a medium as the result of the task 1. The output of task 2 is one of the two classes: (a) $X$, and (b) not $X$ ($X$ = "utterance declaring the results of divination / medium"). The output of task 3-1 is one of the 19 player names listed in Table 5[8]. The output of task 3-2 is one of the two classes: (a) human, (b) and werewolf. Table 7 shows three examples of the task 2, task 3-1, and task 3-2. For those three tasks, the name of the player who utters, the role of the player, and the date of the utterance given to those three tasks as the input are shown.

As illustrated in Figure 2, in the framework of applying classifier learning techniques such as CNN and SVM, we again employ manually crafted rules as well as character level text embeddings when designing feature representations of task 2, task 3-1, and task 3-2.

In task 2, we used 7 manually crafted rules in total. Roughly speaking, those 7 rules judge whether the utterance includes (i) the quoting notation that is commonly used in the werewolf BBS, (ii) player names and the role names such as "human" and "werewolf", (iii) typical Japanese vocabularies representing "acknowledgment" and "identification", (iv) the mixture of (i) and (ii), and (v) the mixture of (i) and (iii). Furthermore, one of those 7 rules judges whether the utterance does not match any of the (i) to (v) above. And, the final one out of those 7 rules represents the order of the utterance among other utterances within the same day. The results of matching those 7 rules are represented as feature value assignments as shown in Figure 2.

Similarly, in task 3-1, we used four rules for each of the 19 player names. Roughly speaking, those four rules judge whether the player is alive or dead, and judge whether the utterance includes the real name or nickname of the player, the quoting notation that is commonly used in the werewolf BBS, and the role names such as "human" and "werewolf". Overall, in task 3-1, we used 77 rules in total (= 19 player names × 4 rules + one rule for detecting that none of the 19 player names matches the input utterance).

In task 3-2, on the other hand, we used 7 rules, out of which 6 are for matching typical Japanese vocabularies representing the roles of "human" and "werewolf", while the final

[8] As shown in Table 1, the werewolf game log data we used in this paper is with 15 players, while the number of the player name candidates is 19 in the werewolf BBS.

