The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

人狼知能大会第一回自然言語部門の開催

Natural Language Task in AI Werewolf Contest

狩野　芳伸[*1]
Yoshinobu Kano

稲葉　通将[*2]
Michimasa Inaba

[*1] 静岡大学
Shizuoka University

[*2] 広島市立大学
Hiroshima City University

We report the first Natural Language Task in the AI Werewolf Contest of the AI Werewolf project. In this task, participants are required to create an agent system that plays the conversation game “Mafia” in Japanese text. We show statistical and subjective evaluation results of automatic play within agents and with a human, discussing future possibility of this task.

1. はじめに

近年、対話システムの利用が注目を集めている。従来多く見られたタスク指向型対話システムに加え、いわゆる雑談的な対話を行う非タスク指向型対話システムも多くみられる。対話システムが必要とする要素はテキスト入出力に限ったとしても、語彙、構文から意味、論理まで、完成度を上げていくと最終的にほとんどあらゆる知的要素が必要となろう。

人狼ゲームは基本的に制約のない会話を通じてのみ行うゲームであり、対話システムに必要な要素を包含している。一方で人狼ゲームを題材にすることは、ゲームの勝敗という目的があるために状況を限定しつつ一歩一歩研究を進めうるという利点がある。本稿では、対話システムの発展に寄与する研究の切り口として、会話ゲーム「人狼」のプレイヤー自動化を試みる人狼知能大会における、自然言語を用いる部門の第一回大会開催とその結果について述べる。なお、全国大会に合わせて第二回のプレ大会開催を企画しており、結果を全国大会のオーガナイズドセッション「不完全情報ゲームと人狼知能」内で発表予定である。

2. 人狼ゲーム

2.1 人狼ゲームの設定

人狼ゲームは数名～十数人程度のプレイヤーで行うゲームである。各プレイヤーは村人陣営と人狼陣営に分かれてそれぞれの陣営の勝利を目指す。ゲーム開始前に所属陣営を決定するが、基本的に村人には他プレイヤーの所属陣営は公開されない。村人陣営の目的は人狼をすべて見つけ出し追放(ゲームから排除)することで、人狼陣営の目的は村人を食べ尽くす(ゲームから排除する)ことである。

人狼陣営のプレイヤーは排除されないよう、村人陣営の振りをすることになる。プレイヤーは基本的に会話のみを通じてお互いの正体を探るが、何の情報も無い状態では人狼側が有利になりすぎるため、村人側に有利な特殊能力を持ったいくつかの「役職」が主に村人側に用意されている。ゲームの設定により異なるが、一日一人だけ指定したプレイヤーが「人狼かどうか」を知ることができる(ただし本人以外にはそれが真かどうか、占ったかどうかも知らされない)「占い師」、人狼陣営だが自身は人狼が誰だかを知らず、能力も持たないため、場を混乱させることによって人狼陣営に貢献することの多い「裏切者」などがある。

連絡先:狩野　芳伸、静岡大学情報学部行動情報学科、静岡県浜松市中区城北 3-5-1、kano@inf.shizuoka.ac.jp

2.2 人狼ゲームの流れ

ゲームは「一日」を基準とするターンを繰り返す。一日の前半を「昼」と呼び、プレイヤーは決められた時間、自由に会話を行う。後半を「夜」と呼び、ゲームから排除するプレイヤーを投票によって決定する。村人陣営は人狼陣営のプレイヤーを排除しようとし、人狼陣営のプレイヤーは村人陣営のプレイヤーを排除しようとするだろうが、表面上はあくまで「人狼と思われる人」を排除するための投票になる。最大得票を得たプレイヤーはゲームから排除(「追放」)される。

次に、全プレイヤーが顔を伏せた状態で、司会者が様々な役職のプレイヤーに声をかけ、声をかけられたプレイヤーはジェスチャーなどで他のプレイヤーにわからないよう、特殊能力に従った行動をとる。 例えば、人狼はゲームから排除するプレイヤーを一人決定する(「襲撃」)。

ここまでが「一日」のターンで、これをどちらかの陣営が勝利条件を満たすまで繰り返す。プレイヤーの勝敗は所属陣営の勝敗できまるので、途中で排除されたかどうかはプレイヤーの勝敗に直接関係がない。村人陣営の勝利条件は「すべての人狼を排除する」こと、人狼陣営の勝利条件は「人狼以外のプレイヤーの数を人狼の数と同数またはそれ未満にする」ことである。