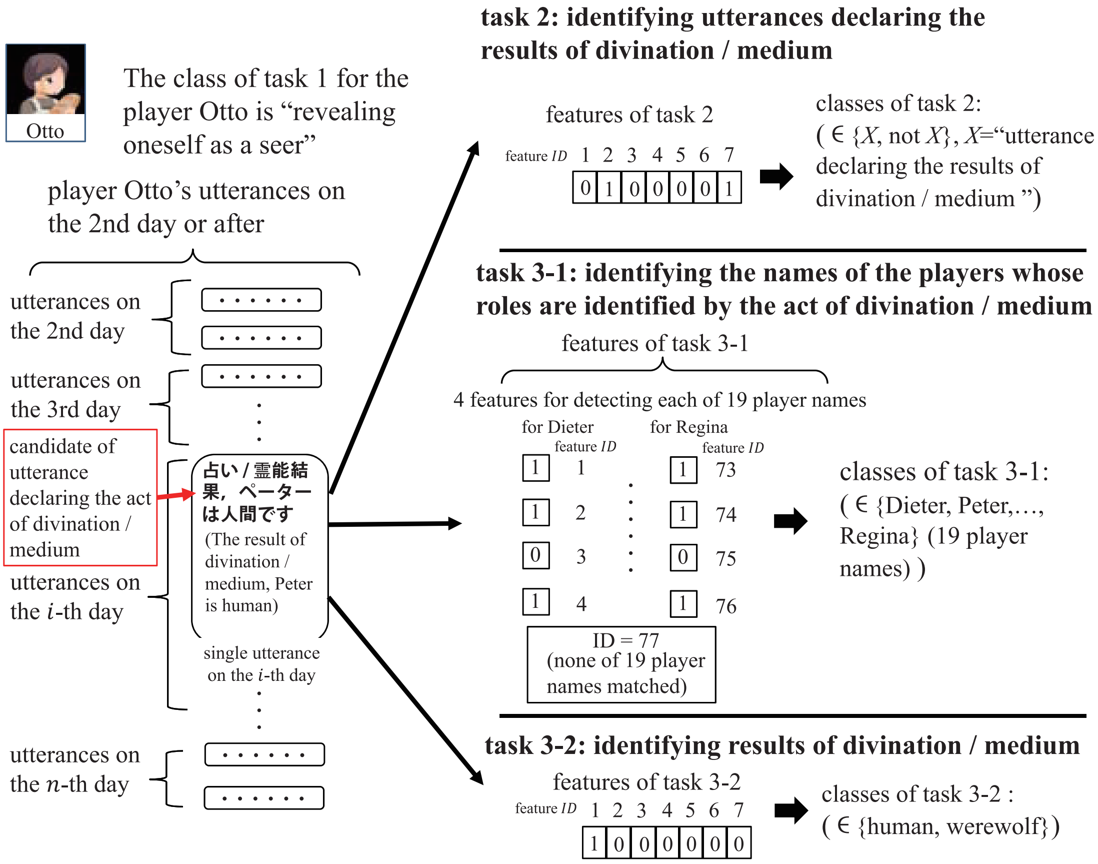

task 2: identifying utterances declaring the results of divination / medium
Otto
The class of task 1 for the player Otto is "revealing oneself as a seer"
player Otto's utterances on the 2nd day or after
utterances on the 2nd day
utterances on the 3rd day
candidate of utterance declaring the act of divination / medium
utterances on the i-th day
占い / 霊能結果，ペーターは人間です
(The result of divination / medium, Peter is human)
single utterance on the i-th day
utterances on the n-th day
features of task 2
feature ID 1 2 3 4 5 6 7
0 1 0 0 0 0 1
classes of task 2: ( ∈ {X, not X}, X="utterance declaring the results of divination / medium ")
task 3-1: identifying the names of the players whose roles are identified by the act of divination / medium
features of task 3-1
4 features for detecting each of 19 player names
for Dieter
for Regina
feature ID
1 1
1 2
0 3
1 4
1 73
1 74
0 75
1 76
ID = 77 (none of 19 player names matched)
classes of task 3-1: ( ∈ {Dieter, Peter,…, Regina} (19 player names) )
task 3-2: identifying results of divination / medium
features of task 3-2
feature ID 1 2 3 4 5 6 7
1 0 0 0 0 0 0
classes of task 3-2 : ( ∈ {human, werewolf})

Figure 2: Feature Representations of Task 2, Task 3-1, and Task 3-2 (when manually crafted 7 (task 2) / 77 (task 3-1) / 7 (task 3-2) rules are employed)

| name of the player who utters, the role of the player, the date of the utterance given to the tasks 2, 3-1, and 3-2 as the input | Joachim, seer, day 3 | Fridel, villager, day 4 | Pamela, werewolf, day 4 |
|---|---|---|---|
| utterance | ··· エルナさんは白でした ···<br>(··· *Erna is white* ···) | ··· ヨアヒムさんが狼だったらどうなるか ···<br>(··· what will happen if Joachim is a werewolf ···) | ··· 羊狼だと違和感 ···<br>(··· it is strange if the role of the player sheep is werewolf ···) |
| class of task 2 $\big( \in \{X, \text{not } X\}$, $X =$ "utterance declaring the results of divination / medium"$\} \big)$ | $X$ ($X$ = "utterance declaring the results of divination / medium") | Since the class of task 1 for the player Fridel is "revealing oneself as neither a seer nor a medium", tasks 2, 3-1 and 3-2 are not applied to any of her utterances. | not $X$ ($X$ = "utterance declaring the results of divination / medium") |
| class of task 3-1 (names of 19 players listed in Table 5) | Erna | | Since the class of task 2 for this utterance is "not $X$", tasks 3-1 and 3-2 are not applied to this utterance. |
| class of task 3-2 $\big( \in \{$ human, werewolf $\} \big)$ | human | | |

Table 7: Examples of Reference Dataset of Task 2, Task 3-1 and Task 3-2 (Underlined part of the utterance is quoted by the player who utters, indicating exactly in that part that he/she declares the results of divination / medium, or just his/her guesses.)

one detects that the utterance does not include any of those vocabularies.

## 5. Classifier

As the classifier, this paper applies CNN and SVM to all of the four tasks: task 1, task 2, task 3-1, and task 3-2.
When applying CNN, the implementation platform within Pytorch[9] is employed, where the following three types of feature representations are evaluated: (i) manually crafted rules are used[10] as shown in Figure 1 and Figure 2, (ii) char-

[9] `https://pytorch.org/`

[10] We use one convolution layer (40 channels and the filter size as 5), one max-pooling layer (the filter size as 2) and two fully connected layers to build the network. We use ReLU activation function, mini-batch size of 10, learning rate: 0.001, number of epochs: 50 epochs. The cross-entropy loss between training labels and predicted ones is minimized and optimization is performed using SGD.

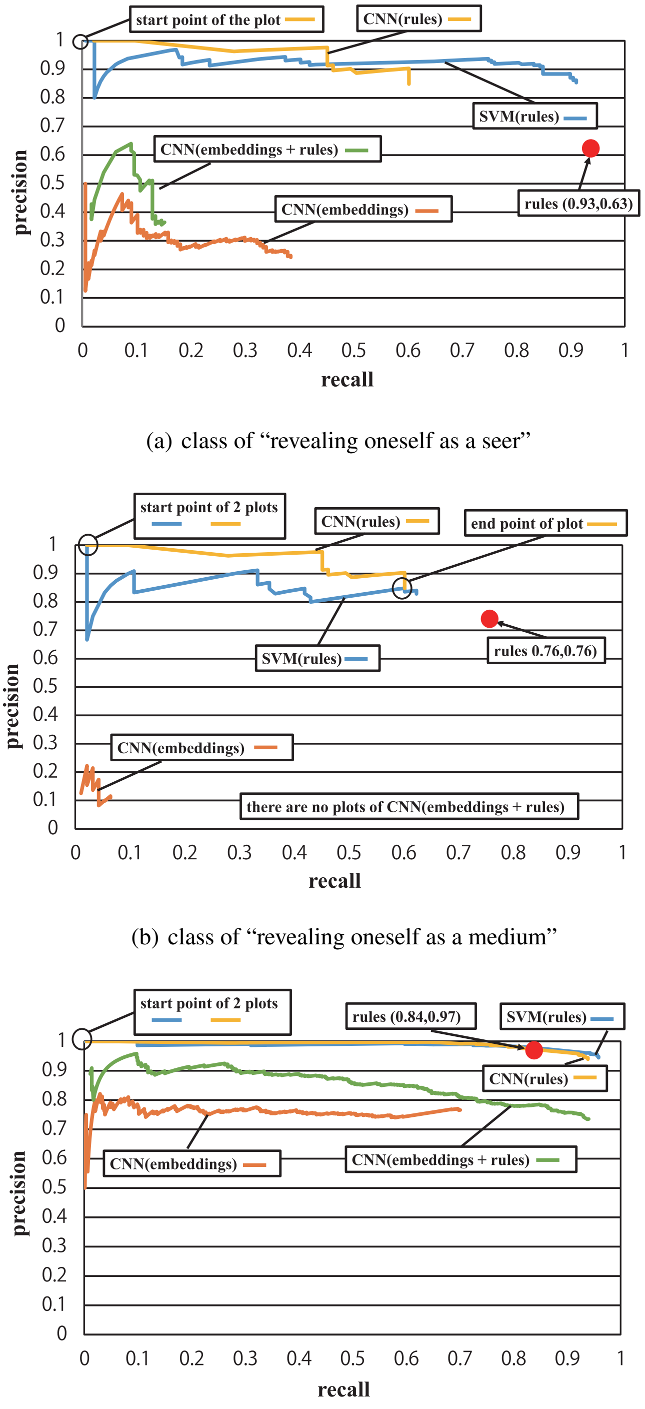
acter level text embeddings[11] are used, (iii) both (i) and (ii) are used together. As character level text embeddings[12], we used the one trained with Wikipedia Japanese text[13] by FastText (Bojanowski et al., 2017)[14], where the character level text embeddings are kept static during the procedure of training the CNN parameters.

When applying SVM[15], the feature representations shown in Figure 1 and Figure 2 are directly used, where 2nd degree polynomial is employed and the hyper parameters $C$ and $\gamma$ are grid-searched.