人狼ゲームは元々ロシアで遊ばれていたゲームでマフィアと呼ばれていた。それがアメリカで商品化され、日本でも「汝は人狼なりや」「タブラの狼」などいくつかのゲームセットが発売されている。現在では、インターネット越しに掲示板形式で人狼ゲームを長期にわたって行う BBS 人狼、短期間で行う短期人狼などがある。人狼をプレイする人口は現在増え続けているが、上の年代になるにつれて知名度は減少しており、若者世代に人気のゲームであると言えよう。

3. 人狼知能プロジェクトと人狼知能大会

3.1 人狼知能プロジェクト

人狼知能プロジェクト[1][2][3][4]では、人狼知能プロトコルと対戦用の人狼知能サーバの情報を公開することで、多く研究者や開発者がエージェントの作成に参加できる環境を整えてきた。

将棋やチェスなど盤面上にすべての情報が開示される完全情報ゲームとは異なり、人狼は不完全情報ゲームである。この特徴と、人狼が基本的に会話を通じてのみ行われるということが、プレイヤーの自動化という視点でとらえたときに特有の興味深い研究テーマを生み出している。

まず、状況判断がプレイヤーの言動に依存するため、おおくの局面で本質的に「正解」を知りえず、「推理」が必要である。推理する際にも、他プレイヤーの意図をモデル化するという高度な作業が必要となる。こうした推理やモデルの上に、人狼陣営はいやおうなく「嘘をついて騙す」ことが求められる一方、村人陣営はそえを「見破る」必要がある。これは見方を変えると、いかに他者を「説得」し「信頼を得る」かということでもある。

さらに、実際のゲームでは表情や音声、ジェスチャーなど非言語情報も大きな役割を果たす。すなわち、人狼プレイヤーの構築の研究は、人工知能を中心に、自然言語、対話、音声、心理学、エージェントなど多様な関連分野の研究とその統合が必要となる挑戦的な課題といえる。実際、人狼知能プロジェクトのオーガナイザーはそうした異なる分野から様々なアプローチで研究を進めている。

4. 人狼知能大会の自然言語部門

2017 年の大会より、従来のプロトコル部門に加え自然言語処理部門のタスクを開催した。プレ大会を GAT (Game AI Tournament) 2017 において、本大会を CEDEC (Computer Entertainment Developers Conference) 2017 において開催、結果の発表を行った。以下では本大会について述べる。

当面の現実的な目標としては、会話を通じたゲームが成立したといえるレベルを達成することである。既存の（雑談）対話システムそのままではシステム間の対話がほとんど成立しないであろう。実際の人狼ゲームでは雑談的な会話も頻繁に発生し、そこからある種の推理を行うこともあるが、自然言語部門ではまずはゲーム勝敗に直結する会話が中心となり、自然と使用する語彙も限定されある程度の意思疎通が可能ではないかと期待している。エージェント内でプロトコルを中間言語的に利用することも考えられる。

4.1 自然言語部門の設定

(1) 人狼ゲームの設定

対話に重点をおき分析や実装を容易にするためプレイヤー数が 5 の 5 人人狼とし、ゲーム当たりのターン数も少なくなっている。役職の内訳は占い師 1、人狼 1、裏切り者 1、村人 2 である。5 人人狼では日数が短く、役職の推測に利用できる情報が少ないのが特徴で、偶然の要素も大きい。

初日（0 日目）は特にゲーム情報を与えない。挨拶などを行うことを想定している。初日の夜（夜は会話以外のゲーム進行処理が行われる）は占いのみを行い、投票と襲撃はない。

1 日目以降は、20 ターン経過するか、全員が over という特別な応答を返すまで 1 日の昼が継続する。人狼知能サーバは同期式で通信を行っており、各エージェントから 1 発話ずつ収集してこれを 1 ターンと呼んでいる。エージェントから見ると 1 ターンは同時の発話であり、前ターンまでの発話情報しか参照できない。発話をしない skip という特別な応答も利用できる。