## 6. Evaluation

The CNN and SVM models described in the previous section are evaluated with the training and evaluation examples whose numbers are as shown in Table 5. As shown in Figure 3 $\sim$ Figure 6, the evaluation results are presented as the recall-precision curves[16] for the evaluation examples of each class of the four tasks. For all the tasks, CNN models with the following three types of feature representations, as well as the SVM model are evaluated and plotted in the figures: (i) manually crafted rules are used, (ii) character level text embeddings are used, (iii) both (i) and (ii) are used together. In addition to those three CNN models and the SVM model, we also dot the recall and precision point when we evaluate the manually crafted rules as they are originally designed to judge the class of each task without incorporating into CNN/SVM.

For task 1, as shown in Figure 3, it is obvious that the CNN model with the feature representation obtained by manually crafted rules performs the best for the two classes: (a) revealing oneself as a seer, and (b) revealing oneself as a medium. For the class of (b) revealing oneself as a medium, 7 rules without incorporating into CNN/SVM achieved the highest recall. One of the reasons why the CNN models having feature representations with character level text embeddings ((ii) and (iii)) performed much worse than those by manually crafted rules only is that the number of utterances on the first three days is too large (up to 60 utterances) when character level text embeddings are incorporated into the CNN model[17].



(a) class of “revealing oneself as a seer”

(b) class of “revealing oneself as a medium”

(c) class of “revealing oneself as neither a seer nor a medium”

Figure 3: Evaluation Result of Task 1

For task 2, overall, the CNN model with feature representations by manually crafted rules as well as the SVM model

[11] For both (ii) and (iii), the fundamental formalization of CNN is based on that of Kim (2014), where one convolution layer (one channel and the filter size as 3, 4, 5), one max-pooling layer (the filter size as 2) and one fully connected layer are used to build the network. We use ReLU activation function, mini-batch size of 10, learning rate: from 0.001 to 0.0001, number of epochs: 100 epochs. The cross-entropy loss between training labels and predicted ones is minimized and optimization is performed using ADAM optimizer (Kingma and Ba, 2015).

[12] We compare word level and character level text embeddings, where the character level embeddings outperformed the word level embeddings.

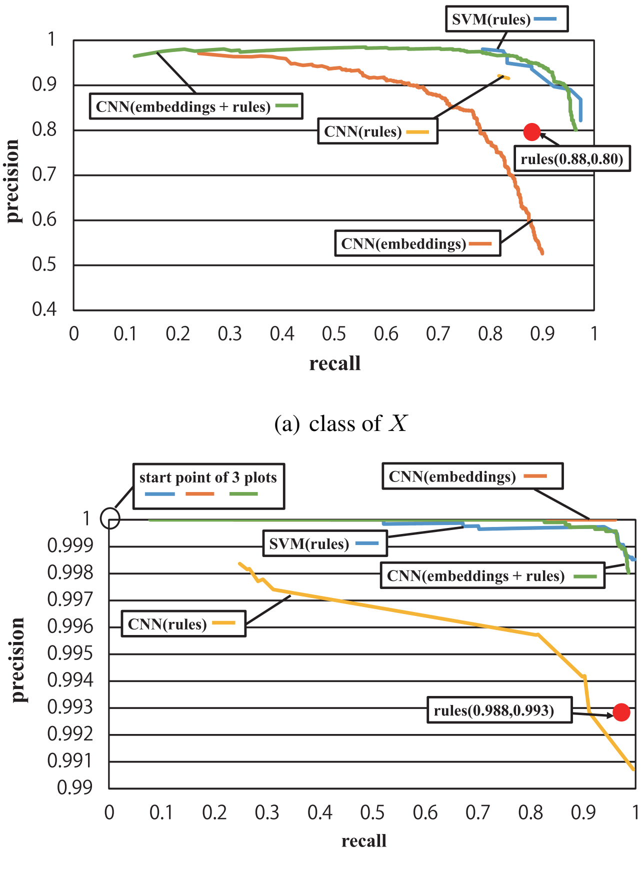
[13] We compared text embeddings trained with the Japanese text data of the 646 werewolf BBS game logs and the one trained with Wikipedia Japanese text, where the one trained with Wikipedia Japanese text outperformed the one trained with the werewolf BBS game logs.

[14] `https://fasttext.cc/docs/en/pretrained-vectors.html`

[15] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[16] The evaluation examples are sorted in descending order of the probability of the softmax function and then the recall-precision curve is plotted by changing the lower bound of the probability of the softmax function.

[17] The CNN model with the feature representation (iii) (both (i) and (ii) are used together) did not predict the class of (b) revealing oneself as a medium for any of the evaluation examples. This is mainly because the numbers of both the training and the evaluation examples are much smaller compared to other two classes.



(a) class of $X$

(b) class of not $X$

Figure 4: Evaluation Result of Task 2 ($X$ = "utterance declaring the results of divination / medium")

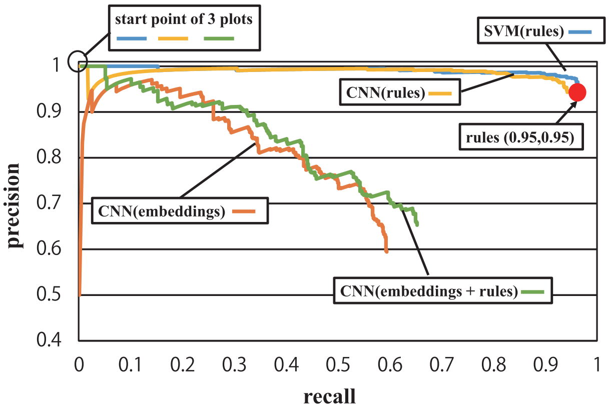
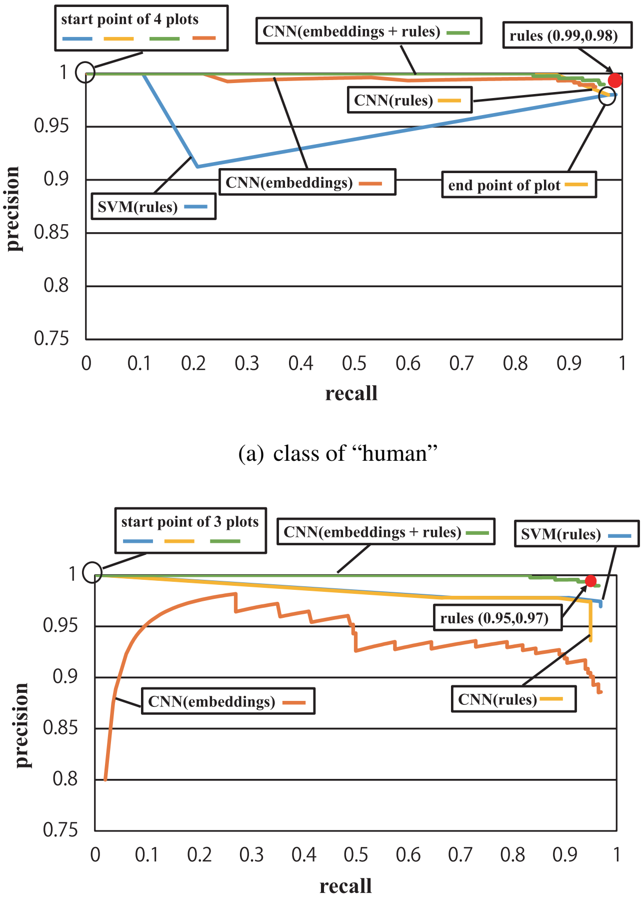


Figure 5: Evaluation Result of Task 3-1



(a) class of "human"

(b) class of "werewolf"

Figure 6: Evaluation Result of Task 3-2

performed the best. Roughly speaking, it can be pointed out that, for the CNN models, the performance improved by incorporating the feature representations (i) manually crafted rules are used, and (ii) character level text embeddings are used, all together into (iii).

For task 3-1, the CNN model with the feature representation obtained by manually crafted rules as well as the SVM model performed the best. Another finding here is that manually crafted rules without incorporating into CNN/SVM achieved almost the highest performance even compared to CNN/SVM. This is mainly because the number of the classes of task 3-1 is 19, which is quite large, and consequently the number of training examples for each of the 19 classes becomes relatively small, especially for the CNN model with the feature representation obtained by the character level text embeddings.

For task 3-2, overall, the CNN model with the feature representation obtained by manually crafted rules performed the best. Also, in task 3-2, manually crafted rules without incorporating into CNN/SVM achieved almost the highest performance even compared to CNN/SVM. This is mainly because, for this task, 6 rules for matching typical Japanese vocabularies representing the roles of "human" and "werewolf" play almost the most important role in this task. And, once the utterance matches one of the 6 rules, it can be stated that the result of the act of divination / medium can be easily obtained even without incorporating the feature representation into CNN/SVM.