投票での得票一位が複数あったときは、単独一位が得られるまで投票を繰り返す。

(2) システムと対戦形式

枠組みとしては従来の人狼知能サーバを流用し、投票などゲーム特有のアクションについては従来通りとしつつ、プロトコルを用いていた「会話」に相当する部分はすべて自然言語（句読点・感嘆符・疑問符以外の記号を含まない日本語テキスト）のみとした。例外として特定のエージェントに対する発話を示せるよう、発話冒頭に「>>Agent[01] 発話本文...」という形のアンカー付与オプションを許した。アンカー自体は発話の一部とはみなさない。

```
1,talk,1,0,4,私は占い師だよ
1,talk,2,0,3,ぼくは占い師。Agent[02]の結果は黒だったよ。
1,talk,3,0,1,Skip
1,talk,4,0,2,俺は占い師だぜ
1,talk,5,1,5,Skip
1,talk,6,1,4,いや、そうは思わないな。
1,talk,7,1,1,僕は占い師だよ。
1,talk,8,1,3,Agent[02]はみんなを騙そうとしてるね。
1,talk,9,1,2,Agent[01]を占った結果人狼だったぜ
1,talk,10,2,5,Over
1,talk,11,2,2,Skip
1,talk,12,2,3,嘘は良くないよ、Agent[01]。
（凡例:日,アクション種別,発言 ID,ターン数,エージェント ID,
発話文字列）
```

図 1. 異種エージェント同士の相互対戦のログ抜粋例

アンカーに限らず、発話内でエージェントを指す際は「Agent[xx]」という形式で表わすこととした。

ネットワーク経由で対戦サーバに接続するリモート対戦を採用し、実装や計算資源は参加者が自由にできるものとした。また、発話リクエストから応答までの制限時間を 5 秒とした。

対戦方法は 3 通りを設定した。ひとつは、同一エージェントを 5 体用意して戦わせる自己対戦である。ふたつめは、5 種類の異なるエージェントを対戦させる相互対戦である。三つめは、4 種類の異なるエージェントと 1 名の人間を混在させて行う対人対戦である。機械同士の対戦は事前に自動的に行ってログを収集し、対人対戦は大会当日にリアルタイムで行った。図 1 に相互対戦のログ抜粋例を示す。

図 2. アバター表示によるプレイ画面の例

対戦表示用の UI として、宝塚大学東京メディア芸術学部渡邉研究室に表情や姿勢の変化するモーション付きアバターを作成いただき、大会当日にはこれを用いたデモンストレーションも行った（図 2）。

4.2 自然言語部門の評価メトリクス

(1) 勝敗率

人狼ゲームは属する陣営ごとに明確に勝敗が決定できる。エージェント間で十分に「意思疎通」がとれているのであれば、勝敗率は言語解析および生成の正確さとゲーム戦略の性能を総合した評価値となりうる。エージェントの実装は役職ごとに異なることが多いため、役職ごと、および全体で勝敗数をカウントし相互対戦による評価を行った。試合数はエージェントあたり 120 ゲームである。表 1 に本大会における勝敗数と勝敗率を示す。

チーム	村人	占い師	裏切者	人狼	総合
AITWolf	24/48 (0.50)	10/24 (0.42)	8/24 (0.33)	7/24 (0.29)	49/120 (0.41)
KELDIC	26/48 (0.54)	14/24 (0.58)	9/24 (0.38)	12/24 (0.50)	61/120 (0.51)
m_cre	14/24 (0.58)	18/24 (0.75)	16/24 (0.67)	16/24 (0.67)	83/120 (0.69)
forst	23/48 (0.48)	14/24 (0.58)	7/24 (0.29)	11/24 (0.46)	55/120 (0.46)
Kanolab	28/48 (0.58)	11/24 (0.46)	13/24 (0.54)	7/24 (0.29)	59/120 (0.49)

表 1. 各参加チームの相互対戦における役職ごとおよび全体の勝敗数と勝敗率

(2) 主観評価

勝敗率はわかりやすい評価尺度であるが、実際には人狼はチームプレイであること、対戦相手のレベルや振る舞いによって同じエージェントでも勝敗率の変動がありうることに注意が必要である。これらいずれの点も、そもそもきちんと「意思疎通」がとれているかが問題であり、そうでなければ勝敗率は対話とは関係のない部分で決定されている可能性もある。そのため、対話システムの評価としては勝敗率はまだ参考程度にみなすべき値であると考えられる。

一般に対話システムの評価は「対話が成立しているようにみえる」という表層的なものになりがちで、結果ある種の逃げに走るほうが良い評価を得られることも多い。人狼知能においては、会話に制約がないと同時にゲームを成立させ勝利に導くという目標があるため、表層的な発話に終始しては内容の一貫したふるまいにならず不自然に映るだろう。そのため適切な項目を設けた主観評価によって、より本質的な要素に高得点をつける評価が可能ではないかと考える。対話システムの評価にとどまらず、人狼ゲーム自体も勝ち負けだけでなく「面白いプレイ」であったかが重要であり、勝敗以外の評価軸をどう設定するかが研究テーマの一つである。