## 7. Evaluation of Applying the Models of Task 1 to Task 3-2 Sequentially

When we actually annotate a werewolf game log text corpus with information that is closely related to state transitions the werewolf game by applying the models proposed in this paper, it is necessary to apply the models of individual tasks one by one sequentially. This section describes the procedure of such a situation and its evaluation results

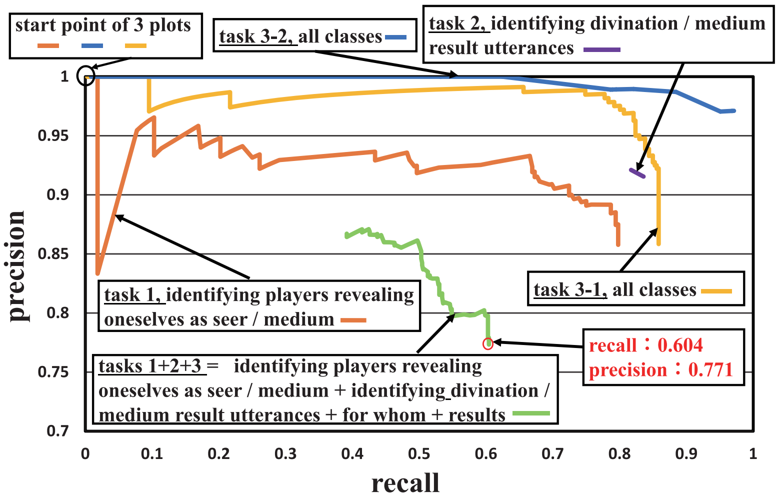


Figure 7: Evaluation Result of Applying the Models of Task 1 to Task 3-2 Sequentially

as in Figure 7[18].

In such a situation, we first apply the model of task 1 to each player (i.e., for each player, to the set of his/her utterances on the first three days), and identify players revealing oneselves as seer/medium (the evaluation result of this individual task is shown as the plot "task 1, identifying players revealing oneselves as seer/medium" in Figure 7), while we ignore players revealing oneselves as neither a seer nor a medium. Next, we apply the model of task 2 to each utterance (on the 2nd day or after) of those identified players, and identify utterances declaring the results of divination/medium (the evaluation result of this individual task is shown as the plot "task 2, identifying divination/medium result utterances" in Figure 7), while we ignore other utterances. Finally, we apply the models of task 3-1 and 3-2 to each utterance $U$ of those identified utterances and obtain the player name whose role is identified by the act of divination / medium in the utterance $U$ and his/her role as the result of the act of divination / medium (the evaluation results of these individual tasks are shown as the plots "task 3-1, all classes" and "task 3-2, all classes" in Figure 7).

When sequentially applying the models of those individual tasks one by one, the recall-precision curve for correctly identifying the outputs of all the four tasks is plotted lower (as shown as the plot "task 1+2+3 = identifying players revealing oneselves as seer/medium + identifying divination/medium result utterances + for whom + result" in Figure 7) compared to those evaluation results of individual tasks. This is obviously because the overall sequential evaluation results are obtained by multiplying each evaluation performance for all the four tasks. However, in this overall evaluation results, we achieved around 60∼70% recall/precision and the highest prevision as over 85% when restricting recall around 40%.

## 8. Related Work

Most previous work related to the werewolf game (and other similar games) studied issues regarding how to design the werewolf game agent which has the ability of joining natural language conversation of the werewolf games (Gillespie et al., 2016; Hirata et al., 2016; Nishizaki and Ozaki, 2016; Toriumi et al., 2016; Nide and Takata, 2017; Xiong et al., 2017; Kano et al., 2019; Nagayama et al., 2019; Sugawara, 2019; Tellols, 2019; Tsunoda and Kano, 2019). Issues studied in those previous work include tendencies in utterances of the executed or attacked players (Nishizaki and Ozaki, 2016) and analyzing the influence of the features such as the number of each player's utterances, number of the players revealing oneselves as seer/medium, etc., against the winning rate of the werewolf side (Nagayama et al., 2019). Among those previous work, the task studied in Sugawara (2019) is relatively similar to those studied in this paper. Sugawara (2019) applied embedding based technique to the task of classifying speech acts of utterances collected from the natural language text based werewolf game log, where their classification performance is much lower than the results we report in this paper. It is obvious from the results we report in this paper that speech act classification performance should improve by incorporating feature representations obtained by manually crafted rules in addition to those embedding based feature representations. This finding is one of the most important differences between this paper and Sugawara (2019).

## 9. Conclusion

This paper studied how to accumulate werewolf game log data annotated with identification of players revealing oneselves as seer/medium, the acts of the divination and the medium and declaring the results of the divination and the medium. In this paper, we divided the whole task into four sub tasks and applied CNN/SVM classifiers to each sub task, where we showed the effectiveness of the proposed CNN/SVM models.

[18] In the evaluation results of Figure 7, to all the tasks, the CNN models with the feature representation obtained by manually crafted rules are applied. Evaluation results for the individual tasks are those against the whole evaluation examples whose numbers are as shown in Table 5.

## 10. Bibliographical References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of ACL*, 5:135–146.

Gillespie, K., Floyd, M. W., Molineaux, M., Vattam, S. S., and Aha, D. W. (2016). Semantic classification of utterances in a language-driven game. In Tristan Cazenave, et al., editors, *Computer Games. CGW 2016, GIGA 2016*, volume 705 of *CCIS*, pages 116–129. Springer, Cham.

Hirata, Y., Inaba, M., Takahashi, K., Toriumi, F., Osawa, H., Katagami, D., and Shinoda, K. (2016). Werewolf game modeling using action probabilities based on play log analysis. In Aske Plaat, et al., editors, *Computers and Games*, volume 10068 of *LNCS*, pages 103–114. Springer, Cham.

Kano, Y., Aranha, C., Inaba, M., Toriumi, F., Osawa, H., Katagami, D., Otsuki, T., Tsunoda, I., Nagayama, S., Tellols, D., Sugawara, Y., and Nakata, Y. (2019). Overview of AIWolfDial 2019 shared task: Contest of automatic dialog agents to play the werewolf game through conversations. In *Proc. AIWolfDial*, pages 1–6.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proc. EMNLP*, pages 1746–1751.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *Proc. ICLR*.

Nagayama, S., Abe, J., Oya, K., Sakamoto, K., Shibuki, H., Mori, T., and Kando, N. (2019). Strategies for an autonomous agent playing the “ werewolf game ” as a stealth werewolf. In *Proc. AIWolfDial*, pages 20–24.

Nide, N. and Takata, S. (2017). Tracing werewolf game by using extended BDI model. *IEICE Transactions on Information and Systems*, E100-D(12):2888–2896.

Nishizaki, E. and Ozaki, T. (2016). Behavior analysis of executed and attacked players in werewolf game by ILP. In *Proc. 26th ILP*, pages 48–53.

Shinoda, T., Chokai, F., Katagami, D., Osawa, H., and Inaba, T. (2014). “Are you a Werewolf?” becomes a standard problem for general artificial intelligence. In *Proc. 28th Annual Conf. JSAI.* (in Japanese).

Sugawara, Y. (2019). Data augmentation based on distributed expressions in text classification tasks. In *Proc. AIWolfDial*, pages 7–10.

Tellols, D. (2019). Are talkative AI agents more likely to win the werewolf game? In *Proc. AIWolfDial*, pages 11–14.

Toriumi, F., Kajiwara, K., Osawa, H., Inaba, T., Katagami, D., and Shinoda, T. (2014). Development of AI wolf server. *Proc. 19th GPW*, pages 127–132. (in Japanese).

Toriumi, F., Osawa, H., Inaba, M., Katagami, D., Shinoda, K., and Matsubara, H. (2016). AI wolf contest — development of game AI using collective intelligence —. In Tristan Cazenave, et al., editors, *Computer Games. CGW 2016, GIGA 2016*, volume 705 of *CCIS*, pages 101–115. Springer, Cham.

Tsunoda, I. and Kano, Y. (2019). AI werewolf agent with reasoning using role patterns and heuristics. In *Proc. AIWolfDial*, pages 15–19.

Xiong, S., Li, W., Mao, X., and Iida, H. (2017). Mafia game setting research using game refinement measurement. In Adrian David Cheok, et al., editors, *Advances in Computer Entertainment Technology. ACE 2017*, volume 10714 of *LNCS*, pages 830–846. Springer, Cham.