本大会では、下記 A～E の 5 つの主観評価項目を設定し、自己対戦および相互対戦ログを対象にそれぞれ 5 段階評価を行った。

A 発話表現は自然か

B 文脈を踏まえた対話は自然か

C 発話内容は一貫しており矛盾がないか（一貫性に関与しない発話は一貫していないとカウント）

D ゲーム行動（投票、襲撃、占いなど）は対話内容を踏まえているか（全行動のなかでの割合）

E 発話表現は豊かか。エージェントごとに一貫して豊かなキャラクター性が出ているか。

参加チーム以外のオーガナイザー5 名、外部から招へいした方 2 名からなる計 7 名の評価委員により主観評価を行った。表 2 に評価軸ごとの平均値を示す。

チーム	A	B	C	D	E
AITWolf*	2.14	1.85	2.14	2.00	1.00
Forst	3.28	3.00	**3.14**	3.14	2.00
Kanolab*	3.00	2.85	3.00	2.85	2.71
KELDIC	2.28	1.85	2.00	1.85	2.57
m_cre**	**3.42**	**3.14**	2.71	**3.85**	**4.00**

表 2. 主観評価の評価軸ごとの平均値（*は個別評価受賞）

さらに、各評価委員ごとにそれぞれの視点で最も良いと判断したチームを選定してもらい、各賞とした。当日は、外部評価委員 1 名が実際に対戦し、その選定によりリアルタイム対戦での評価も行った。

5. おわりに

本稿では、会話ゲーム人狼を日本語のやり取りで自動プレイする初めての試みである人狼知能大会自然言語部門の開催とその結果を紹介した。第一回ということもあり、特に対話システムとしてみると必ずしも十分な言語処理が行えていない面もあるが、まずは人狼をプレイすることはできたといえるのではないだろうか。今回の評価からは、作成者の人狼ゲームへの理解度が、システム動作の評価に大きく影響していたように思われる。評価軸やそのスコア基準については、これからも追加や改善の余地があろう。

今後はさまざまな発展が考えられる。リアルタイム対戦におけるアバターの影響は大きく、こうした実環境とのインタラクションは重要な課題である。直接的な発展としては、音声入出力の導入が考えらえるが、その分の性能低下を考えると当面はテキストのほうが分析が容易であろう。現在、同期式での対戦を行っているが、これを非同期にすればより自然な対話環境となり、異なる側面が見えてくる可能性がある。こうした発展を取り入れつつ、毎年定期的な開催を行い対話システムやエージェントの研究発展に寄与していきたい。

謝辞

本大会の開催には多数の皆様に多大な協力をいただいた。以下でお名前を挙げさせていただく。アバターデザインをご提供くださった宝塚大学東京メディア芸術学部渡邉研究室の渡邉哲意、石川雄仁、森下優香の各氏。自然言語部門の人狼知能サーバと UI 実装・運用に協力くださった静岡大学情報学部狩野研究室の箕輪峻、柴淳、三原直樹、小川ちひろ、滝波秋穂、真木裕子の各氏。評価、議論、大会運営にご参加いただいた井原渉（澪標アナリティクス）、高橋一成（株式会社人狼）、西森魅萌（人狼アイドル）の各氏。CEDEC 運営委員会の三宅洋一郎（スクウェア・エニックス）、今給黎隆（東京工芸大学）の両氏。大会運営と主観評価を担当いただいた人狼知能プロジェクトオーガナイザーの鳥海不二夫（東京大学）、大澤博隆（筑波大学）、片上大輔（東京工芸大学）、篠田孝祐（電気通信大学）、大槻恭士（山形大学）の各氏。ここに深謝申し上げる。

参考文献

[1] 片上大輔, 鳥海不二夫, 大澤博隆, 稲葉通将, 篠田孝祐, 松原仁: 人狼知能プロジェクト, 人工知能, Vol.30, No.1, pp .65-73 (2015)

[2] 鳥海不二夫, 片上大輔, 大澤博隆, 稲葉通将, 篠田孝祐, 狩野芳伸. 人狼知能 だます・見破る・説得する人工知能. 森北出版. (2016)

[3] “人狼知能プロジェクト.” [Online]. http://aiwolf.org/.

[4] 狩野芳伸、大槻恭士、園田亜斗夢、中田洋平、箕輪 峻、鳥海不二夫. 人狼知能で学ぶ AI プログラミング 欺瞞・推理・会話で不完全情報ゲームを戦う人工知能の作り方. マイナビ出版. (2017)